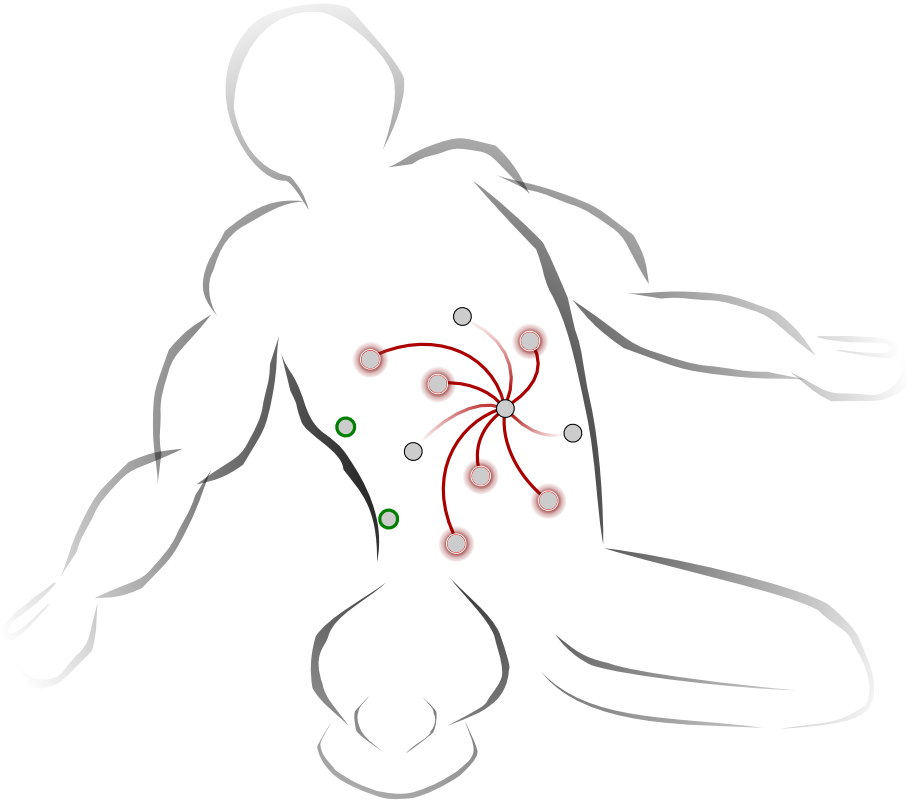
Mining microbiota signatures in human intestinal tract metagenomes



Sebastian Tims

Mining microbiota signatures in human intestinal tract metagenomes

Sebastian Tims

Thesis committee

Promotors
Prof. Dr M. Kleerebezem
Personal chair at the Host Microbe Interactomics Group
Wageningen University

Prof. Dr W.M. de Vos
Professor of Microbiology
Wageningen University

Co-promotor
Dr E.G. Zoetendal
Assistant professor, Laboratory of Microbiology
Wageningen University

Other members
Prof. Dr D. de Ridder, Wageningen University
Prof. J. Raes, University of Leuven, Leuven, Belgium
Prof. H.B. Nielsen, Technical University of Denmark, Lyngby, Denmark
Dr E.E. Vaughan, Sensus, Roosendaal

This research was conducted under the auspices of the Graduate School VLAG
(Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Mining microbiota signatures in human intestinal tract metagenomes

Sebastian Tims

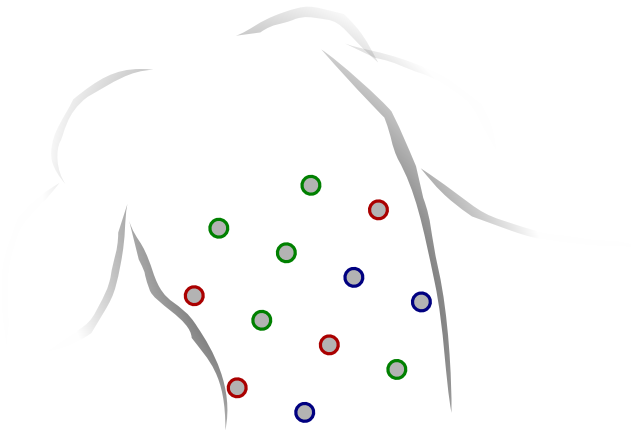
Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 2 May 2016
at 4 p.m. in the Aula.

Sebastian Tims
Mining microbiota signatures in human intestinal tract metagenomes
264 pages.
PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summaries in Dutch and English
ISBN 978-94-6257-693-3

"*Nothing exists for itself alone, but only in relation to other forms of life.*"
-- Charles Darwin

"*The greatest single achievement of nature to date was surely the invention of the molecule DNA.*"
-- Lewis Thomas

Table of contents



Chapter 1

General introduction & Thesis outline

Introduction

Microbial organisms have been functional external partners for thousands of years, aiding us humans with fermentation processes needed to produce various foods and beverages. Production of the long-time popular fermented beverages was already documented 7,000 BC in ancient Mesopotamia [1]. Although this allows us to date the first application of microbial organisms, the realization that microbial cultures can be actively utilized is only a century old. Alexander Fleming's discovery in the 1920s that a compound produced by a mold, i.e. *Penicillium notatum*, was capable of killing the bacteria *Staphylococcus aureus* [2], roused the interest in using microbes beyond food and beverage production processes.

Even though in 1885 Teodor Escherich already isolated the first gastrointestinal microbe [3] and in 1907 Elie Metchnikoff postulated the first links between microbial processes in the gut and human health [4], the microbes inside our own body remained fairly unexplored until the 1970s. This exploration began after the improvement of anaerobic cultivation methods and scientists quickly started to realize that the collective of microbes residing inside our own gastrointestinal tract is actually a complex community. Moreover, the composition of this community was found to be host-specific and even distinctive for the various regions in the GI tract [5-7]. Currently no consensus on the exact composition of the normal gut microbiota has been identified, although it becomes evident from literature that the most dominant bacterial phyla in the human gut are the Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria and Verrucomicrobia [8]. Next to these main phyla the Deinococcus-Thermus group, Lentisphaerae, Planctomycetes, Synergistetes, and the Tenericutes are detected in more intermittent occurrences among studied populations [8]. At deeper phylogenetic levels large variations in abundances are generally observed [7, 9]. Microbial organisms from the two other kingdoms of life, *Archaea* and *Eukarya*, are found in fecal material as well, though usually several orders of magnitude lower in abundance compared to the bacterial inhabitants [8] and therefore have received less attention.

Although in a prenatal stage we might already encounter microbes or their metabolites [10], our birth marks the first drastic impact concerning the gut microbiota due to a swift colonization of maternal and environmental microbes (for a review see [11]). Currently, the knowledge on the mechanisms controlling this colonization and the subsequent successions within the gut community are incomplete. Dietary nutrients supplied by the host are an obvious and longtime recognized factor for impacting the microbiota [12]. Recently different types of human breast-milk, categorized by maternal secretor status due to FUT2 gene alleles, have shown a "dietary" impact on the microbial colonization succession in the early life microbiota [13]. In adults changes in the daily dietary regime allows for the fastest noticeable response in the community profile of the gut microbiota, disregarding the use of devastating drugs like antibiotics. Vegetarian and vegan individuals display a rather distinct microbiota profile, i.e. reduced 16S rRNA gene counts of species belonging to *Bacteroides*, *Bifidobacterium*, and *Enterobacteriaceae*, as compared to omnivorous controls [14, 15]. Moreover, switching from a diet rich in animal products to a diet rich in plant products, either artificial or natural [16, 17], as well as (weight loss) diets with different carbohydrate amounts and polysaccharide types [18] have shown rapid adjustment of

the gut microbiota. Long-term dietary induced changes have been described as well and have a distinct impact on the microbiota [19].

With the gut microbiota acting like a virtual organ, an extensive cross-talk with the host is basically ensured from the moment of colonization and involves several aspects of the host physiology including (hyper)immune status. Therefore, besides diet, antibiotics, and external factors that influence host physiology, such as lifestyle or hygiene, host genotype provides another important factor for shaping the gut microbiota. Several lines of evidence indeed support the influence of host genetics on the microbiota: studies contrasting related and unrelated individuals have shown more similar microbiota community profiles for genetically related individuals [20-22], and studies focusing on specific host genotypes have shown the association of microbiota composition with single host genes such as *APOA1* and *NOD2* in mice [23, 24] and *FUT2*, *MEditerranean FeVer*, and *NOD2* in humans [25-27]. A more detailed discussion on the association of the gut microbiota and the host genotype is provided in **Chapter 2**.

The ever increasingly spectrum of methodologies to study the gut microbes at various levels, e.g. DNA, RNA, proteins, or metabolites, and their continuously improving (technical) throughputs is drastically changing the scientific field of microbial community analysis. This has caused an exponential increase of studies that report on the structure and composition of the microbial community residing in the human gut, especially with DNA-based approaches. Before moving on to the next section, which deals with the methodological "how" to study complex microbial communities, it is worthwhile to mention that current literature can be confusing with respect to the vocabulary used to describe the gut community [28]. This thesis will incorporate the definitions as proposed by Marchesi and coworkers [28], especially regarding the following terms: a microbiota is the full assembly of microbes present in a defined environment such as the gut or fecal sample (in many studies only the bacterial part of a microbiota is measured or considered); a metagenome is the collective genetic content belonging to the members of a microbiota; a microbiome encompasses the entire habitat with all its biotic (*Bacteria*, *Archaea*, lower and higher *Eukarya*, and viruses) and their genetic content as well as abiotic factors. The term microbiome actually describes all life forms (biome) of microorganisms but often is limited to the bacterial metagenome, i.e. only the genetic content of the bacterial part of the system, which is an area of confusion in the field. "Microflora" is another confusing (and outdated) term that has a longstanding history in medical and scientific literature. As microflora either means "microscopic plants" or "all plant life in a microhabitat", this term conflicts with the knowledge of the actual organisms in microbial communities, gut or otherwise. Although the term "microflora" is familiar for the general public and therefore still has commercial and educational value, medical experts and scientists should refrain from using it and make use of the more accurate term "microbiota".

Studying the gut microbiota

In the adult intestine the bacterial cell mass can reach values up to 1.5-2.0 kg, but this mass is comprised of 10^{14} bacterial cells and therefore outnumbers the human cells by a factor of ten [29]. Even more impressive is the millions of genes these microbes bring along [30], which outnumber the human gene pool by at least two orders of magnitude. Since these genes are brought in by at

least 2,500 species [31], most of which are strict anaerobes, it is obvious that advanced methods and approaches are needed to completely cover and describe this microbial community.

Cultivation-dependent methods

For a long time in the lively history of gut microbiology research, isolation and subsequent cultivation and characterization of the microbial strains was the only option. These traditional methods were estimated to cover about 10-30% of the total microbial population present in the gut [32-34]. Hence, as the cultivation captured mainly the "easily cultivable isolates", these methods were deemed unsuitable to provide the complete overview of the gut microbiota. It should be mentioned that recent years have seen advances in cultivation based studies due to high-throughput micro-scale cultivation strategies that may include steps to diminish the amount of "easily cultivable isolates" such as the use of antibiotics or phage cocktail treatments [35]. In general, the recent strategies for high throughput miniaturized culturing of previously uncultured isolates rely on improved mimicking of the natural habitat, which often requires co-culture strategies for multiple isolates from the same habitat (for a review see [36]). Nevertheless, cultivation-independent approaches are currently still the most employed methods and are generally considered to give the most rapid and complete overview of the gut microbiota.

Cultivation independent methods

By the end of the previous century, visualizing the extent and complexity of the remaining majority of the microbiota was made possible due to the rapid development of approaches targeting the 16S ribosomal RNA (rRNA) molecule or its encoding gene [37]. Analogously to other complex microbial ecosystems [38], the majority of the human gut microbiota has only been detected as a (partial) 16S rRNA gene sequence [39]. Hence 16S rRNA gene-based technologies are extremely useful to characterize the gut microbiota composition and its dynamics. For instance both denaturing gel gradient electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) are methods that start with the amplification of (a part of) the 16S rRNA gene, both resulting in a banding pattern on a gel-electrophoresis system that represents the compositional profile of the community from the tested sample. Beyond these initial profiles, however, identification of the organisms corresponding to the visualized bands is rather laborious as all bands must be separately excised, further amplified, and sequenced.

Cloning of 16S rRNA gene amplicons in clone libraries and subsequently sequencing of the (16S rRNA gene) insert is an alternative, yet also very laborious and low throughput, community profiling method. Moreover, it suffers from cloning bias depending on the host used. However, in the beginning of the 21st century phylogenetic, microarrays provided a much more comprehensive and high throughput analysis method for characterizing microbial communities, offering a higher resolution and diagnostic power as compared to the previous profiling methods such as DGGE, TGGE, or clone library sequencing [40]. Most developed phylogenetic microarrays target the 16S rRNA gene [41] and allow diversity assessment and taxonomic profiling, sometimes even allowing discrimination at the strain level [40, 42-44]. Currently, there are several phylogenetic microarrays available, ranging from universal ("all prokaryotes") arrays, such as the PhyloChip

[45], to gut ecosystem-specific arrays such as the Microbiota Array [46], the Human Gut Chip (HuGChip) [47], and the Human Intestinal Tract Chip (HITChip) [48] (Table 1.1). From these gut ecosystem arrays the HITChip is able to detect the largest amount of phylotypes, i.e. more than 1,000.

In the last ten years actual sequencing of (parts of) the 16S rRNA gene-pool has become a popular approach to investigate the gut microbiota composition. Even more can be done at the DNA level, such as the metagenomics approach in which a maximally representative DNA sample from an entire community present in an environmental sample is isolated and subsequently, often in a random fashion, sequenced. Both targeted 16S rRNA gene sequencing and random metagenomic sequencing have been made possible due to the onset of the next generation sequencing (NGS) techniques, discussed below. Although such DNA-based methods have greatly expanded our current knowledge regarding the biotic part of the gut microbiome, other methods are required to better understand the actual functioning and activity in the gut. In order to fully appreciate microbial activity, bacterial gene expression (metatranscriptomics), protein and enzyme production (metaproteomics), as well as the consequential metabolite abundance profiles (metabolomics) should be determined. However, these techniques can be more challenging to perform and are, compared to the DNA-based methodologies, still in their early stages. Hence, the remainder of this chapter primarily focuses on DNA-based methodologies.

Table 1.1 Comparison of four available phylogenetic microarrays.

	All prokaryotes	Gut ecosystem-specific bacteria		
	PhyloChip	Microbiota Array	HuGChip	HITChip
Platform	Affymetrix	Affymetrix	Agilent	Agilent
Target region	Design not restricted to specific V-regions	Design not restricted to specific V-regions	Design not restricted to specific V-regions	V1, V6
Nr. of probes	1,016,064	16,223	4,441	4,809
Probe design database size	59,959 representative sequences *	852 representative sequences †	1,052 sequences	1,140 sequences
Probe length	25-mers (overlapping)	25-mers	25-mers	24-mers (overlapping)
Output range	1,464 families (10,993 sub-families)	775 species-like groups	66 families	1,041 species-like groups
Reference	[45]	[46]	[47]	[48]
Year of publication	2010	2009	2013	2009
Citations ‡	369	53	14	226

*) Operational Taxonomic Units with an average sequence divergence of 0.5%

†) Operational Taxonomic Units with an average sequence divergence of 2.0%

‡) As of January 2016, Scopus [http://www.scopus.com/]

Next generation sequencing technologies

The NGS based community profiling technologies have been quickly adopted and integrated in the majority of gut microbiota studies world-wide, especially due to the various possibilities to outsource sequencing, the ever increasing throughput, and the cost reduction [49]. For community-wide composition profiling, via partial 16S rRNA gene sequencing, the first NGS technology that rapidly became very popular was the 454 pyrosequencing platform (Roche), but in recent years the Illumina based sequencing technologies (Illumina, USA) have taken over [50]. Besides 454 pyrosequecing and Illumina based technologies three more technologies can be considered to be from the, at the time of writing, same line of NGS technologies: the SOLid system (Applied Biosystems), the Ion platforms (Life Technologies), and SMRT system (Pacific Biosystems). The principles as well as the advantages and disadvantages, with respect to gut microbiota research, of these five NGS technologies are briefly described below (for a more extensive review see [51]). To multiplex samples, most NGS technologies support the so-called "barcoding" in which a barcode sequence, unique for every sample, can be placed in front of the (primer) sequence via PCR or ligation.

454 pyrosequencing

In 2005, 454 pyrosequencing was the first NGS technology that became available [52] and was for several years the preferred NGS approach in microbiota research due to its relatively long reads, eventually up to 1,000 bases, allowing a more phylogenetic discrimination and identification with each 16S rRNA gene sequence read generated. This technology starts with preprocessing the target DNA by attaching adapter sequences to it, either via PCR primers that contain the adapter sequence or via ligation. These adapters are used to bind the preprocessed target DNA to beads, i.e. one DNA molecule per bead, which are subsequently subjected to emulsion PCR to populate the bead with a multitude of copies of the initial DNA molecule. After the emulsion PCR, the beads are spread over a so-called PicoTiterPlate, which is a chip with micro-wells that can each hold a single bead. Each of the micro-wells also contains the required enzymes for a sequencing-by-synthesis reaction: a DNA-polymerase, an ATP sulfurase, and a luciferase. Next, the actual pyrosequencing is initiated by controlled additions of each nucleotide (A, C, G, T), which trigger pyrophosphate release and a subsequent light pulse due to the luciferase activity. Through optic fibers connected to each micro-well the light pulses are recorded and the corresponding sequence can be inferred from this information. When the target sequence contains stretches of the same nucleotide in a row (homopolymers), they will all be added in the same step and this process will generate more intense light pulse. Although in theory such homopolymer detection is feasible, it turns out in practice that the detection of larger homopolymers are frequently incorrect and present one of the largest drawbacks of 454 pyrosequencing.

Another disadvantage of 454 pyrosequencing is the relatively high costs per base read, which is substantially higher as compared to other NGS technologies of the same generation, i.e. Illumina, SOLiD, or Ion Torrent, although these generate shorter read-lengths (Table 1.2). The incredible rate of innovative improvements in the available sequencing technologies is illustrated

by the notion that approximately 10 years after its introduction, the 454 pyrosequencing platform will not be supported anymore in 2016.

Illumina

The first NGS platform competing with 454 pyrosequencing came in 2006 with the Illumina Genome Analyser (GA) [www.illumina.com/technology/solexa_technology.ilmn]. Currently, Illumina technologies have expanded GA with HiSeq, MiSeq and most recently with NextSeq chemistries. The general method of the Illumina technologies is sequencing-by-synthesis. Similar to 454 pyrosequencing, Illumina technologies starts with preprocessing the target DNA by attaching adapter sequences to it, either by PCR or by ligation. Subsequently, the pool of adapter-labeled target DNA is immediately washed over a flow cell containing a field of oligos complementary to the adapter sequence, enabling binding of the target DNA. After binding each target DNA molecule is replicated, generating clusters of identical copies of the original molecule. Next the sequencing-by-synthesis starts with a process remotely similar to "old-fashioned" Sanger sequencing [53], except that the dye terminators on the free nucleotides are reversible and thus allow repetitive dye-terminator incorporation combined with imaging and subsequent terminator removal to proceed to the next step [54]. Multiplexing can be achieved via two strategies: 1) adding a barcode sequence similar as in 454 sequencing, 2) through one or two separate indexing reads which are located outside the primer region, requiring separate sequencing runs (that read "away" from the target DNA molecule). For throughput details of each of the Illumina chemistries, at the time of writing, see Table 1.2. In general the Illumina chemistries are more cost effective per generated base, have a lower error rate and do not suffer from homopolymer detecton issues, but runs may take somewhat longer (depending on the desired depth) and produce shorter reads as compared to 454 pyrosequencing.

SOLiD

The third NGS technology that came to market was SOLiD [55]. For SOLiD the library preparation is similar to 454 pyrosequencing and the Illumina technologies, i.e. adapter sequences are added to the target DNA molecules, which are subsequently captured on beads (with the same one DNA molecule per bead standard) and amplified via emulsion PCR. These beads are, via the adapter on the 3' end of the DNA molecules, covalently attached to a glass slide. In contrast to 454 pyrosequencing and the Illumina technologies this method is a "sequencing-by-ligation" approach. SOLiD uses a mixture of labeled oligonucleotides, accounting for all possible di-nucleotides, which are used to "query" the target DNA molecule with a ligase after the initial addition of a random primer. In order to completely comprehend SOLiD technology one must understand the labeled oligonucleotide mixture, the use of random primers, as well as the rinse-and-repeat approach (for a more extensive review see [56]). In short, via addition of a random primer, nucleotide readings at regular spaced intervals are possible with the labeled oligonucleotides. After this first round, the random primer and added oligonucleotides are removed from the target DNA molecule and the "query" is repeated with a random n-1 primer allowing further bases to be identified. With the readings of all ligation rounds the actual target

DNA molecule sequence can be deduced of the fragmented DNA. Due to the short read-length inherent to this NGS technology multiplexing of samples through barcode sequences attached to the actual target sequence is ill-advised, instead multiplexing through a separate indexing run, similar to Illumina technologies, is available. SOLiD has the lowest error-rate of the NGS technologies and typically generates a larger numbers of sequences per run compared to 454 pyrosequencing but less than Illumina, while the read length is the shortest of the NGS technologies and this has been the biggest drawback of SOLiD. The costs of SOLiD sequencing per base generated is relatively low, although Illimuna has the lowest cost per base.

Ion Torrent

Chronologically the last NGS technology introduced is Ion torrent, which became available in 2010 with the personal genome machine (PGM) and was aimed to bring NGS to smaller laboratories as it brought the highest throughput for the lowest price. Ion Torrent-based sequencing shares a great deal of overlap with 454 pyrosequencing: target DNA is captured on beads, the beads undergo an emulsion PCR procedure and are subsequently distributed over a micro-well plate where a sequencing-synthesis-approach is initiated that records the addition of every nucleotide. Multiplexing of samples is achieved via the same barcoding principle as well. However, the actual recording system is unique for Ion Torrent as the micro-well plate is a semiconductor chip which contains pH sensitive micro-detectors in the wells. Whenever a free A, C, G, or T base is incorporated in the growing DNA strand a proton is released that causes a slight pH decrease. Hence for every controlled nucleotide addition on the plate the wells in which the nucleotide is incorporated can be detected and the target DNA molecule sequence can readily be inferred. Despite the different detection method, Ion Torrent has a similar homopolymer detection problem as 454 pyrosequencing since the accuracy of measuring multiple proton releases is not quantitatively reliable enough. Interestingly, Ion Torrent systems principally can generate sequence reads much longer DNA lengths compared to 454 pyrosequencing; however the practical operating procedure imposes a strict size selection limiting the longest read at approximately 400 bases (in the PGM system). Bidirectional sequencing of the target DNA molecule is possible with the Ion Torrent system, though currently this appears to be relatively unreliable [57], which in combination with the target sequence length restriction prevented the Ion Torrent systems to take a larger position in microbiome research, despite the high throughput and relative low sequence costs.

SMRT

Ultimately, researches would like sequencing directly from the target DNA molecule as they appear in *in situ* or *in vivo*, without the potential biases introduced by PCR amplification preprocessing steps. Such single-molecule sequencing platforms are currently available with the SMRT technology (Pacific Biosciences) currently being the most successful on the market. Due to the use of DNA polymerases that are directly linked to zero-mode waveguide detectors on a chip, SMRT sequencing is able to image in real time the incorporation of phosphor-linked dye-labeled nucleotides during DNA synthesis. Hence no PCR amplification step is required and this

feature discriminates SMRT sequencing from the other NGS technologies. Single sequencing reads on the SMRT sequencing (PacBio system) are plagued by a 15% error rate in base calling that is currently negated by the system by the addition of a circular consensus sequencing (CCS), which repeatedly sequences the circularized DNA target molecule providing a high fold coverage of the target molecule that enables the correction of erroneous base calls. Multiplexing with the PacBio system could theoretically be achieved by ligation of a barcode sequence, but this is not frequently employed.

Table 1.2 Comparison of typical high throughput sequencing systems from the five major platforms of the current generation. Adapted from [56-59].

Technology/ Platform	Amplification type	Chemistry	Error rate	System	Seq per run	Run time
454	Emulsion PCR	Pyrosequencing (seq-by-synthesis)	1%	GS-FLX +	1 M (2 x 600 bp)	20 h
Illumina	Bridge Amplification	Reversible dye terminator (seq-by-synthesis)	~0.1%	Miseq	25 M (2 × 300 bp)	5-55 h
				NextSeq	130–400 M (2 × 150 bp)	12-30 h
				HiSeq	2,500 M (2 × 150 bp)	24-48 h
SOLiD	Emulsion PCR	Oligonucleotide 8-mer chained ligation (seq-by-ligation)	≤0.1%	5500 SOLiD™	60-120 M (2 x 35 bp)	24 h
Ion Torrent	Emulsion PCR	Proton detection (seq-by-synthesis)	2%	PGM	1.5-3 M (1 x 400 bp) *	2–3 h
SMRT sequencing	N/A (single molecule)	Phospho-linked fluorescent nucleotides (seq-by-synthesis)	10-15%	PacBio	35 M (1 x 8,500 bp)	1.5 h

* paired-end sequencing available but not advisable due to low quality reverse read

Community-wide profiling strategies

As mentioned above, the two most used ecosystem-wide sequencing strategies currently are targeted 16S rRNA gene sequencing and random metagenomic shotgun sequencing. With targeted 16S rRNA gene sequencing approaches, which utilize the conserved and variable regions within the 16S rRNA gene that is present in all microbes [60], the ultimate question to be answered is “who is present?”. Once it is established “who is there” often a potential functional profile is inferred from this. With random metagenomics, which is performed on a maximally representative DNA sample extracted from an environmental sample, the ultimate question to be answered is “what can they do?” or perhaps more accurate, as DNA reflects the potential capacity, “what *could* they do?”. After assembly of metagenomic sequence data, genes are assigned in the resulting contigs, and are annotated by computational intense algorithms [61], which enable a more complete understanding of the genetic potential within the gut microbiota than any preceding technology. Processing of the enormous amounts of data generated by NGS is more and more dependent on the available bioinformatics capacities (for a review see [58]), not only related to the hardware of the computational platforms, but in particularly related to the

limited number of experts. Therefore, the computational biology aspects of microbiota research pose an increasing barrier for effective and successful analysis, particularly because the resources needed for the data generation of either targeted 16S rRNA gene sequencing or the random metagenomic shotgun sequencing are continuously declining.

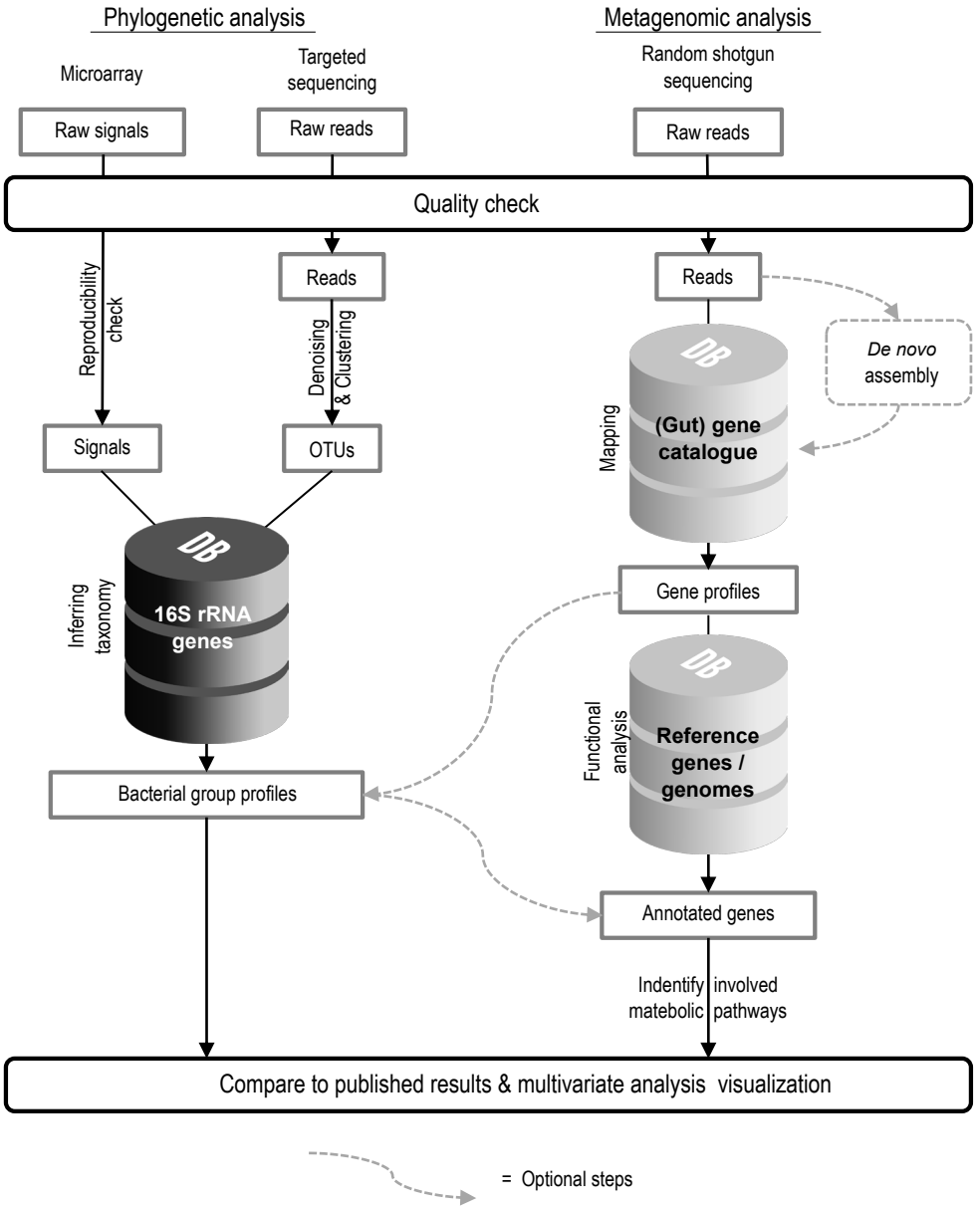


Figure 1.1 Schematic representation of the flow of sequence generation and analysis is in microbiota composition analysis at the level of the phylogenetic composition or the function gene repertoire composition.

The two ecosystem-wide sequencing strategies target different aspects of a microbial ecosystem, each with its own advantages and disadvantages. Targeted 16S rRNA gene-based approaches that yield high dimensional data, such as phylogenetic microarrays and 16S rRNA gene amplicon sequencing, visualize the compositional profile and thus provide insight on the identity of the micro-organisms present. Basically, metagenomic shotgun sequencing sacrifices such compositional resolution but provides a direct view of the functional potential of the studied microbial ecosystem. Both approaches should not be regarded as competing strategies as they both can yield complementary information. Figure 1.1 provides a schematic overview, generalizing the various alternative routes, of typical analysis paths for both strategy types and shows potential cross-over steps. Both analysis paths start with suitable quality checks on the generated data. For microarray analysis sample are usually analyzed *in duplo*, which allows an assessment of the technical reproducibility, e.g. the HITChip samples are loaded with different dyes on independent arrays and only samples with a >0.98 reproducibility are further processed [62]. In targeted 16S rRNA gene amplicon sequencing the quality filtered reads are usually subjected to further denoising and are subsequently clustered into operational taxonomic units (OTUs). Both the probe sequences of microarrays and the representative sequences of the OTUs are aligned to a 16S rRNA gene reference database to assess phylogeny and taxonomic identity is inferred, resulting in bacterial group profiles. Optionally bioinformatics tool such as PICRUSt [63] can be used to infer a functional pathway prediction.

Quality filtered reads of random shotgun metagenomic sequencing studies can be mapped to an annotated reference gene/genome databases, assembled into contiguous sequences (contigs), or both depending on the read lengths and quality. Through the mapping procedure the functional potential of the samples is captured and important metabolic pathways can be distilled from these profiles. Several options for the mapping procedure as well as the reference gene/genome databases exist. Taxonomic identity of the reads can be inferred from the mapping to the annotated reference gene/genome databases, although most genes offer much less phylogenetic resolution as compared to phylogenetic marker genes such as the 16S rRNA gene. Finally, for both gut community profiling methodologies the data can be treated individually or combined with data from other relevant studies. The abundances from both the taxonomic or functional profiles are subjected to various multivariate analyses and visualization. At this stage, it is possible to incorporate sequence data from other relevant studies, or the data can be treated individually. The abundance of the various OTUs is then subjected to elaborate (multivariate) statistics and the resulting patterns are visualized. Table 1.3 compares the technical aspects of the two afore mentioned targeted 16S rRNA gene based approaches and random shotgun metagenomics.

Noteworthy is that the first mentioning of metagenomics in literature was by Handelsman and coworkers [63], which was an application that used the genetic content of environmental samples for functional screening after expression in a cloning host and this can be regarded as functional metagenomics. Functional metagenomics provides a bridge between predictions based on DNA sequences and actual functional characterization of these DNA sequences. Currently, there are two options for functional characterization of metagenomic libraries available: 1) Random cloning of small DNA fragments (<3-15 kb) or large DNA fragments (25+ kb) into

a suitable cloning host that allows (inducible) expression to generate an expression library for enzyme or pathway screening; 2) Random cloning of either small or large DNA fragment into a cloning host that allows (inducible) expression in conjunction with an enrichment strategy. Next, both options will be briefly decribed.

Table 1.3 Pros and cons of approaches targeting genomes and microbes.

	Targeted 16S rRNA gene-based approaches		
	Microarray	**Amplicon sequencing**	**Random shotgun metagenomics**
Phylogenetic resolution	Known & related phylotypes	New & known phylotypes	Depending on read N & read length (& assembly ability)
Reproducibility	High reproducibility	Depending on read N (and filtering settings)	Depending on read N (and filtering settings)
Community depth	Fixed depth (0.1 – 0.01%)	Read N determines depth	Read N determines depth Relatively lowest depth
Relative costs (at comparable depth)	Low	Low	High
Relative computational requirements	Low (Simple pipeline)	Medium ‡	High (mapping) - very high (*de novo* analysis)
Community target complexity	16S rRNA gene (± 3 copies / genome)	16S rRNA gene (± 3 copies / genome)	Microbial genes (± 3,000 gens / genome)
Functional (potential) analysis	Deductive †	Deductive †	SNP analysis Metabolic pathway identification

‡) OTU clustering is the most computational intensive step

†) Automated functional prediction tools available, e.g. PICRUSt [62]

Functional mining by screening a metagenomic expression library containing either small or large inserts (option 1) could be employed to screen for different types of functional capabilities. Depending on the screening strategy and quality of the (sheared) metagenomic DNA different cloning vectors can be used, each with its own advantages and limitations. For instance "normal" plasmids can generally not harbor inserts larger than 10-15 kb, but can achieve higher expression levels due to higher copy numbers, whereas the low- to single-copy cosmids and fosmids (25-40 kb), or BACs (100-200 kb) cloning systems can potentially retain entire pathways [64]. Function-based screening has the obvious advantage that the identified clones are selected on basis of the function they encode, and allow identification of completely novel genes or pathways.

The potential of metagenomic expression libraries still faces constraints both at DNA sequence level as well as on the expression level. From the DNA sequence perspective the limitations are straightforward and DNA sequences need to be intact, i.e. containing the beginning and end of the entire sequence needed to express a gene or pathway, need to be in the right orientation if a small cloning vector (plasmid) is used, and translation signals and codon usage must be compliant with the translation machinery of the cloning host. Although, theoretically, these limitations can be overcome provided that enough DNA can be isolated and, in order to capture all DNA fragments from a community, only the amount of clones that can be obtained would be the real constraint. Indeed there have been various reports describing

successful functional screening studies of metagenomic expression libraries (for a review see [65]). Although incubations of an environmental sample with a stable isotope labeled substrate [66] can reduce the amount of clones needed to detect a given function by physically isolating the DNA of metabolically active bacteria prior to the cloning step, such methodologies are inherently challenging to perform correctly with fecal samples containing a large proportion of strictly anaerobic bacteria.

Various difficulties at the protein expression level will pose further practical limitations which are often directly related to the cloning host used and its translation machinery's constraints [67]. Even when translated successfully, the host employed for heterologous expression will have to provide appropriate conditions for the protein to achieve its proper folding and also should provide potential essential cofactors to allow the protein to display its function. At the same time, the expressed protein should not be detrimental to the cloning host, and in many cases may need to reach its proper subcellular localization in order to function. Presently, there are three types functional selection or enrichment approaches reported for metagenomic expression libraries (option 2): selection based on measurable, phenotypic, changes caused by the functional activity of interest; selection based on heterologous acquisition of functions complementary for the utilized cloning host; selection based on induced gene expression. Each of these section approaches have their own advantages and disadvantage and are nicely reviewed elsewhere [64].

Host health & the gut microbiota

Since the anaerobic cultivation advances in the 1970s and especially due to employment of the (high-throughput) cultivation independent methodologies described in the previous sections, the gut microbiota has increasingly been recognized for being a functional counterpart to its host genome [29, 68, 69]. With a metabolic activity status that is postulated to equal that of the liver [70], the gut microbiota can be considered to be a virtual organ that complements host-function by conversion and fermentation of indigested food components [71], production of specific essential nutrients [72-74], conversion of bile and certain steroid hormones [75, 76], maturation and modulation of the immune system [77, 78], and colonization resistance to pathogens [79]. Aberrant microbiota compositions or decreased stability have been linked to various diseases and syndromes, ranging from the more obvious GI disorders such as Crohn's disease [80] to, on first sight, less direct ailments such as autism [81] and Parkinson's disease [82]. Given the role for the microbiota in nutrition, particularly the conversion of nutrients, it is not surprising that the microbiota is also associated to host energy homeostasis [83-87]. Interestingly, the various possible mechanisms explaining the host-microbe association in energy homeostasis expand far beyond the microbiota "just" modulating the direct energy supply to its host, such as microbial modulation of host satiety and low-grade chronic inflammation (for a review see [88, 89]). Considering the vast amount of processes to which the gut microbiota is linked, it is not surprising that a lot of effort has been put into mapping the factors that influence the microbiota and establishing how the microbiota can be modulated. In the next sections, some highlights are provided on the associations of the gut microbiota and several disorders that have received the main focus in the field in the last years: the metabolic syndrome components obesity and type 2 diabetes (T2D),

inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), and colorectal cancer (CRC) (Table 1.4).

Table 1.4 Examples of documents microbial associations with human diseases and disorders. Table adapted from [90].

Disorder or condition	Microbial group or microbiota characteristic association		Reference
	Increase	Decrease	
Obesity	• Firmicutes:Bacteroidetes ratio † • *Bacteroides* †		[22, 91, 92]
		• Bacterial diversity & bacterial (gene) richness	[22, 91-94]
	• Actinobacteria • *Prevotellaceae*	• *C. leptum* group (*Ruminococcus flavefaciens*) • *Bifidobacterium* • *Methanobrevibacter*	
T2D	• Bacteroidetes • *Bacteroides* spp. • *Clostridium* spp. • *Desulvovibrio* spp. • *Eggerthella lenta* • *Escherichia coli* • Beta-*proteobacteria* class	• Firmicutes: - *Clostridia* class - Butyrate-producing organisms (*Roseburia* spp., *Faecalibacterium* spp., *Eubacterium* spp.)	[95, 96]
IBD		• Bacterial diversity	[97-105]
	• *Clostridium* • *Enterobacteraceae*^{UC} • Gamma–*proteobacteria*:^{CD} - adherent invasive *Escherichia coli*^{CD}	• Firmicutes: - *Clostridium leptum* & *coccoides* group - *Faecalibacterium prausnitzii* - *Roseburia* - *Phascolarctobacterium* • Bacteroidetes	
IBS	• Firmicutes:Bacteroidetes ratio^{A}		[106-108]
	• *Alistipes*^{P} • *Clostridium*^{A} • *Dorea*^{AP} • *Haemophilus influenza*^{P} • *Ruminococcus*^{AP}	• *Bacteroides*^{P} • *Bifidobacterium*^{A} • *Faecalibacterium*^{A}	
CRC	• *Fusobacterium* spp. • *E. coli* (pks+)		[109-112]

IBS, irritable bowel syndrome; A, found in adults with IBS; P, found in pediatric patients with IBS; IBD, inflammatory bowel disease; CD, only found in Crohn's disease patients; UC, only found in ulcerative colitis patients; CRC, colorectal cancer; pks+, polyketide synthase positive; T2D, type 2 diabetes. † Varying results, both in line and in opposition against the initial lower Firmicutes:Bacteroidetes ratio (mainly driven by the lower levels of Bacteroides) found in mice [113].

Metabolic syndrome: obesity and type 2 diabetes

A little over ten years ago, the research on the gut microbiota and its relationship with obesity was brought to a next level by studies in germ-free mice. The original finding was that at phylum level the Bacteroidetes to Firmicutes (B:F) abundance ratio was lower in obese mice [113] and this phenotype could be transferred by fecal transplantation to lean germ-free mice. The microbiota relationship with obesity was subsequently confirmed in a small cohort of

obese human individuals (body mass index, BMI>30) that during a dietary intervention trial indeed lost weight and concomitantly gained elevated relative abundance of Bacteroidetes [85]. However, various studies have followed and contradicting findings were reported for this B:F ratio [22, 91, 92]. The latter does not come as a surprise as beyond BMI differences human study populations are quite heterogeneous in terms of exposed to various other influences such as lifestyle and diet [114], host genotype, and hormonal factors [115], all of which can influence the gut microbiota. Moreover, BMI itself is a rather inaccurate indicator of health status, as 25% of obese subjects are in fact considered to be metabolically healthy [116]. The remaining 75% is often not only challenged by their overweight but also one or several other metabolic conditions such as glucose intolerance or diabetes, hyperlipidaemia, hypertension, and low grade inflammation [117]. Hence obesity is a complex and heterogeneous condition, which encompasses much more than only weight or BMI.

The shortcomings of BMI as a marker for unhealthy obesity was further underlined by a metagenomic study where bacterial richness, as measured in normalized gene counts, in obese subject displayed a bimodal pattern which correlated to metabolic syndrome marker measurements that were associated with obesity-status but not with weight or BMI [94]. Interestingly, obese individuals with a high gene count were marked by a higher prevalence of alleged anti-inflammatory species, more genes potentially involved in organic acid production (among which butyrate pathways), reduced prevalence of alleged pro-inflammatory species, and metagenomes with reduced gene frequencies of genes potentially involved in oxidative stress response as compared to the obese individuals with low gene counts. In a separate study, diet-induced weight-loss was shown to be increased in individuals with high gene richness as compared to those with low gene richness, which was accompanied with an enhanced improvement of the corresponding metabolic health status in the high gene richness individuals [118]. Moreover, two independent metagenomic studies in T2D cohorts [95, 119] resulted in the identification of metagenomic co-abundant gene clusters that in terms of predictor quality for T2D outperformed classical T2D markers such as those identified in the human genome through genome wide association studies (GWAS) and BMI. Analogous to the findings of the metagenomic studies related to obesity the abundance of genes predicted to be involved in oxidative stress responses appeared to be increased in T2D patients [119], whereas genes predicted to be involved in butyrate biosynthesis were decreased [95]. Interestingly, at a phylogenetic level both metagenomic studies in T2D cohorts resulted in different marker species [95, 119], which suggests that such biomarkers are population dependent.

Despite the contradicting findings that have been published, together these studies on obesity and the associated metabolic diseases testify for the significance of the gut microbiota in host energy homeostasis. While the exact mechanisms remain unclear, in general the gut microbiota in these metabolic conditions are less diverse and/or rich, have a more inflammatory prone composition profile, and often appear to have less butyrate production potential.

Inflammatory bowel diseases

In the early days, research on IBD was focused on finding a single perpetrator but in the last decade it has become evident that the entire gut microbiota is affected in IBD patients. Conflicting results have been reported regarding the level of the Bacteroidetes phylum [102, 120] though it appears that there may be a shift within this phylum in IBD patients where *Bacteroides fragilis* comes out as the most dominant species [121]. While two main types of IBD, i.e. Crohn's disease (CD) and ulcerative colitis (UC), have some distinct microbiota characteristics, in both groups of patients a consistent reduction in clostridia members was observed [122]. In addition, in UC patients a marked decrease in *Akkermansia muciniphila* was detected, both in mucosal and fecal samples [123, 124]. Next to the alterations of the Firmicutes and Bacteroidetes, often various members from the Proteobacteria phylum increase with the chronic inflammation status [125, 126].

The Firmicutes reduction has also been shown to be accompanied with a decrease in diversity within this phylum [99] as well as an overall reduction in metagenomic gene counts. A recent extensive study analyzed the 16S rRNA gene sequences of 231 IBD patients and healthy controls as well as 11 metagenomes from a sub-selection of this cohort [105]. From the extensive data set the relation between disease status and Firmicutes and *Enterobacteriaceae*, a family of Proteobacteria, levels was in line with prior publications. However, a striking amount of (inferred) microbial pathways were associated with heightened disease status [105], such as increase in oxidative stress pathways (presumably to counter host oxidative bursts that are used by innate defense), a decrease in the normal pathways for carbohydrate metabolism, amino acid and SCFA biosynthesis. Considering the similar microbial pathway shifts detected in T2D [95], there appears to be a general response of the gut microbiota when it resides in environment marked by chronic inflammation and heightened immune activity.

Irritable bowel syndrome

IBS is a so-called functional bowel disorder, which means it is diagnosed solely on symptom-criteria, after ruling out signs of other diseases or disorder. Although the IBS etiology is still largely unknown, it is regarded as a multi-factorial disease and several lines of evidences indicate that (alterations of) the gut microbiota play a role, potentially through low-grade intestinal inflammation [127-130]. The main indication lies in the fact that IBS symptoms can be alleviated by the use of antibiotics, prebiotics, and probiotics [131, 132]. Unfortunately, most of the early microbiota studies were limited in analysis depth and numbers of subjects included [133]. For post-infectious IBS (PI-IBS) a causal relationship between the gut microbiota and PI-IBS has been proposed in literature [134] and the microbiota profile appears to be different form healthy controls and more similar to diarrhea predominant IBS [135]. Notably, the latter results may indicate in general that infectious organisms, which are not part of the normal microbiota, can have pronounced effects even after they have supposedly been eliminated from the gut ecosystem. Combination of the microbiota data of different PI-IBS cohorts confirmed the robustness of the observations and provided suggestions for patient stratification and treatment [136].

An altered gut microbiota composition in IBS patients as compared to healthy controls has recently been established with high throughput DNA technologies, both in adult [106, 137-139] and pediatric populations [108, 140]. Nevertheless, a typical IBS microbiota composition or specific traits thereof has not been established, mainly due to the heterogeneity of IBS itself but also due to the miscellaneous approaches that have been employed [141]. However, common observation have been reported and often bacteria related to the Firmicutes species *Ruminococcus gnavus* are found in high levels [139, 142, 143], whereas some *Bacteroides* groups are found in low levels amongst IBS patients [106, 139, 144]. Interestingly, high-throughput sequencing studies have identified subgroups of IBS patient based on their microbiota composition, which was in dependent from the more traditional symptom based subtypes [106]. Taken together, the more in-depth and high-powered studies from the last 5 - 10 years indeed strongly suggest that IBS patients harbor an altered gut microbiota and encourage future therapeutic avenues to take the microbiota into account.

Colorectal cancer

Current literature indicates that the gut microbiota plays an important role in (the onset of) CRC, both for colitis-associated [112] or sporadic CRC [145]. Similar to the other diseases described in the sections above, no single causative micro-organism has been found although several studies have incriminated *Fusobacterium* members [109-112]. Metagenomic and metatranscriptomic analysis revealed a highly variable and subject specific microbiome, yet specific bacteria were found to colonize the actual tumor tissue or the surrounding healthy tissue [146]. Intriguingly the tumor tissue was suggested to be selective for commensal species and even bacteria related to probiotic strains, while potential pathogens such as *Enterobacteria* were hardly present [146].

Global studies comparing African Americans with Africans have further implicated the gut microbiota in the onset of CRC. Africans have a *Prevotella*-dominated microbiota whereas African Americans, who have a much higher risk of CRC development which, have a *Bacteroides* dominated microbiota [19]. Since the genetic predisposition of African Americans is unlikely to be drastically different from that of Africans, factors such as diet, lifestyle and environmental exposures are more likely to have modulated the microbiota to a more CRC prone composition. Indeed, African Americans subjected to a 2-week high fiber (>50 g/day) diet decreased the levels of mucosal biomarkers of cancer risk, whereas Africans subjected to a high fat, high protein, low fiber Western diet in the same time-span increased the levels of mucosal biomarkers of cancer risk [17]. These changes were accompanied by vast changes in fecal microbiota co-occurrence network structures and fecal metabolome [17], strongly suggesting that the gut microbiota plays an important role in CRC, although changes in the intake of red-meat and the associated chemical routes cannot be excluded (see below).

In line with the global microbiota studies, others have reported on the metabolic functioning of the gut microbiota in relation to CRC etiology and most often fiber metabolism is mentioned as a key activity [147]. Reduction of butyrate-producing bacteria is frequently identified in CRC patients [148] and this has been reproduced in animal models [149]. Furthermore, bacterial sulfate metabolism is implicated in CRC development as hydrogen sulfide, a bacterial end product,

stimulates mucosal hyper-proliferation and is hardly detoxified in the colon of CRC patients [150]. This mucosal hyperproliferation is due breaking of the disulfide bonds in polymeric mucin by (bacterial) hydrogen sulfide, which increases the mucus layer permeability and exposes the epithelial surface cells to cytotoxins (heme derived and other) and subsequently initiates hyperproliferation to compensate the cell loss [151]. Interestingly, it has been hypothesized that genetic predisposition for CRC may also act via the carbohydrate landscape in the distal gut, which is depends on host genotype [152], to ultimately recruit a (metabolically) modified microbiota that contributes to CRC development.

Modulating the gut microbiota for health benefits

Diet has received increasing attention as a modulation tool for the gut microbiota composition and function. After all, short-term diet-induced changes can already be detected after 24 hours [18]. Having carbohydrate fermentation processes as a core activity, the gut microbiota is the primary driver of energy and carbon fluxes in the colon. Often these saccharolytic fermentation processes consist of four stages performed by different members of the microbial community: First, the (usually) most prevalent bacterial species that perform the initial steps in the breakdown of complex plant polysaccharides into oligosaccharides [153]; next, specialized oligosaccharide degraders, such as bifidobacteria, provide the final breakdown steps of these carbohydrate chains generating short chain fatty acids (SCFAs) and gases; secondary fermentation takes place by "scavenging" microbes that utilize the final breakdown products of the previous step and convert these into SCFAs and gasses as well; finally, the specialized "keystone" species, such as methanogens, reductive acetogens, and sulfate-reducing bacteria, use these SCFAs and gases as energy sources and substrates [154]. The main SCFAs resulting from the fermentation processes have been shown to affect intestinal physiology (butyrate and propionate), or play a role in lipogenesis and gluconeogenesis [155]. Moreover, these bacterial metabolites have been shown to be a means to communicate with the host immune system, influencing the regulation of pro- and anti-inflammatory response mechanisms [156]. For most people consuming a Western style diet the carbohydrate fermentation is considered to mainly take place in the first, proximal, part of the colon, where most of the carbohydrates are depleted. Consequently, in the distal part of the colon, proteins and amino acids become the primary substrate for growth for the microbiota. Although the beneficial SCFAs can, to a certain extent, also be formed from amino acids and derivatives [157], such proteolytic fermentation processes tend to produce more (potentially) harmful metabolites such as amines, ammonia, p-cresol, phenols, and hydrogen sulfide. These proteolytic fermentation metabolites have indeed been shown to induce or enhance DNA damage, inflammation, leaky gut, and even cancer development in animal models [158]. Interestingly, an increased fiber or plant-based proportion in the diet can prevent such detrimental effects [158]. Therefore, the scientific basis for functional foods often aims to maintain or improve the saccharolytic fermentation profile throughout the length of the colon.

Currently dietary strategies to modulate the composition or the activity of the human gut microbiota, among which prebiotics and probiotics are probably the most well-established dietary compounds [159], are becoming increasingly popular. Prebiotics, which usually are non-digestible oligosaccharides (NDOs), are defined as "selectively fermented ingredients that result in specific changes in the composition and/or activity of the gastrointestinal microbiota, thus conferring benefit(s) upon host health" [160]. Examples of non-digestible oligosaccharides that are currently used in a wide range of food products are: inulin, resistant starch, galacto-oligosaccharides and fructo-oligosaccharides (FOS). Next to a direct impact on the microbial community, prebiotics and fibres have an impact on availability of other nutrients, water content, and viscosity, which have various physiological consequences within and beyond the GI tract (for a review see [161]). Even though such physiological factors in the host GI tract are often ignored in study designs, they can have a vast impact on the microbiota as was found for gut transit time which was strongly associated to various frequently used phylogenetic markers of the gut microbiota status [162]. Prebiotics have been shown to have pronounced and reproducible health effects in various animal models for colon cancer, cardiovascular disease, IBD, IBS, obesity, type 2 diabetes (T2D) and more [159]. However, data in human studies has not yet produced definitive results, which is partly due to scarce amount of (well-powered) clinical trials. Moreover, the effective doses in animal models, usually about 10% w/w of the diet, would translate to about 50 g of fiber in the human situation which a rather challenging dose to reach in Western countries [159]. On the other hand, the utilization of prebiotics targets the saccharolytic fermentation profile in the colon and gives us further clues of the complex ecological interactions within the complex gut ecosystem. Such insights help defining what a healthy microbiota profile should consist of, although with current knowledge this definition is far from complete.

Probiotics are the second type of influential dietary components and are actual microbial organisms themselves, which can be either part of normal food products such as yoghurts or fermented milk-drinks, or can be supplemented in capsules or powders. The FAO/WHO definition of a probiotic is "live microorganisms which when administered in adequate amounts confer a health benefit on the host" [163]. Probiotics present perhaps a more well-known and, with the general population, a more popular method to modulate the gut microbiota. On the market the main probiotics available are members of the lactic acid bacteria (e.g. *Lactobacillus*, *Streptococcus*, and *Bifidobacterium* spp.) but other microbes, such as *Bacillus clausii*, *Bacillus oligonitrophilis*, *Enterococcus faecium*, *Escherichia coli*, *Lactococcus lactis*, *Propionibacterium freudenreichii*, and the yeasts *Saccharomyces cerevisiae* and *Saccharomyces boulardii*, are employed as well [164]. Probiotics are hypothesized to have antibacterial effects, anti-inflammatory effects and/or the ability to alter the microbiota composition [165]. Moreover, probiotic bacteria have been shown to inhibit pathogen adhesion, probably due to competitive binding and stimulation of mucin production [166]. Many of the used probiotic strains produce antimicrobial substances like organic acids, bacteriocins, and diacetyl, however none of these substances have been reported to play key functions *in vivo* [167]. Despite this variety of molecular characteristics and mechanisms that have been postulated for the beneficial effects of probiotics [164], the exact mode of action remains to be determined as often in-depth characterization of the health promoting properties

of utilized probiotic strains is not performed [168]. The effectiveness of *a priori* characterization is shown by *Lactobacillus reuteri* NCIMB 30242, which was chosen for its bile salt-hydrolysing activity and indeed presented a reduction of the total and low density lipoprotein cholesterol in hypercholesterolaemic individuals upon ingestion [169]. It appears that consumption of probiotics does not show a tremendous (consistent) impact on the fecal microbiota composition in most studies [129, 170], although probiotic consumption has shown drastic impact on the small intestinal microbiota [171]. Moreover, a multi-species probiotic has shown that the community-wide functionality, measured by gene expression and metabolite profile, was affected in mice [172]. This change in functionality provides some insight on what the functioning of a healthy gut microbiota entails.

Another more radical gut microbiota modulation procedure, both in terms of scope and acceptance, is fecal microbiome transplantation (FMT). Until now this procedure, which replaces the whole gut microbiota of a patient by the fecal microbiota of a healthy donor, has mainly been used as a last resort in patients suffering from recurring otherwise unmanageable *Clostridium difficile* infections [173, 174]. Here it has been shown unequivocally that FMT outperformes conventional (antibiotic) therapy in the treatment of this disease [175]. Due to the extremely high success rate of this treatment that went up to approximately 90%, interest has been generated to include FMT in treatment of other chronic disease states such as inflammatory bowel diseases (first UC studies, see [176]), irritable bowel syndrome, and even metabolic syndrome conditions. Interestingly, while most gut microbiota studies are mainly correlative with respect to host health status, FMT has shown a causal relationship between glucose homeostasis and the microbiota composition as volunteers with metabolic syndrome that received a FMT experienced an improvement of their insulin sensitivity [177]. This improvement was accompanied by an overall microbiota diversity increase, a rise in the levels of bacteria related to *Roseburia intestinalis* (a butyrate producer), and an increase in fecal butyrate content [177], which is in line with the emerging picture regarding the gut microbiota and obesity associated metabolic diseases discussed above. Though the recipient's gut microbiota changes upon FMT, it usually does not completely reach the exact composition of the donor yet the recipient's health is impacted [173, 177]. Hence the question remains what microbes or microbial functions are exactly required for an individual to end up with a "healthy gut microbiome".

Healthy gut microbiome

Initial and successful pioneering studies on the human intestinal metagenome [99, 178, 179] were quickly followed by large-scale research initiatives, such as the European Metagenomics of the Human Intestinal Tract (MetaHIT) [9] and the Human Microbiome Project (HMP) [180], that have provided enormous amounts of metagenomic sequence data. Scientific efforts in the MetaHIT project concentrated on the gut microbiome and its associations with health, obesity and IBD states. Qin and coworkers published the first extensive fecal microbial gene set of 3.3 million non-redundant genes in 2010 [9], providing an unprecedented depth of functional insight in this microbial community. Mapping sequence reads per subject showed that the majority of the subjects shared approximately 40% of the genes of this catalogue, which

suggested the existence of a core metagenome [9]. Currently, the 3.3 million gene catalogue has received two substantial updates [30, 181] of which the last has increased the size to 9.8 million genes [30]. The HMP has generated an overview of the microbiota composition across 15 body sites, including the gut, from 242 healthy adults [182], as well as a set of approximately 800 partial reference genomes from human associated microbes [180]. Both large-scale initiatives have generated vast and complementary resources which pave the way for understanding the "healthy gut microbiome".

With the ever increasing awareness on the vast functional capacity residing within the microbiota and its possible roles in orchestrating human physiology through molecular and/or metabolic host microbiota interactions (for a review see [183]), interest in the gut microbiota is currently stretching across a variety of scientific fields. From every angle a basic question is continuously inquired: "What is the core of a healthy gut microbiome?". Moreover, can such a core microbiome be represented by one or more key bacterial members within our gut communities, or, in other words, does a core microbiota (composition) exist? Older, cultivation-dependent study results do in general suggest that the majority of healthy adults have several bacterial isolates in common. However, cultivation-independent studies have, as of yet, failed to find a consensus on a human gut microbiota core. The latter is partly due to different definitions of such a core microbiota and the phylogenetic level at which the core is defined, but also due to variations in both the study populations investigated (in terms of health status, age, geography, and diet) and the techniques employed to study the microbiota composition, and their substantial differences in depth of analysis [7, 22, 182, 184-187].

Besides variations in study design and methodologies, the extremely high subject specific community composition profiles at species level complicates the identification of a common core microbiota even further [188]. Hence, there appears to be only a small overlap in the gut microbiota composition between human adults but the often observed high temporal stability within individuals [7, 189, 190] does suggest that the gut microbiota forms core bacterial populations that exist at a subject specific level. However, even subject specific core microbiota compositions are susceptible to age, diet, and environmental factors such as antibiotic usage. Nevertheless, the existence of a subject-specific microbiota core could explain why human individuals display highly variable metabolic and gut microbiota composition responses to dietary interventions [18, 191, 192]. Although these responses appear poorly predictable, some consequences are known. For instance, the generally accepted bifidogenic effect of dietary FOS supplementation appears to be inversely correlated to the *Bifidobacterium* levels prior to the supplementation period, i.e. the lowest baseline *Bifidobacterium* levels having the highest increase upon FOS consumption [193, 194]. Furthermore, from a meta-analysis of three dietary interventions the baseline levels of several bacterial groups could predict the responsiveness of the gut microbiota [195]. Notably, these predictive bacteria mostly belonged to the Firmicutes phylum and their response was non-linear but showed a high association with the serum cholesterol read-outs [195]. Therefore, information of an individual's baseline gut microbiota could be key for successful nutritional and pharmaceutical treatments of metabolic disorders [191, 195, 196].

When the first 3.3 million non-redundant gut microbial gene catalogue was published by Qin and coworkers [9], the existence of a general core was also addressed. The metagenomic analysis compared data against 675 complete genomes of *Bacteria* and *Archaea* that were known at that time, of which 194 genomes were of microorganisms originating from the intestinal tract. Depending on bacterial core prevalence definitions (percentage of subjects in which bacteria were detected) 24 to 51 species were found to represent a common core using this metagenomic approach (Table 1.5) in combination with 100 and 90% prevalence in the study-population, respectively. The same samples used in that study were also subjected to microbiota composition analysis by the HITChip [6] and this resulted in the detection of 83 and 135 common core species (only analyzing signals from representatives of cultivated bacterial isolates), with 100 and 90% prevalence in the study-population, respectively (Table 1.5). Because many of the genomes of these cultured isolates are not yet complete or not determined to date, they are absent from the 650 complete genomes used in the metagenomic analysis. The latter discrepancy explains why the HITChip analysis identifies a larger common core microbiota compared to the metagenomic genome mapping approach (Table 1.5). The 135 species detected in more than 90% of the subjects, cover 49 genus-like groups (Figure 1.2) and displayed a large overlap with the species that were detected by metagenomic mapping, i.e. >86% of the species detected through metagenomic mapping that are also represented by specific probes on the HITChip were indeed detected by the latter platform (Table 1.5). Additionally, in the subjects used for the 3.3 million gene catalogue 197 uncultured species-like groups can be detected by HITChip analysis in more than 90% of the subjects, next to the 135 common species-like groups that are cultured (Figure 1.2).

Table 1.5 Core bacterial species as found with metagenomic mapping (with a genome coverage of >1%) and HITChip analysis in the samples that were used for the construction of the 3.3 million gut gene catalogue [9]. In this comparison only the HITChip data representing the 329 cultured isolates that can be detected by the array were evaluated.

Core bacterial species	All subjects	>90% of subjects
Core species detected by metagenomic-mapping of the 3.3 M gene catalogue	24	51
Core species detected by HITChip	83	135
Core species detected by both metagenomic-mapping and HITChip	14 (87.5 %)	33 (86.8%)
(Number of core species detected by metagenomic mapping that are represented by the HITChip probe sets)	(16)	(38)

Previous studies underpin that although microbiota compositional profiles may be highly diverse among individuals, the corresponding functional gene repertoire of these communities is quite similar [22]. Such a microbial gene repertoire conservation evaluated at the level of COG categories, is essentially a reflection of functional redundancy within different bacteria, which is commonly observed in various habitats [197]. The latter is not surprising considering the fact that all essential household genes, such as DNA replication and repair, should be present in every living organism. Additionally in different subjects one can expect that despite the phylogenetic differences of their microbiota composition, the microbial community will still exert largely the

same functionality within the large intestinal niche. These observations question the biological importance of a common microbial core at a phylogenetic level (a microbiota core) and argue that a common microbial core exists at the gene repertoire level (a metagenomic core). After

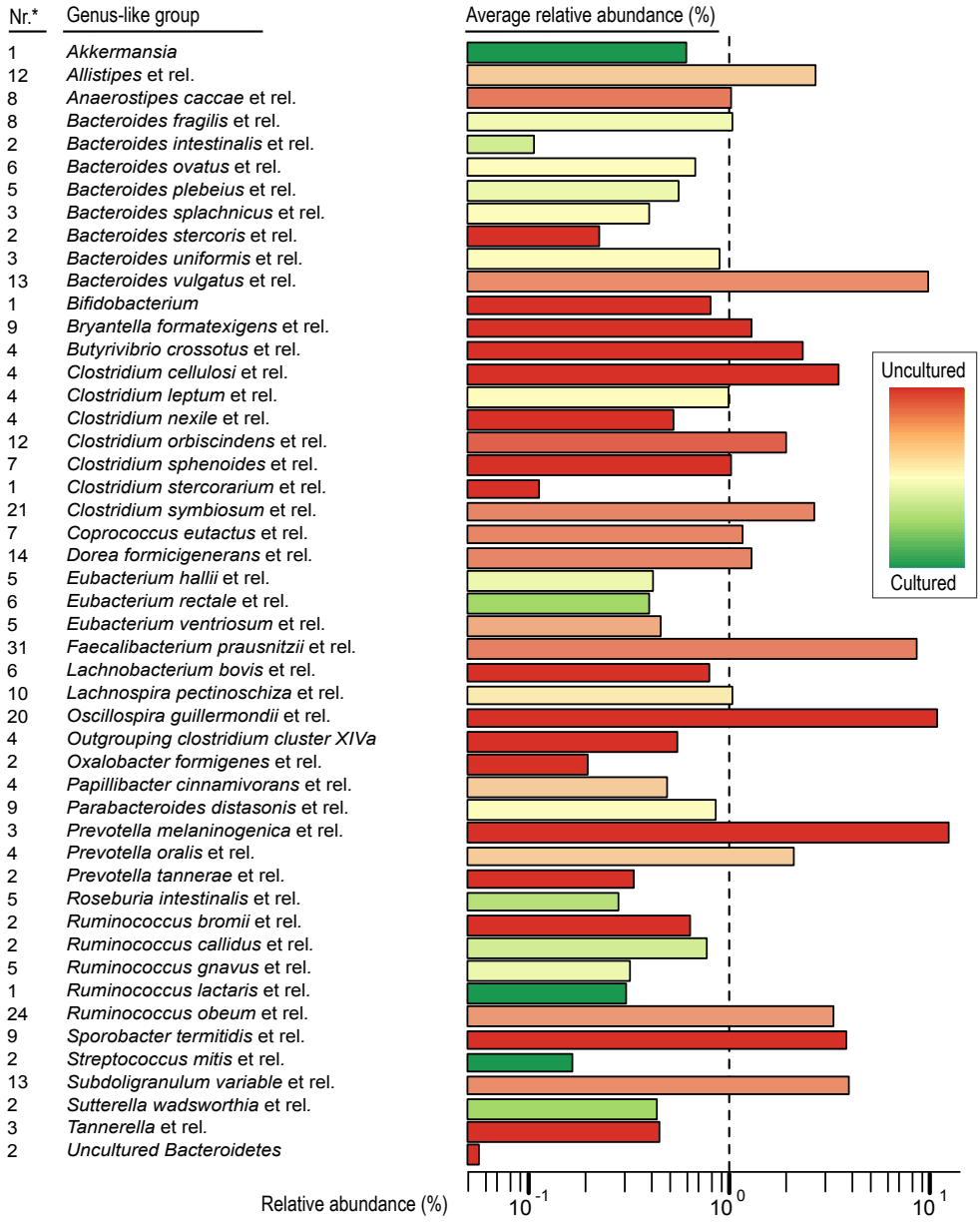


Figure 1.2 Number of core species-like groups and the average relative abundance (in %) of their corresponding 49 genus-like groups that are > 90% prevalent in the study-population (as deduced from HITChip profiling) used for the construction of the 3.3M bacterial gene catalogue. Bar colors represent the ratio of the core species-like groups that have been cultured (green) versus uncultured isolates (brown) within each of their corresponding genus-like groups. *) Number of species-like groups for each genus-like group.

the establishment of metagenomics, which captures DNA of both cultured and uncultured isolates, some have argued that microbial ecology no longer needed the traditional laboratory approaches like cultivation to complete our knowledge of the gut microbiota [198]. Nonetheless, metagenomics has its pitfalls and limitations which can actually only be solved or complemented with traditional, and cultivation based approaches. Moreover, our current understanding of the microbiome functionality is far from complete, which is largely due to the fact that the most prevalent functionalities that are specific to the gut environment are unknown, and the biggest limitation is that significant amounts of metagenomic sequence data (approximately 60% [9]) have no close match in reference databases, which are predominantly filled by sequences from cultured isolates, leaving these metagenomics sequences without (functional) annotation. Therefore, it may be too early to draw conclusions from metagenomics derived functional profiles due to the incompleteness in functional knowledge of the genes encoded by the microbiota. Besides being incomplete due to the lack of genetic information from uncultured isolates, this information gap is also a reflection of the limited understanding and characterization of the (anaerobic) metabolic processes and pathways executed by the microbial community members. The latter is particularly relevant considering the fact that while trying to assess a phylogenetic core, as discussed above, the numbers of uncultured species-like groups outnumber the cultured representatives by at least 2-3 fold (Figure 1.2). Furthermore, various bacteria-bacteria or bacteria-host interactions can still only be discovered in laboratory setting [199, 200], as they could not (yet) be predicted with the current algorithms or systems biology approaches. Nevertheless, metagenomics has significantly improved our knowledge of the gut microbiota [201] and various connections have been made with factors shaping the microbiota, such as diet, health status of the host, and age. Yet the fundamental basis of these advancement stems from the knowledge gained from reference databases containing cultivation and characterization studies. Hence, it would be recommendable that (ideal) future studies employ both sequence-based technologies and cultivation methodologies to complement each other and further advance our insight into the gut eco-system [202].

Thesis outline

From the various factors that can shape the intimate cross-talk between the host and its gut microbiota, the host genotype is probably the most stable factor. **Chapter 2** aims to describe the influence of the host genotype and shows why the genotype should still be taken into consideration even though environmental factors, like diet, may have a more noticeable effect on the microbial gut community. Next, **Chapter 3** both illustrates the power of the host genotype as well as addressing the link between the host energy homeostasis and the gut microbiota in a cohort of monozygotic twins, among which half of the twin pairs are discordant for their Body Mass Index (BMI). The global microbiota composition was studied by using a phylogenetic microarray and the results of this approach in the twin cohort has led to the hypothesis that several bacterial groups may vary in abundance along with the host BMI. Nevertheless, the profiles of the members within these taxonomic groups were surprisingly conserved, likely due to the host genotype.

The second part of this thesis aims to map the modulation of the gut microbiota that can be brought about by different types of dietary intervention. Treatment of Irritable Bowel Syndrome (IBS), a globally prevalent and chronic disorder, is shifting its focus to modulation of the gut microbiota. Although several trials with probiotic treatment have not shown a consistent success rate, a multi-species probiotic consisting of *Lactobacillus rhamnosus* GG, *Lactobacillus rhamnosus* Lc705, *Propionibacterium freudenreichi* PAJ, and *Bifidobacterium animalis* BB12 was successful in reducing IBS symptoms. In **Chapter 4** phylogenetic microarray analysis of the gut microbiota of the participants of this multi-species probiotic trial is reported. No consistent drastic shifts in the bacterial composition were observed, yet subtle changes measurable in the co-occurrences between bacterial taxa, were exclusively identified in the treatment group, which suggests the treatment induced a re-alignment or re-structuring of the microbiota. Modulation of the gut microbiota with the prebiotic FOS showed both drastic shift in microbiota composition as well as structural changes in the bacterial co-occurrence networks as studied by phylogenetic microarray analysis, which is described in **Chapter 5**. Interestingly, knowledge on cultured representatives of the taxa involved in the co-occurrence changes due to FOS consumption raises a new hypothesis regarding butyrate, an important short chain fatty acid for host physiology, production routes in the gut.

The third part of this thesis combines molecular or sequence-based technologies with cultivation and characterization approaches in a laboratory setting. **Chapter 6** shows how combining two complementary metagenomic approaches, i.e. phylogenetic microarray analysis and untargeted shotgun metagenomics, could guide development of media for future isolation and cultivation attempts, as well as targeting various uncultured organisms based on their remarkable genetic potential. **Chapter 7** provides a novel targeted approach to mine metagenomic gene catalogues and harvest genes of interest (or domains of genes) to characterize their predicted functionality as well. This targeted approach was validated with fibronectin binding domains.

Finally, **Chapter 8** summarizes the research described in this work and its contribution to current knowledge and understanding of gut microbiota. Further integration and future improvement of the approaches presented here is provided as well.

References

1. **McGovern PE, Zhang J, Tang J, et al.** Fermented beverages of pre- and proto-historic China. Proc Natl Acad Sci U S A 2004;101(51):17593-8.
2. **Fleming A.** On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. Bull World Health Organ 2001;79(8):780-90.
3. **Escherich T.** Die darmbakterien des neugeborenen und sauglings. Fortschr. Med. 1885;3:515-522.
4. **Metchnikoff E.** The prolongation of life; optimistic studies. 1907. London: Heinemann.
5. **Zoetendal EG, Cheng B, Koike S, et al.** Molecular microbial ecology of the gastrointestinal tract: from phylogeny to function. Curr Issues Intest Microbiol 2004;5(2):31-47.
6. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
7. **Jalanka-Tuovinen J, Salonen A, Nikkila J, et al.** Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One 2011;6(7):e23035.
8. **Rajilic-Stojanovic M and de Vos WM.** The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 2014;38(5):996-1047.
9. **Qin J, Li R, Raes J, et al.** A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59-65.
10. **DiGiulio DB, Romero R, Amogan HP, et al.** Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: a molecular and culture-based investigation. PLoS One 2008;3(8):e3056.
11. **Mackie RI, Sghir A, and Gaskins HR.** Developmental microbial ecology of the neonatal gastrointestinal tract. The American Journal of Clinical Nutrition 1999;69(5):1035s-1045s.
12. **Hooper LV, Xu J, Falk PG, et al.** A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. Proc. Natl. Acad. Sci. USA 1999;96(17):9833-9838.
13. **Lewis ZT, Totten SM, Smilowitz JT, et al.** Maternal fucosyltransferase 2 status affects the gut bifidobacterial communities of breastfed infants. Microbiome 2015;3:13.
14. **Kabeerdoss J, Devi RS, Mary RR, et al.** Faecal microbiota composition in vegetarians: comparison with omnivores in a cohort of young women in southern India. Br J Nutr 2012;108(6):953-7.
15. **Zimmer J, Lange B, Frick JS, et al.** A vegan or vegetarian diet substantially alters the human colonic faecal microbiota. Eur J Clin Nutr 2012;66(1):53-60.
16. **David LA, Maurice CF, Carmody RN, et al.** Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505(7484):559-563.
17. **O'Keefe SJ, Li JV, Lahti L, et al.** Fat, fibre and cancer risk in African Americans and rural Africans. Nat Commun 2015;6:6342.
18. **Walker AW, Ince J, Duncan SH, et al.** Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J 2011;5(2):220-30.
19. **Wu GD, Chen J, Hoffmann C, et al.** Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science 2011.
20. **Zoetendal EG, Akkermans ADL, Akkermans-van Vliet WM, et al.** The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. Microbial Ecology in Health and Disease 2001;13(3):129-134.
21. **Stewart JA, Chadwick VS, and Murray A.** Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. J Med Microbiol 2005;54(12):1239-1242.
22. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-484.
23. **Zhang C, Zhang M, Wang S, et al.** Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. ISME J 2009;4(2):232-241.
24. **Petnicki-Ocwieja T, Hrncir T, Liu Y-J, et al.** Nod2 is required for the regulation of commensal microbiota in the intestine. Proceedings of the National Academy of Sciences of the United States of America 2009;106(37):15813-15818.
25. **Wacklin P, Tuimala J, Nikkila J, et al.** Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. PLoS One 2014;9(4):e94863.
26. **Khachatryan ZA, Ktsoyan ZA, Manukyan GP, et al.** Predominant role of host genetics in controlling the composition of gut microbiota. PLoS One 2008;3(8):e3064.
27. **Frank DN, Robertson CE, Hamm CM, et al.** Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflamm Bowel Dis 2011;17(1):179-84.
28. **Marchesi JR and Ravel J.** The vocabulary of microbiome research: a proposal. Microbiome 2015;3(1):1-3.
29. **Savage DC.** Microbial ecology of the gastrointestinal tract. Ann. Rev. Microbiol. 1977;31:107-133.

30. **Li J, Jia H, Cai X, et al.** An integrated catalog of reference genes in the human gut microbiome. Nat Biotech 2014;32(8):834-841.
31. **Ritari J, Salojarvi J, Lahti L, et al.** Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 2015;16(1):1056.
32. **Suau A, Bonnet R, Sutren M, et al.** Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. Appl. Environ. Microbiol. 1999;65(11):4799-4807.
33. **Sokol H and Seksik P.** The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. Curr Opin Gastroenterol 2010;26(4):327-31.
34. **Tannock GW.** Molecular assessment of intestinal microflora. The American Journal of Clinical Nutrition 2001;73(2):410s-414s.
35. **Lagier JC, Armougom F, Million M, et al.** Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 2012;18(12):1185-93.
36. **Stewart EJ.** Growing unculturable bacteria. J Bacteriol 2012;194(16):4151-60.
37. **Zoetendal EG, Akkermans AD, and de Vos WM.** Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. Appl. Environ. Microbiol. 1998;64(10):3854-3859.
38. **Schloss PD and Handelsman J.** Status of the microbial census. Microbiol. Mol. Biol. Rev. 2004;68(686-691).
39. **Rajilić-Stojanović M, Smidt H, and de Vos WM.** Diversity of the human gastrointestinal tract microbiota revisited. Environ. Microbiol. 2007;9:2125-2136.
40. **DeSantis TZ, Brodie EL, Moberg JP, et al.** High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb. Ecol. 2007;53(3):371-383.
41. **Bodrossy L and Sessitsch A.** Oligonucleotide microarrays in microbial diagnostics. Curr Opin Microbiol 2004;7(3):245-54.
42. **Guschin DY, Mobarry BK, Proudnikov D, et al.** Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Appl. Environ. Microbiol. 1997;63(6):2397-2402.
43. **Gentry TJ, Wickham GS, Schadt CW, et al.** Microarray applications in microbial ecology research. Microb. Ecol. 2006;52(2):159-175.
44. **Wagner M, Smidt H, Loy A, et al.** Unravelling microbial communities with DNA-microarrays: challenges and future directions. Microb. Ecol. 2007;53(3):498-506.
45. **Hazen TC, Dubinsky EA, DeSantis TZ, et al.** Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. Science 2010;330(6001):204-208.
46. **Paliy O, Kenche H, Abernathy F, et al.** High-Throughput Quantitative Analysis of the Human Intestinal Microbiota with a Phylogenetic Microarray. Applied and Environmental Microbiology 2009;75(11):3572-3579.
47. **Tottey W, Denonfoux J, Jaziri F, et al.** The Human Gut Chip "HuGChip", an Explorative Phylogenetic Microarray for Determining Gut Microbiome Diversity at Family Level. PLoS ONE 2013;8(5):e62544.
48. **Rajilic-Stojanovic M, Heilig H, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
49. **Frank DN and Pace NR.** Gastrointestinal microbiology enters the metagenomics era. Curr Opin Gastroenterol 2008;24(1):4-10.
50. **Caporaso JG, Lauber CL, Walters WA, et al.** Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 2011;108 Suppl 1:4516-22.
51. **Mardis ER.** Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 2013;6:287-303.
52. **Margulies M, Egholm M, Altman WE, et al.** Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437(7057):376-80.
53. **Sanger F, Nicklen S, and Coulson AR.** DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74(12):5463-7.
54. **Bentley DR, Balasubramanian S, Swerdlow HP, et al.** Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456(7218):53-9.
55. **Peckham HE, McLaughlin SF, Ni JN, et al.** SOLiD(TM) Sequencing and 2-Base Encoding. in American Society of Human Genetics. 2007. San Diego.
56. **Mardis ER.** Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics 2008;9(1):387-402.
57. **Hodkinson BP and Grice EA.** Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Advances in Wound Care 2015;4(1):50-58.
58. **Kuczynski J, Lauber CL, Walters WA, et al.** Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 2012;13(1):47-58.

59. **Bikel S, Valdez-Lara A, Cornejo-Granados F, et al.** Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Comput Struct Biotechnol J 2015;13:390-401.
60. **Woese CR and Fox GE.** Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 1977;74(11):5088-90.
61. **Thomas T, Gilbert J, and Meyer F.** Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation 2012;2:3-3.
62. **Langille MGI, Zaneveld J, Caporaso JG, et al.** Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotech 2013;31(9):814-821.
63. **Handelsman J, Rondon MR, Brady SF, et al.** Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 1998;5(10):R245-R249.
64. **Simon C and Daniel R.** Metagenomic analyses: past and future trends. Appl Environ Microbiol 2011;77(4):1153-61.
65. **Uchiyama T and Miyazaki K.** Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol 2009;20(6):616-22.
66. **Schwarz S, Waschkowitz T, and Daniel R.** Enhancement of gene detection frequencies by combining DNA-based stable-isotope probing with the construction of metagenomic DNA libraries. World Journal of Microbiology and Biotechnology 2006;22(4):363-368.
67. **Warren RL, Freeman JD, Levesque RC, et al.** Transcription of foreign DNA in Escherichia coli. Genome Res 2008;18(11):1798-805.
68. **Yang X, Xie L, Li Y, et al.** More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body. PLoS One 2009;4(6):e6074.
69. **Wei C and Brent MR.** Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 2006;7:327.
70. **O'Hara AM and Shanahan F.** The gut flora as a forgotten organ. EMBO Rep. 2006;7(7):688-693.
71. **Cummings JH and Macfarlane GT.** Colonic microflora: Nutrition and health. Nutrition 1997;13(5):476-478.
72. **Metges CC.** Contribution of microbial amino acids to amino acid homeostasis of the host. J. Nutr. 2000;130(7):1857S-1864.
73. **Ramotar K, Conly JM, Chubb H, et al.** Production of menaquinones by intestinal anaerobes. J. Infect. Dis. 1984;150(2):213-218.
74. **Albert MJ, Mathan VI, and Baker SJ.** Vitamin B12 synthesis by human small intestinal bacteria. Nature 1980;238(5749):781-782.
75. **Begley M, Gahan CG, and Hill C.** The interaction between bacteria and bile. FEMS Microbiol Rev 2005;29(4):625-51.
76. **Adlercreutz H, Pulkkinen MO, Hamalainen EK, et al.** Studies on the role of intestinal bacteria in metabolism of synthetic and natural steroid hormones. J. Steroid. Biochem. 1984;20(1):217-229.
77. **Cebra JJ.** Influences of microbiota on intestinal immune system development. Am. J. Clin. Nutr. 1999;69(5):1046S-1051.
78. **Shanahan F.** The host-microbe interface within the gut. Best Practice & Research Clinical Gastroenterology 2002;16(6):915-931.
79. **O'Hara AM, O'Regan P, Fanning A, et al.** Functional modulation of human intestinal epithelial cell responses by Bifidobacterium infantis and Lactobacillus salivarius. Immunology 2006;118(2):202-15.
80. **Seksik P, Rigottier-Gois L, Gramet G, et al.** Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. Gut 2003;52(2):237-42.
81. **Borre YE, O'Keeffe GW, Clarke G, et al.** Microbiota and neurodevelopmental windows: implications for brain disorders. Trends in Molecular Medicine 2014;(0).
82. **Scheperjans F, Aho V, Pereira PA, et al.** Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov Disord 2015;30(3):350-8.
83. **Backhed F, Ding H, Wang T, et al.** The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 2004;101(44):15718-23.
84. **Backhed F, Manchester JK, Semenkovich CF, et al.** Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U S A 2007;104(3):979-84.
85. **Ley RE, Turnbaugh PJ, Klein S, et al.** Microbial ecology: Human gut microbes associated with obesity. Nature 2006;444:1022-1023.
86. **Samuel BS, Shaito A, Motoike T, et al.** Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. Proc Natl Acad Sci U S A 2008;105(43):16767-72.

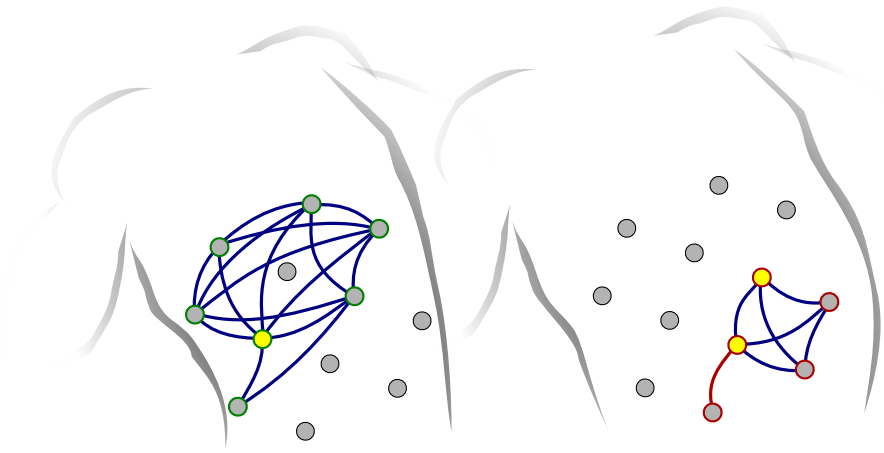
87. **Turnbaugh PJ, Ley RE, Mahowald MA, et al.** An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 2006;444(7122):1027-31.
88. **Harley IT and Karp CL.** Obesity and the gut microbiome: Striving for causality. Mol Metab 2012;1(1-2):21-31.
89. **Arora T, Sharma R, and Frost G.** Propionate. Anti-obesity and satiety enhancing factor? Appetite 2011;56(2):511-5.
90. **Guinane CM and Cotter PD.** Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. Therap Adv Gastroenterol 2013;6(4):295-308.
91. **Duncan SH, Belenguer A, Holtrop G, et al.** Reduced Dietary Intake of Carbohydrates by Obese Subjects Results in Decreased Concentrations of Butyrate and Butyrate-Producing Bacteria in Feces. Applied and Environmental Microbiology 2007;73(4):1073-1078.
92. **Schwiertz A, Taras D, Schafer K, et al.** Microbiota and SCFA in Lean and Overweight Healthy Subjects. Obesity (Silver Spring) 2009.
93. **Zhang H, DiBaise JK, Zuccolo A, et al.** Human gut microbiota in obesity and after gastric bypass. Proc. Natl. Acad. Sci. U S A. 2009;106(7):2365-2370.
94. **Le Chatelier E, Nielsen T, Qin J, et al.** Richness of human gut microbiome correlates with metabolic markers. Nature 2013;500(7464):541-6.
95. **Qin J, Li Y, Cai Z, et al.** A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55-60.
96. **Larsen N, Vogensen FK, van den Berg FWJ, et al.** Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. PLoS ONE 2010;5(2):e9085.
97. **Darfeuille-Michaud A, Boudeau J, Bulois P, et al.** High prevalence of adherent-invasive Escherichia coli associated with ileal mucosa in Crohn's disease. Gastroenterology 2004;127(2):412-421.
98. **Li Q, Wang C, Tang C, et al.** Molecular-Phylogenetic Characterization of the Microbiota in Ulcerated and Non-Ulcerated Regions in the Patients with Crohn's Disease. PLoS ONE 2012;7(4):e34939.
99. **Manichanh C, Rigottier-Gois L, Bonnaud E, et al.** Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006;55(2):205-11.
100. **Sokol H, Seksik P, Rigottier-Gois L, et al.** Specificities of the fecal microbiota in inflammatory bowel disease. Inflamm. Bowel Dis. 2006;12(2):106-111.
101. **Garrett WS, Gallini CA, Yatsunenko T, et al.** Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. Cell host & microbe 2010;8(3):292-300.
102. **Frank DN, St Amand AL, Feldman RA, et al.** Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc. Natl. Acad. Sci. U S A. 2007;104(34):13780-13785.
103. **Martinez-Medina M, Aldeguer X, Gonzalez-Huix F, et al.** Abnormal microbiota composition in the ileocolonic mucosa of Crohn's disease patients as revealed by polymerase chain reaction-denaturing gradient gel electrophoresis. Inflammatory Bowel Diseases 2006;12(12):1136-1145.
104. **Sokol H, Seksik P, Furet JP, et al.** Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflamm. Bowel Dis. 2009;15(8):1183-1189.
105. **Morgan XC, Tickle TL, Sokol H, et al.** Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology 2012;13(9):R79-R79.
106. **Jeffery IB, O'Toole PW, Ohman L, et al.** An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. Gut 2012;61(7):997-1006.
107. **Rajilić-Stojanović M, Biagi E, Heilig HGHJ, et al.** Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology 2011;141(5):1792-1801.
108. **Saulnier DM, Riehle K, Mistretta TA, et al.** Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology 2011;141(5):1782-91.
109. **Castellarin M, Warren RL, Freeman JD, et al.** Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Research 2012;22(2):299-306.
110. **Kostic AD, Gevers D, Pedamallu CS, et al.** Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Research 2012;22(2):292-298.
111. **McCoy AN, Araújo-Pérez F, Azcárate-Peril A, et al.** Fusobacterium Is Associated with Colorectal Adenomas. PLoS ONE 2013;8(1):e53653.
112. **Arthur JC, Perez-Chanona E, Mühlbauer M, et al.** Intestinal inflammation targets cancer-inducing activity of the microbiota. Science (New York, N.Y.) 2012;338(6103):120-123.
113. **Ley RE, Backhed F, Turnbaugh P, et al.** Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 2005;102(31):11070-5.
114. **Claesson MJ, Jeffery IB, Conde S, et al.** Gut microbiota composition correlates with diet and health in the elderly. Nature 2012;488(7410):178-84.

115. **Romero-Corral A, Somers VK, Sierra-Johnson J, et al.** Accuracy of body mass index in diagnosing obesity in the adult general population. Int J Obes 2008;32(6):959-966.
116. **Blüher M.** The distinction of metabolically 'healthy' from 'unhealthy' obese individuals. Current Opinion in Lipidology 2010;21(1):38-43.
117. **Zupancic ML, Cantarel BL, Liu Z, et al.** Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. PLoS ONE 2012;7(8):e43052.
118. **Cotillard A, Kennedy SP, Kong LC, et al.** Dietary intervention impact on gut microbial gene richness. Nature 2013;500(7464):585-588.
119. **Karlsson FH, Tremaroli V, Nookaew I, et al.** Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013;498(7452):99-103.
120. **Walker AW, Sanderson JD, Churcher C, et al.** High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. BMC Microbiology 2011;11(1):1-12.
121. **Swidsinski A, Weber J, Loening-Baucke V, et al.** Spatial Organization and Composition of the Mucosal Flora in Patients with Inflammatory Bowel Disease. Journal of Clinical Microbiology 2005;43(7):3380-3389.
122. **Peterson DA, Frank DN, Pace NR, et al.** Metagenomic Approaches for Defining the Pathogenesis of Inflammatory Bowel Diseases. Cell Host & Microbe 2008;3(6):417-427.
123. **Png CW, Linden SK, Gilshenan KS, et al.** Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. Am J Gastroenterol 2010;105(11):2420-2428.
124. **Rajilic-Stojanovic M, Shanahan F, Guarner F, et al.** Phylogenetic analysis of dysbiosis in ulcerative colitis during remission. Inflamm Bowel Dis 2013;19(3):481-8.
125. **Mukhopadhya I, Thomson JM, Hansen R, et al.** Detection of <italic>Campylobacter concisus</italic> and Other <italic>Campylobacter</italic> Species in Colonic Biopsies from Adults with Ulcerative Colitis. PLoS ONE 2011;6(6):e21490.
126. **Lupp C, Robertson ML, Wickham ME, et al.** Host-Mediated Inflammation Disrupts the Intestinal Microbiota and Promotes the Overgrowth of Enterobacteriaceae. Cell Host & Microbe 2007;2(3):204.
127. **Brint EK, MacSharry J, Fanning A, et al.** Differential Expression of Toll-Like Receptors in Patients With Irritable Bowel Syndrome. Am J Gastroenterol 2011;106(2):329-336.
128. **Pimentel M, Lembo A, Chey WD, et al.** Rifaximin therapy for patients with irritable bowel syndrome without constipation. N Engl J Med 2011;364(1):22-32.
129. **Spiller R.** Review article: probiotics and prebiotics in irritable bowel syndrome. Aliment Pharmacol Ther 2008;28(4):385-96.
130. **Zoetendal EG, Rajilic-Stojanovic M, and de Vos WM.** High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. Gut 2008;57(11):1605-15.
131. **Whelan K.** Probiotics and prebiotics in the management of irritable bowel syndrome: a review of recent clinical trials and systematic reviews. Current Opinion in Clinical Nutrition & Metabolic Care 2011;14(6):581-587.
132. **Pimentel M and Lezcano S.** Irritable Bowel Syndrome: Bacterial Overgrowth--What's Known and What to Do. Curr Treat Options Gastroenterol 2007;10(4):328-37.
133. **Salonen A, de Vos WM, and Palva A.** Gastrointestinal microbiota in irritable bowel syndrome: present state and perspectives. Microbiology 2010;156(Pt 11):3205-15.
134. **Spiller R and Garsed K.** Postinfectious irritable bowel syndrome. Gastroenterology 2009;136(6):1979-88.
135. **Jalanka-Tuovinen J, Salojärvi J, Salonen A, et al.** Faecal microbiota composition and host–microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. Gut 2014;63(11):1737-1745.
136. **Jalanka J, Salonen A, Fuentes S, et al.** Microbial signatures in post-infectious irritable bowel syndrome - toward patient stratification for improved diagnostics and treatment. Gut Microbes 2015;6(6):364-9.
137. **Carroll IM, Ringel-Kulka T, Keku TO, et al.** Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. Am J Physiol Gastrointest Liver Physiol 2011;301(5):G799-807.
138. **Noor SO, Ridgway K, Scovell L, et al.** Ulcerative colitis and irritable bowel patients exhibit distinct abnormalities of the gut microbiota. BMC Gastroenterol 2010;10:134.
139. **Rajilic-Stojanovic M, Biagi E, Heilig HG, et al.** Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology 2011;141(5):1792-801.
140. **Rigsbee L, Agans R, Shankar V, et al.** Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable bowel syndrome. Am J Gastroenterol 2012;107(11):1740-51.
141. **Simren M, Barbara G, Flint HJ, et al.** Intestinal microbiota in functional bowel disorders: a Rome foundation report. Gut 2013;62(1):159-76.

142. **Lyra A, Rinttila T, Nikkila J, et al.** Diarrhoea-predominant irritable bowel syndrome distinguishable by 16S rRNA gene phylotype quantification. World J Gastroenterol 2009;15(47):5936-45.
143. **Malinen E, Rinttila T, Kajander K, et al.** Analysis of the fecal microbiota of irritable bowel syndrome patients and healthy controls with real-time PCR. Am J Gastroenterol 2005;100(2):373-82.
144. **Krogius-Kurikka L, Lyra A, Malinen E, et al.** Microbial community analysis reveals high level phylogenetic alterations in the overall gastrointestinal microbiota of diarrhoea-predominant irritable bowel syndrome sufferers. BMC Gastroenterology 2009;9(1):95.
145. **Ou J, Carbonero F, Zoetendal EG, et al.** Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. The American Journal of Clinical Nutrition 2013;98(1):111-120.
146. **Marchesi JR, Dutilh BE, Hall N, et al.** Towards the Human Colorectal Cancer Microbiome. PLoS ONE 2011;6(5):e20447.
147. **Marchesi JR, Adams DH, Fava F, et al.** The gut microbiota and host health: a new clinical frontier. Gut 2015.
148. **Wu N, Yang X, Zhang R, et al.** Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients. Microbial Ecology 2013;66(2):462-470.
149. **Zhu Q, Jin Z, Wu W, et al.** Analysis of the Intestinal Lumen Microbiota in an Animal Model of Colorectal Cancer. PLoS ONE 2014;9(3):e90849.
150. **Carbonero F, Benefiel AC, Alizadeh-Ghamsari AH, et al.** Microbial pathways in colonic sulfur metabolism and links with health and disease. Frontiers in Physiology 2012;3.
151. **Ijssennagger N, Belzer C, Hooiveld GJ, et al.** Gut microbiota facilitates dietary heme-induced epithelial hyperproliferation by opening the mucus barrier in colon. Proc Natl Acad Sci U S A 2015;112(32):10038-43.
152. **Kashyap PC, Marcobal A, Ursell LK, et al.** Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. Proceedings of the National Academy of Sciences 2013;110(42):17059-17064.
153. **Flint HJ, Scott KP, Louis P, et al.** The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol 2012;9(10):577-589.
154. **Ze X, Le Mougen F, Duncan SH, et al.** Some are more equal than others. Gut Microbes 2013;4(3):236-240.
155. **Macfarlane GT and Macfarlane S.** Fermentation in the Human Large Intestine: Its Physiologic Consequences and the Potential Contribution of Prebiotics. Journal of Clinical Gastroenterology 2011;45:S120-S127.
156. **Arpaia N, Campbell C, Fan X, et al.** Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. Nature 2013;504(7480):451-455.
157. **Bui TP, Ritari J, Boeren S, et al.** Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. Nat Commun 2015;6:10062.
158. **Windey K, De Preter V, and Verbeke K.** Relevance of protein fermentation to gut health. Molecular Nutrition & Food Research 2012;56(1):184-196.
159. **Tuohy KM and Del Rio D.** Diet-microbe interactions in the gut: effects on human health and disease. 2014. Amsterdam Elsevier.
160. **Gibson GR, Scott KP, Rastall RA, et al.** Dietary prebiotics: current status and new definition. Food Science & Technology Bulletin: Functional Foods 2010;7(1):1-19.
161. **Tungland BC and Meyer D.** Nondigestible Oligo- and Polysaccharides (Dietary Fiber): Their Physiology and Role in Human Health and Food. Comprehensive Reviews in Food Science and Food Safety 2002;1(3):90-109.
162. **Vandeputte D, Falony G, Vieira-Silva S, et al.** Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut 2015.
163. **FAO/WHO.** Probiotics in Food. Health and nutritional properties and guidelines for evaluation. Report of a joint FAO/WHO expert consultation on evaluation of health and nutritional properties of probiotics in food including powder milk with live lactic acid bacteria. . 2006. Rome.
164. **Penner R, Fedorak RN, and Madsen KL.** Probiotics and nutraceuticals: non-medicinal treatments of gastrointestinal diseases. Curr Opin Pharmacol 2005;5(6):596-603.
165. **O'Mahony L, McCarthy J, Kelly P, et al.** Lactobacillus and bifidobacterium in irritable bowel syndrome: Symptom responses and relationship to cytokine profiles. Gastroenterology 2005;128(3):541-551.
166. **Mack DR, Michail S, Wei S, et al.** Probiotics inhibit enteropathogenic E. coli adherence in vitro by inducing intestinal mucin gene expression. Am J Physiol 1999;276(4 Pt 1):G941-50.
167. **Rastall RA, Gibson GR, Gill HS, et al.** Modulation of the microbial ecology of the human colon by probiotics, prebiotics and synbiotics to enhance human health: An overview of enabling science and potential applications. FEMS Microbiology Ecology 2005;52(2):145-152.
168. **Bron PA, van Baarlen P, and Kleerebezem M.** Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. Nat Rev Micro 2012;10(1):66-78.

169. **Jones ML, Martoni CJ, and Prakash S.** Cholesterol lowering and inhibition of sterol absorption by Lactobacillus reuteri NCIMB 30242: a randomized controlled trial. Eur J Clin Nutr 2012;66(11):1234-1241.
170. **Lahti L, Salonen A, Kekkonen RA, et al.** Associations between the human intestinal microbiota, Lactobacillus rhamnosus GG and serum lipids indicated by integrated analysis of high-throughput profiling data. PeerJ 2013;1:e32.
171. **El Aidy S, van den Bogert B, and Kleerebezem M.** The small intestine microbiota, nutritional modulation and relevance for health. Current Opinion in Biotechnology 2015;32:14-20.
172. **McNulty NP, Yatsunenko T, Hsiao A, et al.** The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. Science Translational Medicine 2011;3(106):106ra106-106ra106.
173. **Fuentes S, van Nood E, Tims S, et al.** Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. International Society for Microbial Ecology Journal 2014;8(8):1621-33.
174. **Bakken JS.** Fecal bacteriotherapy for recurrent Clostridium difficile infection. Anaerobe 2009;15(6):285-289.
175. **van Nood E, Vrieze A, Nieuwdorp M, et al.** Duodenal infusion of donor feces for recurrent Clostridium difficile. N Engl J Med 2013;368(5):407-15.
176. **Rossen NG, Fuentes S, van der Spek MJ, et al.** Findings From a Randomized Controlled Trial of Fecal Transplantation for Patients With Ulcerative Colitis. Gastroenterology 2015;149(1):110-118.e4.
177. **Vrieze A, Van Nood E, Holleman F, et al.** Transfer of Intestinal Microbiota From Lean Donors Increases Insulin Sensitivity in Individuals With Metabolic Syndrome. Gastroenterology 2012;143(4):913-916.e7.
178. **Gill SR, Pop M, Deboy RT, et al.** Metagenomic analysis of the human distal gut microbiome. Science 2006;312(5778):1355-9.
179. **Kurokawa K, Itoh T, Kuwahara T, et al.** Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. 2007;14(4):169-181.
180. **The Human Microbiome Project C.** A framework for human microbiome research. Nature 2012;486(7402):215-221.
181. **Nielsen HB, Almeida M, Juncker AS, et al.** Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 2014;32(8):822-8.
182. **The Human Microbiome Project C.** Structure, Function and Diversity of the Healthy Human Microbiome. Nature 2012;486(7402):207-214.
183. **Shreiner AB, Kao JY, and Young VB.** The gut microbiome in health and in disease. Current Opinion in Gastroenterology 2015;31(1):69-75.
184. **Hamady M and Knight R.** Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 2009;19(7):1141-52.
185. **Tap J, Mondot S, Levenez F, et al.** Towards the human intestinal microbiota phylogenetic core. Environ Microbiol 2009;11(10):2574-84.
186. **Claesson MJ, Cusack Sn, O'Sullivan O, et al.** Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proceedings of the National Academy of Sciences 2011.
187. **Tims S, Derom C, Jonkers DM, et al.** Microbiota conservation and BMI signatures in adult monozygotic twins. ISME J 2013;7(4):707-17.
188. **Salonen A, Salojarvi J, Lahti L, et al.** The adult intestinal core microbiota is determined by analysis depth and health status. Clin Microbiol Infect 2012;18 Suppl 4:16-20.
189. **Caporaso JG, Lauber CL, Costello EK, et al.** Moving pictures of the human microbiome. Genome Biol 2011;12(5):R50.
190. **Costello EK, Lauber CL, Hamady M, et al.** Bacterial community variation in human body habitats across space and time. Science 2009;326(5960):1694-7.
191. **Lampe JW, Navarro SL, Hullar MA, et al.** Inter-individual differences in response to dietary intervention: integrating omics platforms towards personalised dietary recommendations. Proc Nutr Soc 2013;72(2):207-18.
192. **McOrist AL, Miller RB, Bird AR, et al.** Fecal butyrate levels vary widely among individuals but are usually increased by a diet high in resistant starch. J Nutr 2011;141(5):883-9.
193. **Benjamin JL, Hedin CR, Koutsoumpas A, et al.** Randomised, double-blind, placebo-controlled trial of fructo-oligosaccharides in active Crohn's disease. Gut 2011;60(7):923-9.
194. **Whelan K, Judd PA, Preedy VR, et al.** Fructooligosaccharides and fiber partially prevent the alterations in fecal microbiota and short-chain fatty acid concentrations caused by standard enteral formula in healthy humans. J Nutr 2005;135(8):1896-902.
195. **Korpela K, Flint HJ, Johnstone AM, et al.** Gut Microbiota Signatures Predict Host and Microbiota Responses to Dietary Interventions in Obese Individuals. PLoS ONE 2014;9(3):e90702.
196. **Haiser HJ and Turnbaugh PJ.** Is it time for a metagenomic basis of therapeutics? Science 2012;336(6086):1253-5.

197. **Bodelier PLE.** Toward Understanding, Managing, and Protecting Microbial Ecosystems. Frontiers in Microbiology 2011;2:80.
198. **Ritz K.** The Plate Debate: Cultivable communities have no utility in contemporary environmental microbial ecology. FEMS Microbiology Ecology 2007;60(3):358-362.
199. **Belenguer A, Duncan SH, Calder AG, et al.** Two routes of metabolic cross-feeding between Bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut. Appl Environ Microbiol 2006;72(5):3593-9.
200. **Muñoz-Tamayo R, Laroche B, Walter É, et al.** Kinetic modelling of lactate utilization and butyrate production by key human colonic bacterial species. FEMS Microbiology Ecology 2011;76(3):615-624.
201. **Schloss P and Handelsman J.** Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biology 2005;6(8):229.
202. **Walker AW, Duncan SH, Louis P, et al.** Phylogeny, culturing, and metagenomics of the human gut microbiota. Trends Microbiol 2014;22(5):267-74.



Chapter 3

Microbiota conservation and BMI signatures in the microbiota of adult monozygotic twins

Sebastian Tims, Catherine Derom, Daisy M. Jonkers, Robert Vlietinck, Wim H. Saris, Michiel Kleerebezem, Willem M. de Vos, Erwin G. Zoetendal

The ISME Journal, 2013, 7(4), p.707–717 (minor modifications)

Abstract

The human gastrointestinal tract (GI) microbiota acts like a virtual organ and is suggested to be of great importance in human energy balance and weight control. This study included 40 monozygotic (MZ) twin pairs to investigate the influence of the human genotype on GI microbiota structure as well as microbial signatures for differences in body mass index (BMI). Phylogenetic microarraying based on 16S rRNA genes demonstrated that monozygotic twins have more similar microbiota compared to unrelated subjects ($p < 0.001$), which allowed the identification of 35 genus-like microbial groups that are more conserved between MZ twins. Half of the twin pairs were selected on discordance in terms of BMI, which revealed an inverse correlation between *Clostridium* cluster IV diversity and BMI. Furthermore, relatives of *Eubacterium ventriosum*, *Roseburia intestinalis* were positively correlated to BMI differences, and relatives of *Oscillospira guillermondii* were negatively correlated to BMI differences. Lower BMI was associated with a more abundant network of primary fiber degraders, while a network of butyrate producers was more prominent in subjects with higher BMI. Combined with higher butyrate and valerate contents in the fecal matter of higher BMI subjects, the difference in microbial networks suggests a shift in fermentation patterns at the end of the colon, which could affect human energy homeostasis.

Introduction

The microbial cells inside the human gastrointestinal (GI) tract are collectively called the GI microbiota and provide an extensive genetic function counterpart to the host genome [1-12]. Previous studies have shown that GI microbiota is host-specific and GI tract region-specific [13-15], aberrant in composition and stability in patients suffering from GI disorders such as Crohn's disease [16], and associated to host energy homeostasis [17-21]. Analysis of global fecal microbiota introduced the concept that human GI microbiota appeared to have three distinct structural biome-types called enterotypes [22]. Although the enterotype distinction did not appear to be correlated to health status or host demography, recent 16S rRNA based studies by Wu and co-workers [23] and by Huse and co-workers [24] suggest that the distinctive biomes among the human GI microbiota appear to be more like a continuum with gradients of the main enterotype driving taxa, which could be driven by long-term dietary habits.

Human energy homeostasis varies greatly between persons but monozygotic (MZ) twins show more resemblance in the variations of their energy balance [25]. Furthermore, current literature putatively links a large number of human genes to variations in body mass index (BMI). Recruitment of MZ twins discordant for BMI in studies exploring human transcript profiles has revealed several potential obesity marker-genes [26] and potential (mitochondrial) pathways that are associated with major BMI increases [27]. However, the identified human genes account for a relatively small amount of the observed variance in energy homeostasis [28, 29]. Recent findings suggest that the GI microbiota is important for the energy and metabolic homeostasis of its host. In mice clear links have been observed between energy homeostasis and GI microbiota, for instance: the resistance to obesity development of Germ-free mice [18], stimulation of weight gain by GI tract colonization [20], interaction between GI microbiota and fatty acid storage mechanisms [17, 18], variations between genetic obese (*ob/ob*) and lean mice in the relative abundances of the bacterial phyla Bacteroidetes, Firmicutes and Actinobacteria [21, 30].

In contrast to the associations found in mice, studies on the relation between human energy homeostasis and GI microbiota have generated conflicting results. Analogous to results obtained in mice, Ley and co-workers detected fewer Bacteroidetes and more Firmicutes in obese subjects compared to lean controls [19]. Moreover, this study also revealed that the relative abundances of Bacteroidetes increased while Firmicutes decreased when the subjects decreased their BMI by following either a fat restricted or carbohydrate restricted diet [19]. Although Duncan and co-workers confirmed a significant decrease in Firmicutes when subjects followed a low-carbohydrate weight-loss diet [31], several other human studies did not confirm these differences in Bacteroidetes to Firmicutes (B:F) ratio [31-33]. Schwiertz and co-workers even concluded that relative abundance of Firmicutes was reduced in obese subjects [33], but they also reported that higher levels of short chain fatty acids (SCFAs) were present in fecal material of obese subjects compared to lean controls, possibly suggesting that the amount SCFAs produced is a more prominent determinant of the BMI status than the phylogenic distributions of the microbiota [33]. The conflicting results in these studies may be explained by the heterogeneity among human subjects, with respect to their genotype and lifestyle as well as specificity of an individual's microbiota. Furthermore, these studies have compared subjects on the opposite

extremes of the BMI scale (lean and obese), while the microbiota is exposed to fundamentally different "environmental" factors in both states that go beyond BMI alone, such as diet, host metabolic and hormonal factors [34, 35], and low-grade systemic inflammation [36].

Results of mice studies often lead to the hypothesis that the mammalian host-genotype, in particular factors for the immune system phenotype, has a huge impact on the GI microbiota characteristics [37]. Genotype, however, is not determined so far for human subjects participating in microbiota studies. Consequently, these human studies do not take the host-genotype into account as a determinant of the phenotype, while the genotype is known to be heterogeneous amongst modern human populations.

Genotypic influences could be minimized by evaluating phenotypic variations in MZ twins allowing pair-wise comparisons within a fixed genotype. An early study of separately living MZ twins and their marital partners revealed that for co-twins the within-pair microbiota similarity is significantly higher compared to unrelated individuals, while the microbiota similarity for married couples is not significantly higher compared to unrelated individuals [38]. Similar observations were reported in later studies [39, 40]. Furthermore, differences in GI microbiota could be related to disease phenotype by comparing MZ twin pairs concordant and discordant for inflammatory bowel diseases [41-43]. A study by Turnbaugh and co-workers [40] on a cohort of obese and lean adult females, monozygotic and dizygotic twin pairs, demonstrated that the microbiota in obese pairs was reduced in bacterial diversity and contains an altered representation of bacterial genes and metabolic pathways compared with lean pairs [40]. This study, however, only included twins with concordant phenotypes in terms of BMI, again making both host genetics and absolute BMI values confounding variables between the groups of subjects. To identify specifically microbiota signatures of BMI, we compared the microbiota composition in MZ twin pairs that are concordant and discordant in BMI. This enabled us to achieve our main objective, which was to define microbiota signatures that correlate directly with BMI differences independent of the host genotype and absolute BMI values.

Materials & methods

Subjects, sample size and sampling

This study was approved by the METC of Wageningen University. A selection of MZ twin pairs was contacted from the East Flanders Prospective Twin Study, which presently has over 7.000 twins. Subjects who used medication that may affect the GI microbiota, prebiotics, or probiotics within one month before sampling were excluded. Subjects with pre-existing bowel diseases and subjects that were pregnant or breast feeding were excluded as well. Since there are no previous HITChip studies on MZ twins, the power calculation was based on intra-individual microbiota variation observed in healthy subjects [15] which we expect to be close to that within paired MZ twin microbiotas. Based on the expectation that we will have at least a proportional difference of 0.18 (SD 0.04), we calculated that 19 pairs per group would suffice to address our objective (assuming $\alpha = 0.05$ and power $1- \beta = 0.20$). Therefore, our cohort consisted of 20 twin pairs of varying genders and ages (> 18 years) that were previously recorded to have a BMI difference

of more than 5, and 20 age- and gender-matched control MZ twin pairs with no significant difference in BMI. Subjects were able to understand the written study information and signed an informed consent. Fecal samplings, bodyweight and length measurements were collected from these volunteers (Table S3.1). Furthermore, the volunteers filled in a questionnaire concerning changes in their dietary habits, medication, and gastrointestinal symptoms of the last 4 weeks prior to sampling (Table S3.1). Fecal samples were collected at the volunteers home, frozen immediately, and transported on dry ice to the laboratory where they were kept at -80 °C until further analysis.

Twin pair classification into BMI phenotypes

No subjects were underweight (BMI < 18.5). Only two twin pairs and one sibling of a third pair were obese (BMI > 30). To define which twin pairs were discordant in terms of weight maintenance the recommendation by Stevens and co-workers was used [44] (details see SI Materials and Methods), which in a maximum BMI difference for within-pair concordance of 1.35 kg/m^2 and a minimum BMI difference for within-pair discordance of 2.7 kg/m^2. This definition ensures that twins classified as discordant truly have different phenotypes. Discordant twins were subdivided into two groups based on their relative weight compared to their own sibling. This resulted in a "lower BMI" group for the leaner siblings and a "higher BMI" group for the heavier siblings. At the moments of sampling 18 pairs were discordant, 16 were concordant, and 6 occupied the grey area between our definitions of concordance and discordance (indistinct twins).

Microbial DNA extraction, microarray hybridization and data extraction

Microbial DNA was extracted utilizing the repeated bead beating protocol [45]. Fecal microbial diversity and composition was studied in detail using the Human Intestinal Tract chip (HITChip) as described previously [14] (further details see SI Materials and Methods). This phylogenetic microarray has been shown to be a powerful tool for deep GI tract microbiota composition analysis and has been benchmarked against several classical 16S rRNA gene-based methodologies, such as qPCR, FISH, and 454 pyrosequencing [14, 46-48] as well as metagenomics [22]. HITChip probes are assigned to three phylogenetic levels: level 1, defined as order-like 16S rRNA gene sequence groups; level 2, defined as genus-like 16S rRNA gene sequence groups (sequence similarity > 90%); and level 3, phylotype-like 16S rRNA gene sequence groups (sequence similarity > 98%) [14].

Organic acid and short chain fatty acid concentration measurement

To determine metabolic profiles, fecal samples were diluted in deionized water to a 10% (w/v) concentration and subsequent HPLC analysis was performed as described previously [49] to determine citrate, malate, succinate, lacate, fumarate, formate, acetate, propionate, iso-butrate, butyrate, and valerate concentrations. The HPLC system was equipped with a Shodex S1821 column and temperature was set to 70 °C.

Data analysis and statistical methods

For the total microbiota normalized signal values of all unique HITChip probes were used to calculate Simpson's Diversity index for each sample and the Spearman's correlation coefficient between different samples. Unique probe signal values per level 1 and per level 2 group (with >20 probes) group were used to calculate diversity and similarities for the groups of each phylogenetic level separately. Spearman's correlation coefficients between random unrelated subjects within this cohort were compared to Spearman's correlation coefficients within twin pairs by a Student's t-test with unequal groups.

For each sample relative abundances were calculated for the groups of each specificity level by summing all signal values of the probes targeting a group and dividing by the total of all probe signals for the corresponding sample. All comparisons between the discordant twin groups were pair-wise and significance were assessed with dependent 2-group Wilcoxon signed rank tests. For all statistical tests that were performed on multiple parameters the obtained p-values were adjusted by a Bonferroni correction. All p-values noted in text are adjusted p-values with $p<0.05$ being regarded as significant.

Results

Monozygotic twins have highly similar microbiotas.

A host-genotype controlled setup for this study was realized by recruiting MZ twins. Twin pairs enrolled in this study were contacted from the East Flanders Prospective Twin Study (Flanders, Belgium) [50]. A total of 40 MZ twin pairs volunteered to donate fecal material from which DNA was extracted for microbiota composition analysis with phylogenetic microarray the Human Intestinal Tract Chip (HITChip [14]). This twin cohort consisted of 11 male pairs (age 19-43 years, BMI 18.5-34.7 kg/m^2, see Table S3.1) and 29 female pairs (age 20-43 years, BMI 20.2-34.7 kg/m^2). This selection consisted of 20 twin pairs of varying genders and ages that were previously recorded to have a BMI difference of more than 5 units. In addition 20 age and gender matched control twin pairs with no significant difference in BMI were selected from the twin cohort (Table S3.1).

Similarity of the HITChip profiles between all subjects was calculated to determine the influence of the human genotype influence on the GI microbiota composition. Both, co-twins concordant (ΔBMI < 1.35, n = 16) and discordant (ΔBMI > 2.7, n = 18) for BMI showed a significantly higher similarity of their GI microbiota profile compared to random-paired subjects ($p < 0.001$, Figure 3.1A). Similarity coefficients between co-twins did not significantly correlate to age or the length of time that the twins had been living separately, corroborating the previously reported impact of the host-genotypic factors on the microbiota composition [38-40].

To determine which microbial groups contributed the most to the high within-pair similarity, similarity indices between all subjects were calculated for all microbial subgroups. These subgroups can be defined at different phylogenetic levels, such as level 1 which is defined as order-like 16S rRNA gene sequence groups, and level 2 which is defined as genus-like 16S rRNA gene sequence groups (sequence similarity > 90%), as described previously [14]. Within-pair microbiota profile similarity of each level 1 and 2 subgroup was higher compared to random-paired subjects (Figure S3.1 and S3.2), although this difference was not significant for every subgroup. The co-twins showed a significantly higher within-pair similarity compared to random-paired subjects for 12 order-like (level 1) and 35 genus-like (level 2) groups (Figure 3.1B and C). Moreover, for *Clostridium* clusters XI and XIVa and for 27 genus-like groups the within-pair similarities were even significantly higher than the total microbiota similarity (Figure 3.1B and C, Table S3.2). A high similarity index within a bacterial group means that the probe signals occur in (nearly) the same ratios relatively to one another, implying that the presence and ratios of specific bacteria belonging to certain groups are highly conserved between the subjects. Hence, these structurally conserved bacterial groups within MZ twin pairs acknowledge the existence of a structural core in the human GI microbiota, which correlates with host genetics and shared (early life) environmental exposures and therefore can be considered as "imprinted structural cores".

Next to an imprinted structural core the existence of a general core based on phylotype abundance was investigated. For HITChip data phylotype-like groups are defined as 16S rRNA gene groups with a sequence similarity of > 98%. Above the array background specific signals of 96 phylotype-like were found in all subjects (100% prevalence). These phylotype-like groups, which comprise the general core in this twin cohort, are shown in Table S3.3, and Figure S3.3. In this twin cohort the general core accounts on average for 34.7% (SD 12.8%) of the total microbiota. However, the abundance of this general core varies greatly between the subjects (from 10.6% for twin 10A to 81.5 % for twin 26A), indicating enormous subject specificity at phylotype level in the GI tract.

Clostridium cluster IV is less diverse in higher BMI sibling group.

To evaluate if microbial groups are associated to BMI and host genetic traits, microbiota composition was compared between twins that are concordant and discordant for BMI. At the moment of sampling the BMI difference for 15 of the selected discordant twins was less than the 5 units recorded previously. Six twin pairs were between the ΔBMI thresholds used to determine concordance and discordance (1.35 > ΔBMI > 2.7) and were not taken into account. Each discordant twin pair was split up into two groups: the sibling with the lowest BMI of each pair was placed in the "lower BMI siblings" group and the sibling with the highest BMI of each pair was placed in the "higher BMI siblings" group. At the highest phylogenetic levels no differences were observed between the two discordant twin pair groups. No consistent Bacteroidetes:Firmicutes (B:F) ratio differences were observed in pair-wise comparison of lower- and higher-BMI siblings (Figure S3.1). Similarly, B:F ratios did not correlate with absolute BMI values. Moreover, these

pair-wise comparisons did not reveal consistent differences between the total microbial diversity (inverse Simpson's index of diversity), which appeared to be hugely variable and between pairs ranged from 1.8 (twin pair 2; Table S3.1) to 149.0 (twin pair 30; Table S3.1). This variability indicates that the total microbiota diversity does not relate to the phenotypic differences in this twin cohort.

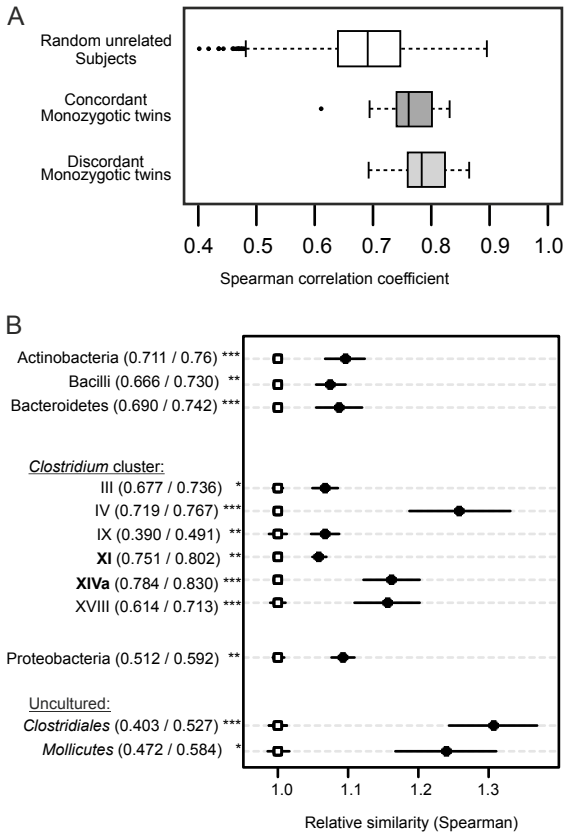


Figure 3.1 Gastrointestinal microbiota similarity in monozygotic twins. (A) Box-whisker plots of total microbiota profile similarity. Spearman's correlation coefficient was calculated for random unrelated subjects and between monozygotic twins. Average microbiota similarity between twins concordant for BMI and twins discordant for BMI are both significantly higher than between unrelated subjects ($p = 1\ e^{-4}$, $p = 1\ e^{-7}$, respectively). Dot-plots are shown of the mean order-like (B) similarity abetween random-paired subjects and monozygotic twins. Order-like groups that were significantly different in similarity index are presented. The similarity indices of all order-like groups are represented in figure S3.1. Mean within-pair Spearman's correlation coefficient values (depicted by black dots) and are relative to the mean Spearman's correlation coefficient values of random unrelated subjects within this cohort (depicted by open squares), for each phylogenetic group. The plot-labels describe the actual mean Spearman's correlation coefficient value (unrelated subjects / monozygotic twins). Error bars represent 95% confidence intervals around the respective mean values. Asterisks indicate the level of significance of the corrected p-value: *) $p < 0.05$, **) $p < 0.01$, ***) $p < 0.001$.

However, diversity at order-like (level 1) groups indicated that *Clostridium* cluster IV was significantly lower in diversity in the higher BMI siblings group (p = 0.012), indicating that *Clostridium* cluster IV diversity decreases when BMI increases, independent of the absolute BMI value of the lower-BMI-sibling (Figure 3.2A). Overall, the diversity of *Clostridium* cluster IV in the lower BMI siblings group is more comparable to the control group than the diversity in the higher BMI group.

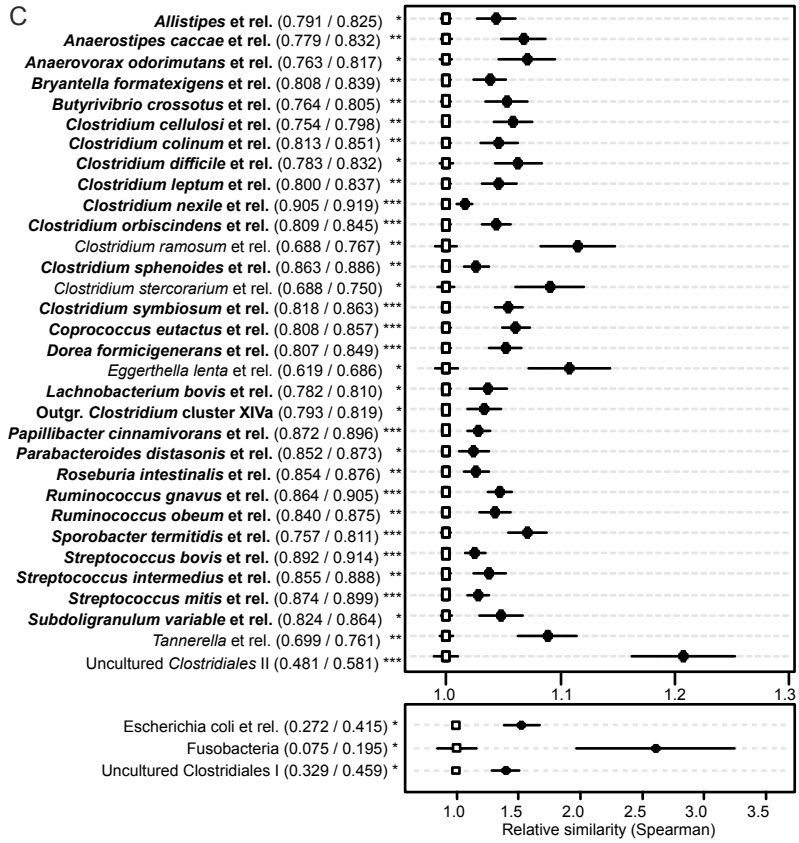


Figure 3.1 (continued) Gastrointestinal microbiota similarity in monozygotic twins. Dot-plots are shown of the mean genus-like (C) similarity between random-paired subjects and monozygotic twins. Genus-like groups that were significantly different in similarity index are presented. The similarity indices of all genus-like groups are represented in figure S3.2. Mean within-pair Spearman's correlation coefficient values (depicted by black dots) and are relative to the mean Spearman's correlation coefficient values of random unrelated subjects within this cohort (depicted by open squares), for each phylogenetic group. The plot-labels describe the actual mean Spearman's correlation coefficient value (unrelated subjects / monozygotic twins). Error bars represent 95% confidence intervals around the respective mean values. Asterisks indicate the level of significance of the corrected p-value: *) p < 0.05, **) p < 0.01, ***) p < 0.001.

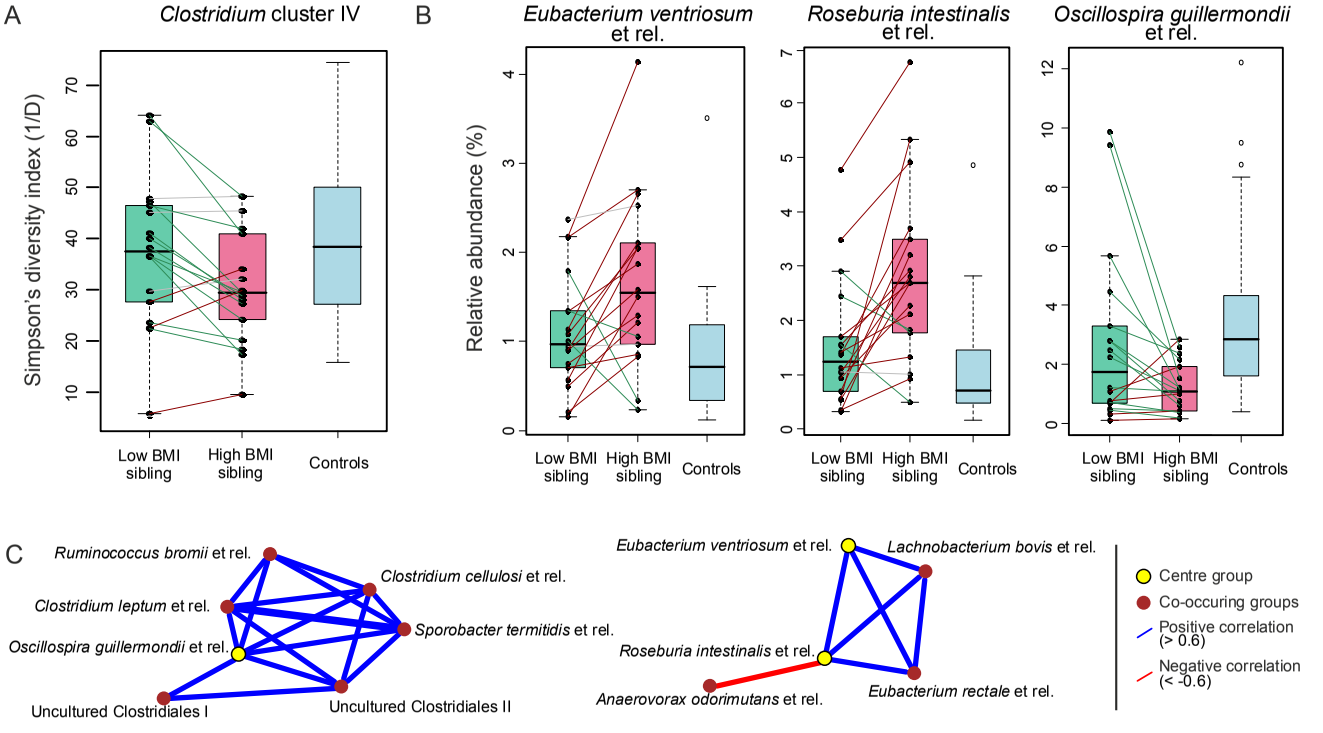


Figure 3.2. BMI phenotype influences on gastrointestinal microbiota. A. Box-whisker plots of inverse Simpson's index of diversity for *Clostridium* cluster IV in discordant and concordant twins. The results of each subject from the two discordant twins groups are visualized with black dots. Colored lines connect the dots of each twin pair. Grey lines indicate no change in diversity (or less than a factor of 0.1). A green line indicates that the diversity is higher in the lower BMI sibling, while a red line indicates a higher diversity in the higher BMI sibling. B. Box-whisker plots of the relative abundances of the genus-like groups significantly differing in discordant twins: *Eubacterium ventriosum* et rel., *Roseburia intestinalis* et rel., and *Oscillospira guillermondii* et rel. The results of each subject from the two discordant twins groups are visualized with black dots. Colored lines connect the dots of each twin pair. Grey lines indicate no change in relative abundance (or less than a factor of 0.1). A green line indicates that the relative abundance is higher in the lower BMI sibling, while a red line indicates a higher relative abundance in the higher BMI sibling. C. Co-occurrence networks in all subjects of the genus-like groups significantly differing in discordant twins. *Eubacterium ventriosum* et rel. and *Roseburia intestinalis* et rel. (more abundant in higher BMI siblings) appear in a network of butyrate producers, and *Oscillospira guillermondii* et rel. (more abundant in lower BMI siblings) appears in a network of primary fibre degraders.

BMI phenotype correlates with signatures in genus-like microbial groups.

To determine if quantitative differences in microbiota composition were consistently different between the higher BMI siblings and the lower BMI siblings groups given an identical genetic background, relative abundances of specific microbial groups were compared and contrasted. This revealed that the genus-like (level 2) groups *Eubacterium ventriosum* et rel. and *Roseburia intestinalis* et rel. were significantly more abundant in the higher BMI siblings ($p = 0.014$ and $p = 0.003$, respectively; Figure 3.2B), while and *Oscillospira guillermondii* et rel. was significantly more abundant different in the lower BMI siblings ($p = 0.014$; Figure 3.2B). These three genus-like (level 2) groups seem to co-occur in two distinct ecological networks with several other genus-like groups in all subjects when using a Spearman's correlation coefficient cut-off of > 0.6 and < -0.6 (Figure 3.2C). Such ecological networks may visualize potential cooperation (mutualism or commensalism) and competition between the microbial groups. The first network, which is enriched in lower BMI sibling group, is centered around *Oscillospira guillermondii* et rel. and contains three other genus-like (level 2) groups which encompass isolates associated with degradation of plant material: *Clostridium cellulosi* et rel. [51], *Ruminococcus bromii* et rel. [52, 53], and *Sporobacter termitidis* et rel. [54]. Therefore, this network seems to be specialized in degradation of complex fibers, marking this network as a primary degrader network. Degradation of complex fibers can yield fermentation products, such as partially degraded oligosaccharides, acetate and lactate, which in turn can be used as substrates for those butyrate producers that act like scavengers [55]. The second network, which is enriched in the higher BMI sibling group, includes *Eubacterium ventriosum* et rel. and *Roseburia intestinalis* et rel. which both contain known butyrate producing isolates capable of degrading fibers themselves [56, 57]. Furthermore this network also includes *Eubacterium rectale* et rel. which is another group with known butyrate producers. Therefore, the second network seems to be a butyrate producing network.

Short chain fatty acid profiles show within-pair differences in discordant twins.

Since both discordant siblings were found be enriched in different fermentation networks metabolic profiling was performed to confirm if these differences were visible in the fermentation products of these sibling groups as well. Although several organic acids were not detected in this cohort (i.e. citrate, lactate, fumarate, and formate), acetate, propionate and butyrate were the most dominating metabolites in both groups From all detected organic compounds only butyrate and valerate were present at significantly higher concentrations in higher BMI siblings compared to their lower BMI siblings (Table 3.1), which is in line with the predicted networks (Figure 3.2C).

Table 3.1. Fecal organic acid and short chain fatty acid concentration of twins discordant and concordant for BMI. Citrate, lactate, fumarate, and formate were not detected. Statistical testing of between the concentrations of lower BMI sibling and higher BMI sibling groups was performed with paired dependent 2-group Wilcoxon signed rank tests and p-values are reported (significant p-values are in bold).

Component	Lower BMI siblings (n = 16) (mean ± SDE)	Higher BMI siblings (n = 16) (mean ± SDE)	p-value	Controls (n = 32) (mean ± SDE)
Acetate	56.44 +- 6.38	74.41 ± 9.04	0.252	56.97 ± 6.49
Butyrate	13.76 +- 1.59	25.12 ± 3.51	0.013	16.40 ± 1.90
Iso-butyrate	0.23 +- 0.14	1.15 ± 0.33	0.067	1.09 ± 0.29
Malate	10.56 +- 2.60	25.71 ± 8.3	0.403	11.81 ± 2.84
Propionate	16.82 +- 1.76	25.25 ± 4.04	0.193	15.86 ± 1.42
Succinate	0.64 +- 0.35	0.03 ± 0.03	0.178	1.41 ± 1.21
Valerate	1.47 +- 0.30	3.43 ± 0.62	0.006	2.00 ± 0.30

3

Discussion

Human GI microbiota and its relation to differences in BMI was investigated in a host-genotype controlled setup by analyzing fecal samples of MZ twins. The cohort comprised of MZ twins discordant and concordant for BMI, allowing us to assess the influence of BMI differences independent of the absolute BMI value of the twins. With this MZ twin control study we were also able to control for gender, age, birth weight and other prenatal and postnatal exposures shared by the co-twins. Furthermore, with 80 subjects containing 40 different human genotypes, this cohort allowed to assess more generic topics like the human GI microbiota core.

Due to this variety of core definitions and molecular techniques employed in current literature no consensus on a general human GI microbial core has emerged [15, 40, 58-62]. In our cohort we could define a general core microbiota of 96 phylotype-like groups prevalent in all subjects that accounts for 34.7% of the total microbiota (SD 12.8%). However, in line with recent observations [62], this general core is very dependent on detection threshold and furthermore highly subject specific, as even the most prominent phylotype-like groups comprised only 0.1 - 0.22% of the total microbiota in some subjects. Our assessment of a general phylogenetic GI microbiota core agrees with various previous studies, however Turnbaugh and co-workers and Tap and co-workers did not find a common microbial core [40, 59]. Notably Turnbaugh and co-workers used a higher relative abundance threshold than we used here, i.e. 0.5% [40]. Although Tap and co-workers defined a phylogenetic core that accounted for 35.8% of the total sequences from their cohort [59], their core definition only accounts for 8.1% of the total signals in our twin cohort. Next to methodology and the actual core definitions, the high subject specificity of core phylotypes complicates assessment of the general GI microbiota core.

In contrast to a general genotype independent core based on phylotype abundance, a genotype-dependent structural core (at higher phylogenetic levels) was much more pronounced in our data set. This study demonstrates that the human genotype and possibly early-life stimuli exhibit a strong influence on the GI microbiota structural composition and extends earlier findings on twins and their relatively high degree of inter-pair microbiota similarity [38, 63].

Host metadata (Table S3.1) revealed no factors that significantly influenced the microbiota profile similarity. Yet MZ twins have GI microbiota profiles that are more similar to each other than to random unrelated subjects, despite the fact that half of them is discordant in terms of BMI (Figure 3.1A). Within-pair microbiota similarity was not equally represented over the phylogenetic subgroups (Figure S3.1 and S3.2). From different phyla several subgroups displayed within-pair profiles which where conserved among MZ twin pairs (Figure 3.1B and 1C, Table S3.2). Interestingly, most of the conserved phylogenetic subgroups are significantly higher compared to the total within-pair microbiota similarity. Several other genus-like groups were found to be very dissimilar compared to the total microbiota similarity, which include some facultative anaerobes that are known to be fastidious (Figure S3.2). Although this could suggest that these groups likely respond quickly to changing conditions, this remains speculative. The conserved profiles do not necessarily mean that the corresponding groups are present at similar abundance levels, for instance, *Roseburia intestinalis* et rel. and *Eubacterium ventriosum* et rel. are less abundant in the lower BMI siblings compared to their higher BMI siblings (Figure 3.2B). Similarities of the conserved genus-like groups were independent of host phenotype (BMI related or otherwise). Therefore, the human microbiota appears to have an imprinted structural core. Some of the microbial groups of the imprinted structural core are more strongly genotype dependent, like *Clostridium ramosum* et rel., *Escherichia coli* et rel., *Eggerthella lenta* et rel., and genus-like groups of *Fusobacteria* and uncultured *Clostridiales*, as their within-pair similarity is at least 10% higher compared to random-paired subjects. From the imprinted structural core *Clostridium nexile* et rel., *Ruminococcus gnavus* et rel. and *Streptococcus bovis* et rel. have the highest within-pair similarity (>0.9 Spearman's correlation coefficient) but they are also highly similar in random-paired subjects. The existence of an imprinted structural core strengthens and extends earlier studies which link human genotype to GI microbiota composition [38, 63]. Moreover, scientific findings have not yet established the exact potency of early-life epigenetic imprinting on adult human beings. MZ co-twins have likely experienced the same stimuli and conditions during early-life. Whether or not early-life dietary influences are important, next to the host-genotype, for GI microbiota composition and the imprinted structural core remains an outstanding question. By design, this study circumvents early-life influences by pair wise analyses between co-twins with an identical genetic background.

Remarkably, approximately half of the genera that are structurally conserved within MZ twin pairs were previously reported to be driving genera for the classification of the three enterotypes [22]. Although enterotypes are classifications based on microbial abundance levels [22] while the imprinted structural core members are not necessarily highly abundant in every subject, it is noteworthy that we observed a significant overlap between bacterial groups that are important for enterotype classification and the imprinted structural core. Finding these particular members to be part of the imprinted structural core adds to the debate on the classification of the GI microbiota structure, i.e. whether the GI microbiota can truly be classified into distinct types (enterotypes) or if the GI microbiota should be regarded as a "state" (a part of a continuum). Our data indicate that the imprinted structural core harbors the capacity to form each enterotype (Table S3.2), thereby favoring the previous observations that enterotypes can best be seen as

distinct states, rather than distinct types. This finding deserves more attention in longitudinal dietary studies, especially since drivers from all three enterotypes are found to be structurally conserved per genotype and not just the drivers of one of the proposed enterotypes.

In previously reported studies human genotype and phenotype could not be distinguished when reporting on the GI microbiota of lean and obese individuals. High expectations are raised in current literature about the role of GI microbiota on host energy homeostasis and weight control. This may not be justified given the fact that in this twin population with an identical genetic background discordance for BMI do not spectacularly differ in GI microbiota. It appears to be unlikely for genetically identical people to diverge in BMI as much as the subjects in cross sectional studies on high and low BMI. Moreover, extreme differences in BMI and drastic weight-loss regimes are accompanied by many other potential confounding factors, such as (extreme) changes in diet, host physiology, host health status, and change of physical activity and their consequences on the overall host metabolism. Our study design enabled the elimination of genotype influences, by pair-wise comparison of MZ twins with discordant BMIs independent of absolute BMI values. This allowed us to strongly corroborate the influence of the host-genotype on the GI microbiota structure, but this also unveiled specific microbiota differences associated to host phenotype.

In our twin cohort no trend between BMI differences and total microbiota diversity or B:F ratio was detected, which agrees with the findings reported by Duncan and co-workers [31]. At lower phylogenetic levels, however, microbial differences related to differences in BMI phenotype were detected. A consistent pair wise difference in diversity within the discordant twin pairs can be found for *Clostridium* cluster IV. Our results indicate that this group is associated with phenotypic changes in BMI, decreasing in diversity as BMI increases. Furthermore, from this *Clostridium* cluster we found the relatives of the long known, yet uncultivated *Oscillospira guillermondii* to be significantly higher in the lower BMI siblings. Interestingly, previous findings link low *Clostridium* cluster IV diversity to Crohn's disease [64]. Therefore it appears that *Clostridium* cluster IV diversity can be affected by systemic changes of the host phenotype. Moreover, depletion of several members of *Clostridium* cluster IV, in our case *Oscillospira guillermondii* et rel., is associated with phenotypic changes.

Members of the morphologically distinct genus *Oscillospira* are frequently seen in cattle and sheep rumen. Several *Oscillospira* species react to the diets of their host, increasing strongly when the hosts are feeding on fresh green fields [65]. It is likely that these organisms are adapted to degrade the fibers of young plants. Hence the plant fiber content in human diet might also influence the presence and abundance of *Oscillospira* species. Without further knowledge on the metabolism of *Oscillospira guillermondii* et rel. it is hard to elucidate the extent of their role in the host energy homeostasis. Possibly, the presence of *Oscillospira* species has an impact on the metabolism of nutritional fibers. In line with this hypothesis, is the finding of three other genus-like (level 2) groups that co-occur with *Oscillospira guillermondii* et rel. (Figure 3.2C) which have isolates associated with degradation of plant material: *Clostridium cellulosi* [51], *Ruminococcus bromii* [52, 53], and *Sporobacter termitidis* [54]. It seems the co-occurrence network of *Oscillospira guillermondii* et rel. is specialized in fermenting complex (plant) materials. Fermentation products

from such a primary degrader network can be used by scavenging butyrate producing bacteria [55].

Both the *Eubacterium ventriosum* et rel. and *Roseburia intestinalis* et rel. groups were significantly higher in the higher BMI siblings compared to their corresponding lower BMI siblings. Cultured isolates belonging to these two taxonomic groups are known butyrate producers [56, 57]. However, other results from the HITChip data did not reveal additional differences in butyrate production (potential), as based on similar levels of other butyrate-producing organisms in the discordant twins. Analogously, the co-occurrence network of the *Eubacterium ventriosum* et rel. and *Roseburia intestinalis* et rel. (Figure 3.2C) did not include other potential butyrate producing organisms that are detected by HITChip, such as: *Butyrivibrio crossotus* et rel., *Coprococcus eutactus* et rel., *Eubacterium hallii* et rel., *Faecalibacterium prausnitzii* et rel,. *Megasphaera elsdenii* et rel., or *Mitsuokella multiacida* et rel. Moreover, *Roseburia intestinalis* et rel. displays anti-occurrence with *Anaerovorax odorimutans* et rel., which has an representative isolate capable of butyrate production as well [66].

In contrast to the primary degrader network of *Oscillospira guillermondii* et rel., the network of *Eubacterium ventriosum* et rel. and *Roseburia intestinalis* et rel. is composed of butyrate producers that are likely able to degrade complex material on their own. We hypothesize that host BMI increase is accompanied by GI microbiota changes and therefore a metabolic shift in butyrate production structure inside the colon: from mainly scavenging fermentation products produced by primary degraders to produce butyrate to mainly fermenting fibers directly into butyrate. Such a metabolic shift is likely to alter the net energy production by the GI microbiota and subsequently affect host energy harvest. This hypothesis is strengthened by finding significantly more butyrate in the fecal content of the higher BMI siblings, which are enriched for the butyrate producing network, compared to their lower BMI co-twins. Moreover, more valerate is found in the higher BMI siblings as well. Valerate can be formed by fermentation of the amino acid proline [67], which could indicate two (not mutually exclusive) possibilities: 1) no carbohydrates are left for a part of the microbial community, forcing this part to switch to amino acid fermentation; 2) the microbial community is more efficient in utilizing different types of (polymer) nutrients. Hence, the valerate results are in line with the network predictions in both discordant sibling groups and add to our hypothesis on the net energy production by the GI microbiota. However, from our data it is not possible to determine causality. If primary degraders outcompete the butyrate producers that are capable of degrading fibers, then subjects with lower BMI may only have scavenging butyrate producers left due to a diet rich in complex fibers. This is in line with the observation that a higher intake of fiber leads to lower levels of BMI. On the other hand, subjects with lower BMI could have much less butyrate producers, scavenging or otherwise, to begin with and therefore these subjects do not possess a microbiota that can harvest all available energy from the diet. Furthermore, the higher BMI siblings could also have consumed more protein and therefore show the increase in valerate. Therefore, more information on dietary intake and fermentation products in the GI tract is needed to further elucidate the mechanism between GI microbiota and host energy harvest.

Overall, this study revealed the existence of an imprinted structural core, to which several enterotype drivers belong, in the human GI microbiota and that BMI-phenotypic signatures are observed that can be related to energy harvest potential. However, the results are not such that a clear weight regulatory effect can be explained. Nonetheless, we have shown that different microbial networks are associated to changes in BMI. A network of primary degraders was more prominent in subjects with lower BMI, while a network of butyrate fermenters was more prominent in subjects with higher BMI. Our data suggest that primary degraders, not capable of producing butyrate, could play a more important role in energy homeostasis than initially expected. Previous studies on obesity have mostly compared cross-sectional the microbiota of lean and obese people and reported contradicting results. Besides the differences in host-genotypes in these studies the subject are at different ends of the BMI scale, i.e. the lean and obese "states". Here, we studied genetically identical human hosts and assessed BMI differences independent of the lean or obese characteristic of the individuals, enabling the detection of BMI-phenotype signatures in the GI microbiota. Known genetic background, or minimizing its influence in studies like this, will be pivotal in the deciphering of the effects of GI microbiota on mechanisms underlying phenotypic traits of the host, such as changes in BMI as reported here.

Acknowledgements

We would like to thank Lut De Zeure for her excellent technical assistance and of course we would like to thank the twins that participated. Since its origin the East Flanders Prospective Survey has been partly supported by grants from the Fund of Scientific Research, Flanders (Belgium), and by the Association for Scientific Research in Multiple Births (Belgium). The funding sources had no role in the design and conduct of the study, in the collection, analysis and interpretation of the data, or in the preparation, review or approval of the manuscript. Authors ST and EGZ received funding from the European Community's Seventh Framework Programme (FP7/2007-2013): MetaHIT, grant agreement HEALTH-F4-2007-201052, coordinated by S. Dusko Ehrlich (Institut National de la Recherche Agronomique, France).

Supplementary information

Supplementary materials & methods

Twin pair classification

To define which twin pairs were discordant in terms weight maintenance the recommendation by Stevens and co-workers was used: in adults long-term weight maintenance is considered as a weight change of < 3% of body weight [44]. The average weight of the recruited subjects was 68 kg, therefore the normal variation of 3% would on average correspond to 2,0 kg for one subject. Both siblings from one twin pair may vary in weight independently. Therefore twice the normal variation, i.e. 4 kg, is used as a maximum allowed weight maintenance variation to consider twin siblings concordant in body mass. To apply this weight concordance limit to all subjects the 4 kg variation was converted into a BMI difference (using the average subject length of 1.72 m), which is 1.35 kg/m^2. Twice the maximal normal weight variation was taken used to define discordance in weight maintenance, which is 2.7 kg/m^2.

Microbial DNA extraction, microarray hybridization and data extraction

Fecal microbial diversity and composition was studied in detail using the Human Intestinal Tract chip (HITChip) as described previously [14]. The DNA extracts were diluted to 20 ng/µl. Bacterial 16S rRNA genes were amplified in PCR using primers T7prom-Bact-27-for (5'-TGAATTGTAATACGACTCACTATAGGGGTTTGATCC TGGCTCAG–3') and Uni-1492-rev (5'-CGGCTACCTTGTTACGAC-3'). In vitro transcription of 500 ng of T7-16S rRNA gene amplicons for each sample were performed with the Riboprobe System (Promega, Madison, WI, USA). Subsequently, the 16S rRNA gene transcripts were labeled with CyDyes using the Post-Labeling Reactive Dye (Amersham Biosciences, Little Chalfont, UK) dissolved in 84 µl DMSO. For each sample Cy3-labelled RNA and Cy5-labelled RNA were fragmented and hybridized to the HITChip microarrays, which were produced by Agilent technologies (Agilent Technologies, Palo Alto, CA., USA). Analysis of the microarray data was performed in a custom-designed relational database which runs under the MySQL database management system [http://www.mysql.com/; 14]. Data normalization and analysis was performed using a set of R based scripts [http://www.r-project.org/] as described before [14]. As a quality threshold only those samples were accepted for which the different Cy-dye measurements displayed a Pearson's correlation coefficient of 0.98 or higher at probe signal intensity level.

Supplementary Results

Phylogenetic group conservation in twins

Monozygotic twin pairs have a more similar microbiota composition compared to random unrelated individuals. At lower phylogenetic levels the similarity compared to random unrelated subject is higher for all subgroups as well, although the degree of similarity varies extensively (Figure S1 and S2).

Core phylotypes detection

Phylotype level HITChip data shows a core microbiota can be defined (Figure S3). Different core definitions have been used in literature [40, 58, 59]. The studied individuals' core phylotypes can be detected at different minimal relative abundance thresholds. The highest relative abundance thresholds that still had at least one core phylotype with in all subjects (n=80) was 0.22%. Employing quite different definitions Tap and co-workers do find a phylogenetic core: 66 operational taxonomic units (OTUs) with a prevalence of >50% (n = 17) account for 35.8% of the sequences of their cohort [59]. The HITChip contains probes capable of detecting 62 of the 66 core OTUs defined by Tap and co-workers [59], which account for 8.1% of the total signals in our twin cohort. Although the HITChip does not detect all of the OTUs defined by Tap and co-workers, it is unlikely that the 4 missing OTUs would account more than 25% of the bacterial population in our cohort.

Bacteroidetes : Firmicutes ratio

B:F ratios were calculated by dividing the sum of signals of the level 1 Bacteroidetes group by the sum of signals of all level 1 groups belonging to the Firmicutes combined (i.e. *Bacilli*, all *Clostridium* clusters, uncultured *Clostridiales*, and uncultured *Mollicutes*). In four discordant twin pairs the Bacteroidetes to Firmicutes (B:F) ratio is found to be higher in the high BMI twin, while in four other pairs this ratio is higher in the low BMI twin. In six, concordant, control pairs one of the twins has a much higher B:F ratio compared to its respective sibling. However most twin pairs, discordant or concordant, do not show large differences in B:F ratio (Figure S4).

Supplementary figures

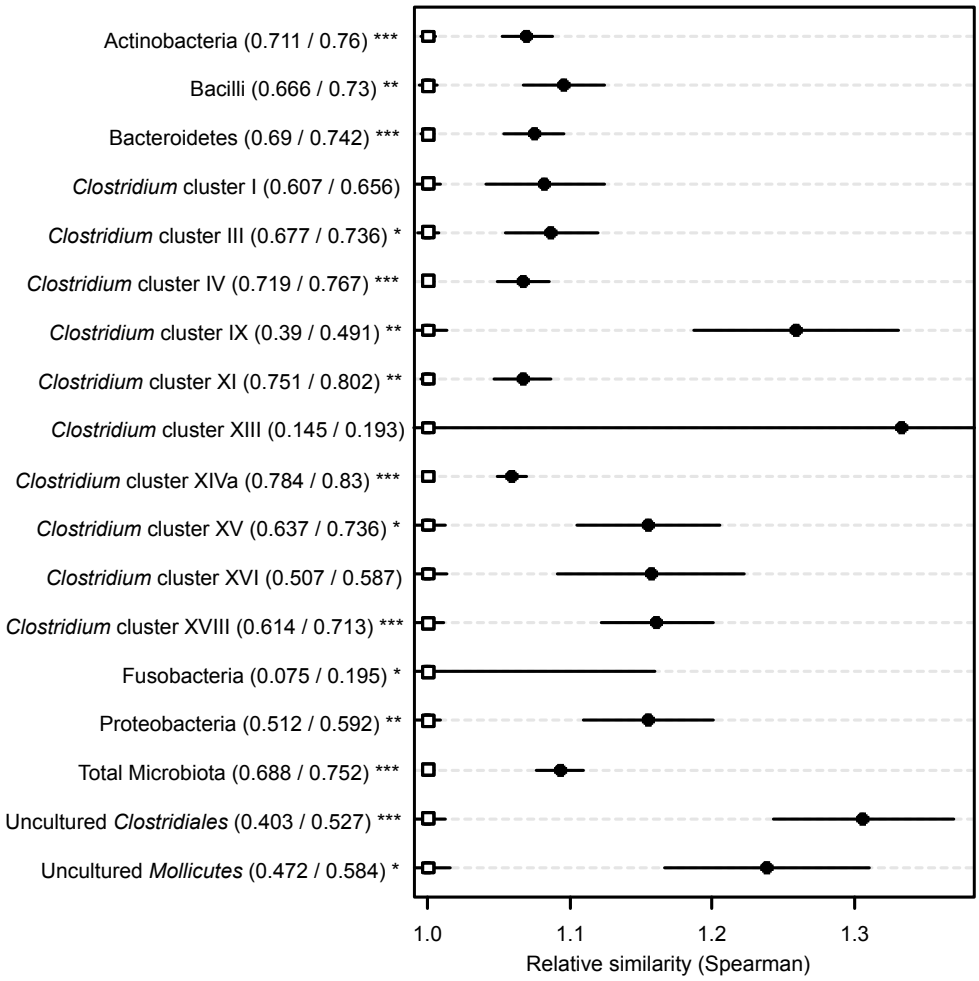


Figure S3.1 Gastrointestinal microbiota similarity in monozygotic twins of order-like (level 1) groups. Dot-plots of the mean total microbiota profile similarity and the mean phylogenetic level 1 group similarity between monozygotic twins. Only level 1 groups with more than 20 probes are shown here. Mean Spearman's correlation coefficients values for phylogenetic subgroups are depicted with black dots. Error bars represent 95% confidence intervals around the respective mean values. Open squares indicate average Spearman's correlation coefficients of random unrelated subjects within this data set. Asterisks indicate the level of significance of the corrected p-value: *) $p < 0.05$, **) $p < 0.01$, ***) $p < 0.001$

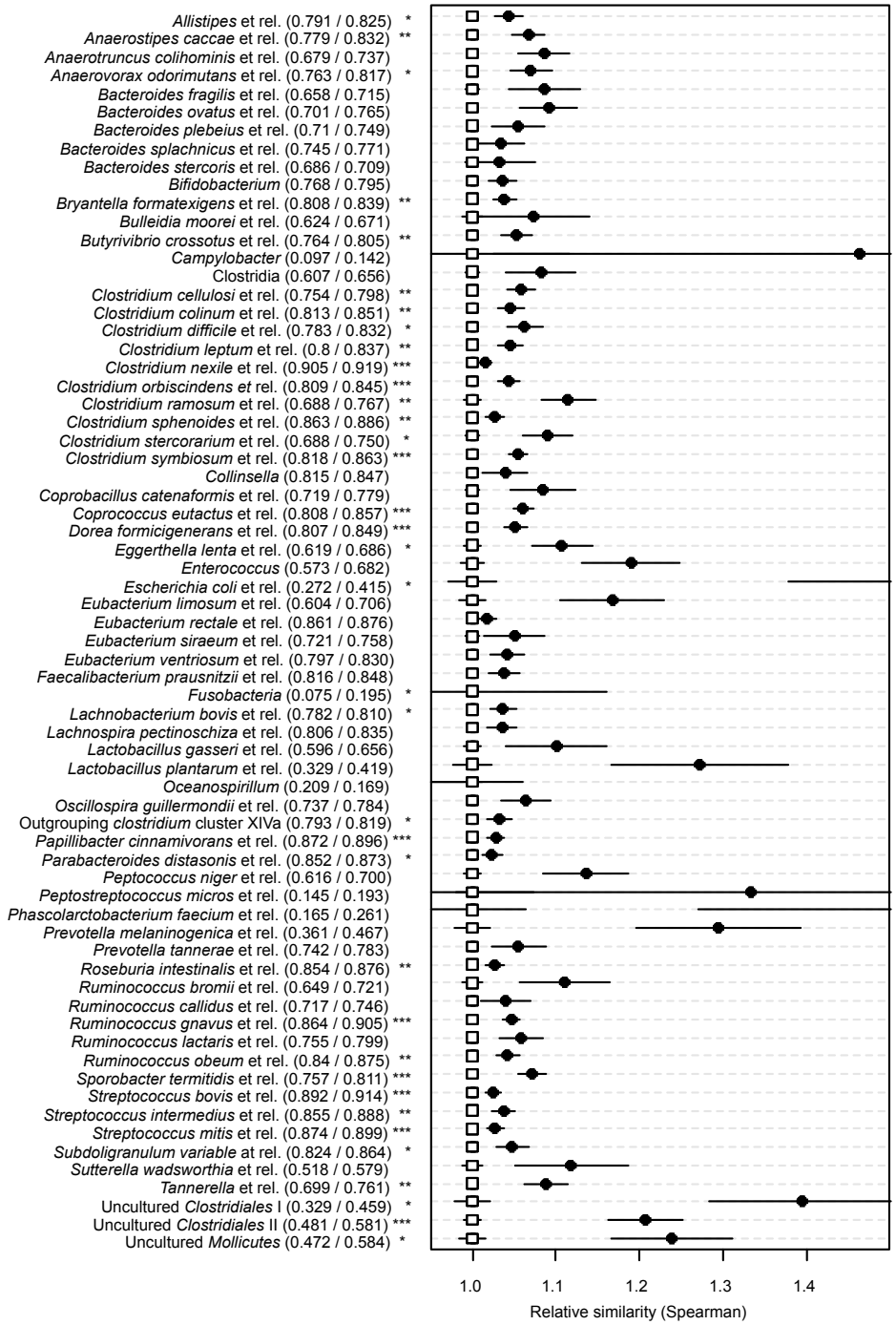


Figure S3.2 Gastrointestinal microbiota similarity in monozygotic twins of genus-like (level 2) groups. Dot-plots of the mean total microbiota profile similarity and the mean phylogenetic level 2 group similarity between monozygotic twins. Only level 2 groups with more than 20 probes are shown here. Mean Spearman's correlation coefficients values, error bars, level of significance as in Figure S3.1.

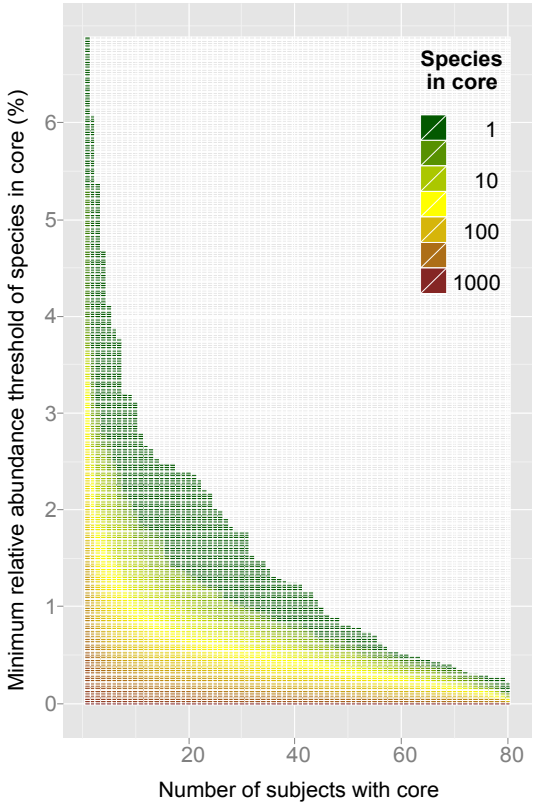


Figure S3.3 Number of core phylotypes. Core phylotypes were identified at a step-wise gradient of minimal relative abundances levels, in steps of 0.02%. Number of core phylotypes at varying prevalence and minimum phylotype abundance thresholds is visualized.

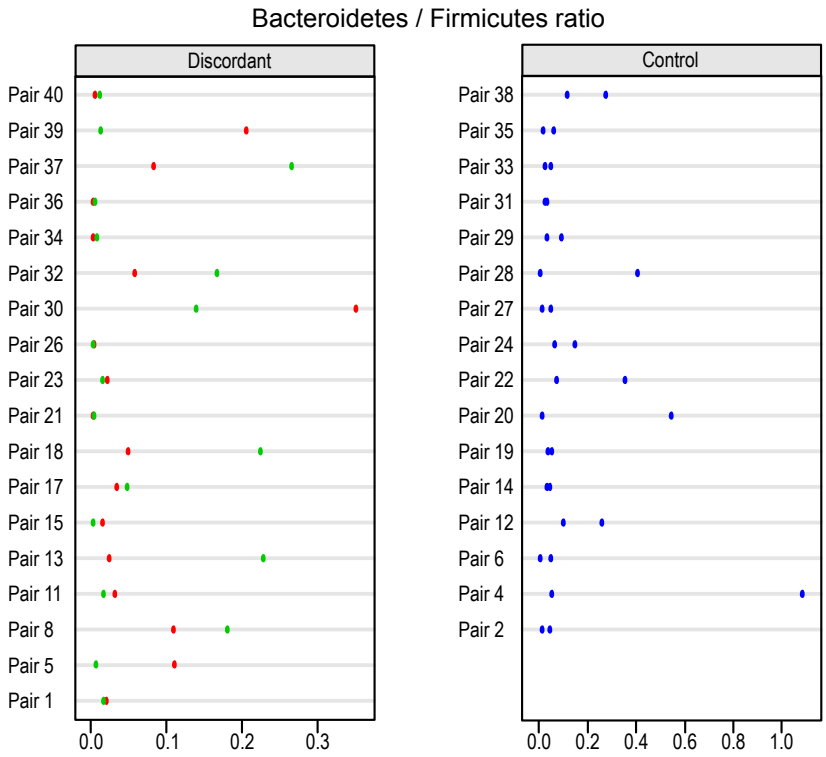


Figure S3.4 Bacteroidetes to Firmicutes (B:F) relative abundance ratios. The relative abundances of Bacteroidetes and Firmicutes were determined by dividing the sum of probe signals targeting these phyla. A. The results from the discordant twin pairs are visualized with red and green dots, representing the heavier and leaner sibling respectively. B. The results from the concordant control twins pairs are visualized with blue dots.

Table S3.1 Twin subject metadata. Gender (M = male, F = female), year of sampling, age at time of sampling (in years), length (in cm), weight (in kg), years apart, BMI, received breast feeding (0 = no, 1 = yes), obesitas family history (0 = no, 1 = yes), smoking (0 = no, 1 = yes), medication in last 4 weeks (0 = no, medication indicated otherwise). Twin subject classification and results. BMI classification and inverse Simpson index of diversity (1/D).

Twin	Gender	Year of sampling	Age	Length	Weight	Years apart	BMI	Breast fed?	Medication	BMI class	Diversity 1/D
01A	F	2008	35	160	63.5	10	24.8	0	probiotica	Higher BMI sibling	106.12
01B	F	2008	35	163	54	10	20.3	0	probiotica	Lower BMI sibling	212.83
02A	F	2009	35	167	59	12	21.2	0	0	Control	134.294
02B	F	2009	35	167	62	12	22.2	0	0	Control	132.412
03A	F	2009	35	163	60	14	22.6	0	0	Indistinct	99.424
03B	F	2009	35	158	52	14	20.8	0	0	Indistinct	44.497
04A	F	2009	35	170	57.5	10	19.9	0	0	Control	67.544
04B	F	2009	35	174	58	10	19.2	0	0	Control	58.412
05A	F	2008	33	164	84	12	31.2	0	0	Lower BMI sibling	221.307
05B	F	2008	33	162	91	12	34.7	0	0	Higher BMI sibling	182.36
06A	M	2008	32	172	68	11	23	0	probiotica	Control	120.298
06B	M	2008	32	172	68	11	23	0	0	Control	65.161
07A	F	2009	33	180	72	11	22.2	0	probiotica	Indistinct	80.58
07B	F	2009	33	174	62	11	20.5	0	probiotica	Indistinct	171.463
08A	F	2009	33	157	62	9	25.2	0	0	Higher BMI sibling	170.731
08B	F	2009	33	159	55	9	21.8	0	0	Lower BMI sibling	165.429

Table S3.1 (continued) Twin subject metadata.

Twin	Gender	Year of sampling	Age	Length	Weight	Years apart	BMI	Breast fed?	Medication	BMI class	Diversity 1/D
09A	F	2009	33	168	67	2	23.7	0	0	Indistinct	90.383
09B	F	2009	33	167	70	2	25.1	0	probiotica	Indistinct	107.585
10A	F	2009	33	168	66.5	11	23.6	0	0	Indistinct	57.957
10B	F	2009	33	157	52	11	21.1	0	0	Indistinct	50.353
11A	F	2009	33	170	71	14	24.6	1	probiotica	Higher BMI sibling	218.35
11B	F	2009	33	170	54	14	18.7	1	probiotica	Lower BMI sibling	166.612
12A	F	2009	32	173	56	9	18.7	1	probiotica	Control	44.495
12B	F	2009	32	177	58	9	18.5	1	probiotica	Control	120.051
13A	F	2009	32	167	68	7	24.4	0	0	Higher BMI sibling	175.257
13B	F	2009	32	168	60	7	21.3	0	0	Lower BMI sibling	172.732
14A	F	2009	32	165	60	8	22	0	0	Control	217.005
14B	F	2009	32	165	58.5	8	21.5	0	0	Control	154.848
15A	F	2008	30	164	57	8	21.2	0	0	Lower BMI sibling	119.964
15B	F	2008	30	165	67	8	24.6	0	0	Higher BMI sibling	111.274
16A	M	2009	30	187	100	8	28.6	0	0	Indistinct	96.452
16B	M	2009	30	187	95	8	27.2	0	0	Indistinct	54.083

Table S3.1 (continued) Twin subject metadata.

Twin	Gender	Year of sampling	Age	Length	Weight	Years apart	BMI	Breast fed?	Medication	BMI class	Diversity 1/D
17A	F	2009	30	164	73	5	27.1	0	0	Higher BMI sibling	49.661
17B	F	2009	30	165	65	5	23.9	0	probiotica	Lower BMI sibling	101.97
18A	F	2008	28	163	57	7	21.5	1	0	Lower BMI sibling	151.162
18B	F	2008	28	161	72	7	27.8	1	0	Higher BMI sibling	154.395
19A	F	2009	43	176	97	18	31.3	0	0	Control	100.754
19B	F	2009	43	174	96	18	31.7	0	0	Control	98.806
20A	F	2009	27	164	57	4	21.2	1	0	Control	148.109
20B	F	2009	27	165	57	4	20.9	1	0	Control	96.569
21A	F	2008	25	178	75	0	23.7	0	probiotica	Higher BMI sibling	104.231
21B	F	2008	25	175	62	0	20.2	0	probiotica	Lower BMI sibling	115.549
22A	F	2009	25	167	53	2	19	1	probiotica	Control	239.705
22B	F	2009	25	165	53	2	19.5	1	probiotica	Control	140.031
23A	M	2008	24	180	76	0	23.5	0	0	Higher BMI sibling	116.013
23B	M	2008	24	175	62	0	20.2	0	0	Lower BMI sibling	177.011
24A	M	2008	24	177	70	0	22.3	0	0	Control	148.281
24B	M	2008	24	178	74	0	23.4	0	0	Control	222.273

Table S3.1 (continued) Twin subject metadata.

Twin	Gender	Year of sampling	Age	Length	Weight	Years apart	BMI	Breast fed?	Medication	BMI class	Diversity 1/D
25A	M	2008	22	188	82	0	23.2	0	0	Indistinct	133.996
25B	M	2008	22	187	73	0	20.9	0	0	Indistinct	112.939
26A	F	2008	22	168	55	1	19.5	0	0	Lower BMI sibling	33.532
26B	F	2008	22	168	69	1	24.4	0	probiotica	Higher BMI sibling	91.257
27A	F	2008	21	171	58	0	19.8	1	0	Control	159.644
27B	F	2008	21	168	56	0	19.8	1	0	Control	187.998
28A	M	2009	23	187	85	0	24.3	0	0	Control	87.313
28B	M	2009	23	186	82	0	23.7	0	probiotica	Control	68.805
29A	F	2009	20	173	71	0	23.7	1	probiotica	Control	133.115
29B	F	2009	20	173	67.5	0	22.6	1	0	Control	100.279
30A	M	2008	19	194	78	0	20.7	1	0	Lower BMI sibling	276.668
30B	M	2008	19	191	95	0	26	1	0	Higher BMI sibling	127.63
31A	F	2008	41	167	62	18	22.2	0	probiotica	Control	206.721
31B	F	2008	41	167	62	18	22.2	0	probiotica	Control	219.692
32A	F	2008	43	172	74	33	25.0	0	0	Lower BMI sibling	132.388
32B	F	2008	43	172	84	33	28.4	0	0	Higher BMI sibling	106.323
33A	M	2008	43	180	70	21	21.6	0	probiotica	Control	203.366
33B	M	2008	43	180	74	21	22.8	0	0	Control	210.752

Table S3.1 (continued) Twin subject metadata.

Twin	Gender	Year of sampling	Age	Length	Weight	Years apart	BMI	Breast fed?	Medication	BMI class	Diversity 1/D
34A	M	2008	43	182	115	20	34.7	0	anti-diaretics	Higher BMI sibling	128.046
34B	M	2008	43	182	79	20	23.8	0	0	Lower BMI sibling	114.504
35A	F	2008	37	176	64	14	20.7	0	0	Control	171.271
35B	F	2008	37	176	66	14	21.3	0	probiotica	Control	47.969
36A	F	2008	37	164	73	14	27.1	0	laxatives	Higher BMI sibling	100.429
36B	F	2008	37	167	63	14	22.6	0	probiotica	Lower BMI sibling	77.747
37A	M	2009	38	176	76	10	24.5	0	0	Lower BMI sibling	135.018
37B	M	2009	38	181	90	10	27.5	0	0	Higher BMI sibling	146.948
38A	F	2009	37	160	53.5	15	20.9	0	probiotica	Control	153.95
38B	F	2009	37	161	54.6	15	21.1	0	probiotica	Control	174.2
39A	F	2008	36	165	66.5	13	24.4	0	0	Higher BMI sibling	176.316
39B	F	2008	36	162	53	13	20.2	0	0	Lower BMI sibling	143.408
40A	M	2008	19	187	89	0	25.5	0	0	Higher BMI sibling	100.618
40B	M	2008	19	184	73	0	21.6	0	0	Lower BMI sibling	115.271

Table S3.2 HITChip level 2 taxonomic groups targeted by more than 20 probes which have a significantly higher (p < 0,05) Spearman's correlation coefficient pairwise compared to the total microbiota profile similarity. Enterotypes are abbreviated as followed: ET1 = Enterotype 1, ET2 = Enterotype 2, and ET3 = Enterotype 3.

Level 1	Level 2 group	Nr of probes	Enterotype driver	Average Spearman's correlation coefficient	Average relative abundance (%)	Adjusted p-value
Actinobacteria	*Collinsella*	24	ET3	0.847	0.465	0.003
Bacilli	*Streptococcus bovis* et rel.	52	ET2	0.914	1.470	<0.001
	Streptococcus intermedius et rel.	34	ET2	0.888	0.645	<0.001
	Streptococcus mitis et rel.	43	ET2	0.899	1.310	<0.001
Bacteroidetes	Allistipes et rel.	74	ET3	0.825	1.195	<0.001
	Parabacteroides distasonis et rel.	45	ET1	0.873	0.763	<0.001
Clostridium cluster IV	*Clostridium cellulosi* et rel.	78	-	0.798	2.365	0.029
	Clostridium leptum et rel.	61	-	0.837	1.765	<0.001
	Clostridium orbiscindens et rel.	104	-	0.845	2.346	<0.001
	Faecalibacterium prausnitzii et rel.	81	ET1	0.848	5.395	<0.001
	Papillibacter cinnamivorans et rel.	36	-	0.896	1.537	<0.001
	Sporobacter termitidis et rel.	93	-	0.811	1.753	<0.001
	Subdoligranulum variable at rel.	51	ET3	0.864	3.656	<0.001
Clostridium cluster XI	*Anaerovorax odorimutans* et rel.	35	-	0.817	0.507	0.022
	Clostridium difficile et rel.	45	-	0.832	1.240	0.008
Clostridium cluster XIVa	*Anaerostipes caccae* et rel.	33	ET1	0.832	1.249	0.003
	Bryantella formatexigens et rel.	110	-	0.839	2.905	<0.001
	Butyrivibrio crossotus et rel.	88	-	0.805	1.835	0.001
	Clostridium colinum et rel.	33	-	0.851	1.432	<0.001
	Clostridium nexile et rel.	24	-	0.919	2.405	<0.001

Table S3.2 (continued) HITChip level 2 taxonomic groups targeted by more than 20 probes which have a significantly higher (p < 0,05) Spearman's correlation coefficient pairwise compared to the total microbiota profile similarity. Enterotypes are abbreviated as followed: ET1 = Enterotype 1, ET2 = Enterotype 2, and ET3 = Enterotype 3.

Level 1	Level 2 group	Nr of probes	Enterotype driver	Average Spearman's correlation coefficient	Average relative abundance (%)	Adjusted p-value
Clostridium cluster XIVa	*Clostridium sphenoides* et rel.	68	-	0.886	3.051	<0.001
	Clostridium symbiosum et rel.	135	-	0.863	4.188	<0.001
	Coprococcus eutactus et rel.	97	ET3	0.857	3.987	<0.001
	Dorea formicigenerans et rel.	96	ET3	0.849	4.230	<0.001
	Eubacterium rectale et rel.	56	ET3	0.876	3.811	<0.001
	Clostridium colinum et rel.	33	-	0.851	1.432	<0.001
	Clostridium nexile et rel.	24	-	0.919	2.405	<0.001
	Eubacterium ventriosum et rel.	22	ET3	0.830	1.066	0.002
	Lachnobacterium bovis et rel.	41	-	0.810	1.340	0.03
	Lachnospira pectinoschiza et rel.	83	ET2	0.835	4.163	0.001
	Outgrouping *Clostridium* cluster XIVa	79	-	0.819	1.877	<0.001
	Roseburia intestinalis et rel.	29	ET1	0.876	1.550	<0.001
	Ruminococcus gnavus et rel.	43	ET3	0.905	2.991	<0.001
	Ruminococcus obeum et rel.	158	ET3	0.875	7.019	<0.001

Table S3.3 Core phylotypes at different minimal relative abundances levels (core phylotypes were identified at a stepwise relative abundance gradient, in steps of 0,02%).

	Relative abundance threshold (%)	Number of core phylotype-like groups
Anaerofustis stercorihominis	0.02	
Aneurinibacillus aneurinolyticus	0.02	
bacterium A21	0.02	
bacterium adhufec13	0.02	
bacterium adhufec250	0.02	
bacterium adhufec296	0.02	
bacterium adhufec68	0.02	
Bryantella formatexigens	0.02	
butyrate-producing bacterium SR1/1	0.02	
Clostridium leptum	0.02	
Dorea formicigenerans	0.02	
Dorea longicatena	0.02	
Eubacterium cellulosolvens	0.02	
Eubacterium rectale	0.02	
Lachnobacterium sp. wal 14165	0.02	
Roseburia intestinalis	0.02	
Ruminococcus obeum	0.02	
Streptococcus agalactiae	0.02	
Streptococcus bovis	0.02	96
Streptococcus equi subsp. *zooepidemicus*	0.02	
*Streptococcus equi*nus	0.02	
Streptococcus infantarius subsp. *coli*	0.02	
Streptococcus intermedius	0.02	
Streptococcus mutans	0.02	
Streptococcus pyogenes	0.02	
Streptococcus uberis	0.02	
Streptococcus viridans	0.02	
Subdoligranulum variabile	0.02	
uncultured bacterium C583	0.02	
Uncultured bacterium clone Eldhufec100	0.02	
Uncultured bacterium clone Eldhufec157	0.02	
Uncultured bacterium clone Eldhufec170	0.02	
Uncultured bacterium clone Eldhufec262	0.02	
Uncultured bacterium clone Eldhufec268	0.02	
Uncultured bacterium clone Eldhufec302	0.02	
uncultured bacterium D416	0.02	
uncultured bacterium HuCA22	0.02	

Table S3.3 (continued) Core phylotypes at different minimal relative abundances levels (core phylotypes were identified at a step-wise relative abundance gradient, in steps of 0,02%).

	Relative abundance threshold (%)	Number of core phylotype-like groups
uncultured bacterium HuCA28	0.02	
uncultured bacterium HuCA8	0.02	
uncultured bacterium HuCB21	0.02	
uncultured bacterium HuCB5	0.02	
uncultured bacterium HuCC15	0.02	
uncultured bacterium HuCC43	0.02	
uncultured bacterium HuRC12	0.02	
uncultured bacterium inhufecA-27	0.02	
uncultured bacterium K094	0.02	
uncultured bacterium LC79	0.02	
uncultured bacterium M431	0.02	
uncultured bacterium MB66	0.02	
uncultured bacterium MD61	0.02	
uncultured bacterium MI29	0.02	
uncultured bacterium MN45	0.02	
uncultured bacterium N322	0.02	
uncultured bacterium NL49	0.02	
uncultured bacterium NQ96	0.02	
Uncultured bacterium UC7-7	0.02	
bacterium A57	0.04	
bacterium adhufec295	0.04	
bacterium adhufec57	0.04	
butyrate-producing bacterium L2-10	0.04	
butyrate-producing bacterium T2-132	0.04	
Streptococcus constellatus	0.04	
Streptococcus salivarius	0.04	
Uncultured bacterium clone Eldhufec185	0.04	
uncultured bacterium E177	0.04	
uncultured bacterium HuCA17	0.04	40
uncultured bacterium HuCB12	0.04	
uncultured bacterium HuCC34	0.04	
uncultured bacterium KS90	0.04	
uncultured bacterium KZ22	0.04	
uncultured bacterium L068	0.04	
uncultured bacterium MO17	0.04	
Uncultured bacterium UC7-35	0.04	
Uncultured bacterium UC7-36	0.04	

Table S3.3 (continued) Core phylotypes at different minimal relative abundances levels (core phylotypes were identified at a step-wise relative abundance gradient, in steps of 0,02%).

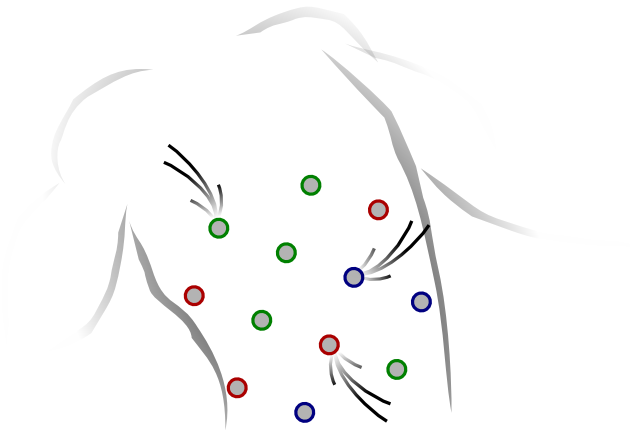
	Relative abundance threshold (%)	Number of core phylotype-like groups
bacterium A54	0.06	
bacterium adhufec25	0.06	
Clostridium aminovalericum	0.06	
Streptococcus mitis	0.06	
Streptococcus parasanguinis	0.06	
Streptococcus pneumoniae	0.06	
Streptococcus sanguis	0.06	
uncultured bacterium A20	0.06	22
Uncultured bacterium clone Eldhufec267	0.06	
uncultured bacterium HuCA2	0.06	
uncultured bacterium K375	0.06	
uncultured bacterium M510	0.06	
uncultured bacterium ME10	0.06	
uncultured bacterium NP09	0.06	
uncultured bacterium OLDB-D1	0.06	
Eubacterium hallii	0.08	
uncultured bacterium HuCB26	0.08	7
uncultured bacterium K379	0.08	
Ruminococcus gnavus	0.1	
uncultured bacterium cadhufec20a04	0.1	4
uncultured bacterium HuCA15	0.1	
Uncultured bacterium UC7-62	0.22	1

References

1. **Savage DC.** Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 1977;31:107-33.
2. **Yang X, Xie L, Li Y, et al.** More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body. PLoS One 2009;4(6):e6074.
3. **Wei C and Brent MR.** Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 2006;7:327.
4. **Cummings JH and Macfarlane GT.** Colonic microflora: nutrition and health. Nutrition 1997;13(5):476-8.
5. **Metges CC.** Contribution of Microbial Amino Acids to Amino Acid Homeostasis of the Host. J. Nutr. 2000;130(7):1857S-1864.
6. **Ramotar K, Conly JM, Chubb H, et al.** Production of menaquinones by intestinal anaerobes. J Infect Dis 1984;150(2):213-8.
7. **Albert MJ, Mathan VI, and Baker SJ.** Vitamin B12 synthesis by human small intestinal bacteria. Nature 1980;283(5749):781-2.
8. **Begley M, Gahan CG, and Hill C.** The interaction between bacteria and bile. FEMS Microbiol Rev 2005;29(4):625-51.
9. **Adlercreutz H, Pulkkinen MO, Hamalainen EK, et al.** Studies on the role of intestinal bacteria in metabolism of synthetic and natural steroid hormones. J Steroid Biochem 1984;20(1):217-29.
10. **Cebra JJ.** Influences of microbiota on intestinal immune system development. Am J Clin Nutr 1999;69(5):1046S-1051S.
11. **Shanahan F.** The host-microbe interface within the gut. Best Practice & Research Clinical Gastroenterology 2002;16(6):915-931.
12. **O'Hara AM, O'Regan P, Fanning A, et al.** Functional modulation of human intestinal epithelial cell responses by Bifidobacterium infantis and Lactobacillus salivarius. Immunology 2006;118(2):202-15.
13. **Zoetendal EG, Cheng B, Koike S, et al.** Molecular microbial ecology of the gastrointestinal tract: from phylogeny to function. Curr Issues Intest Microbiol 2004;5(2):31-47.
14. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009.
15. **Jalanka-Tuovinen J, Salonen A, Nikkila J, et al.** Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One 2011;6(7):e23035.
16. **Seksik P, Rigottier-Gois L, Gramet G, et al.** Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. Gut 2003;52(2):237-42.
17. **Backhed F, Ding H, Wang T, et al.** The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 2004;101(44):15718-23.
18. **Backhed F, Manchester JK, Semenkovich CF, et al.** Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U S A 2007;104(3):979-84.
19. **Ley RE, Turnbaugh PJ, Klein S, et al.** Microbial ecology: human gut microbes associated with obesity. Nature 2006;444(7122):1022-3.
20. **Samuel BS, Shaito A, Motoike T, et al.** Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. Proc Natl Acad Sci U S A 2008;105(43):16767-72.
21. **Turnbaugh PJ, Ley RE, Mahowald MA, et al.** An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 2006;444(7122):1027-31.
22. **Arumugam M, Raes J, Pelletier E, et al.** Enterotypes of the human gut microbiome. Nature 2011;473(7346):174-80.
23. **Wu GD, Chen J, Hoffmann C, et al.** Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science 2011.
24. **Huse SM, Ye Y, Zhou Y, et al.** A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. PLoS One 2012;7(6):e34242.
25. **Bouchard C, Tremblay A, Despres JP, et al.** The response to long-term overfeeding in identical twins. N Engl J Med 1990;322(21):1477-82.
26. **Naukkarinen J, Surakka I, Pietilainen KH, et al.** Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes. PLoS genetics 2010;6(6):e1000976.
27. **Pietilainen KH, Naukkarinen J, Rissanen A, et al.** Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. PLoS medicine 2008;5(3):e51.
28. **Saris WH and Tarnopolsky MA.** Controlling food intake and energy balance: which macronutrient should we select? Curr Opin Clin Nutr Metab Care 2003;6(6):609-13.

29. **Heid IM, Jackson AU, Randall JC, et al.** Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. 2011;43(11):1164-1164.
30. **Ley RE, Backhed F, Turnbaugh P, et al.** Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 2005;102(31):11070-5.
31. **Duncan SH, Lobley GE, Holtrop G, et al.** Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes (Lond) 2008;32(11):1720-4.
32. **Zhang H, DiBaise JK, Zuccolo A, et al.** Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A 2009;106(7):2365-70.
33. **Schwiertz A, Taras D, Schafer K, et al.** Microbiota and SCFA in Lean and Overweight Healthy Subjects. Obesity (Silver Spring) 2009.
34. **Manson JE, Willett WC, Stampfer MJ, et al.** Body Weight and Mortality among Women. New England Journal of Medicine 1995;333(11):677-685.
35. **Stevens J, Cai J, Pamuk ER, et al.** The Effect of Age on the Association between Body-Mass Index and Mortality. New England Journal of Medicine 1998;338(1):1-7.
36. **Visser M, Bouter LM, McQuillan GM, et al.** ELevated c-reactive protein levels in overweight and obese adults. JAMA: The Journal of the American Medical Association 1999;282(22):2131-2135.
37. **Benson AK, Kelly SA, Legge R, et al.** Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci U S A 2010;107(44):18933-8.
38. **Zoetendal EG, Akkermans ADL, Akkermans-van Vliet WM, et al.** The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. Microbial Ecology in Health and Disease 2001;13(3):129-134.
39. **Stewart JA, Chadwick VS, and Murray A.** Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. J Med Microbiol 2005;54(12):1239-1242.
40. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-4.
41. **Dicksved J, Halfvarson J, Rosenquist M, et al.** Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J 2008;2(7):716-27.
42. **Lepage P, Hasler R, Spehlmann ME, et al.** Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. Gastroenterology 2011;141(1):227-36.
43. **Willing BP, Dicksved J, Halfvarson J, et al.** A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology 2010;139(6):1844-1854 e1.
44. **Stevens J, Truesdale KP, McClain JE, et al.** The definition of weight maintenance. 2005;30(3):391-399.
45. **Salonen A, Nikkila J, Jalanka-Tuovinen J, et al.** Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods 2010;81(2):127-34.
46. **Booijink CC, Zoetendal EG, Kleerebezem M, et al.** Microbial communities in the human small intestine: coupling diversity to metagenomics. Future Microbiol 2007;2(3):285-95.
47. **Claesson MJ, O'Sullivan O, Wang Q, et al.** Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One 2009;4(8):e6669.
48. **van den Bogert B, de Vos WM, Zoetendal EG, et al.** Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. Appl Environ Microbiol 2011;77(6):2071-80.
49. **Stams AJ, Van Dijk JB, Dijkema C, et al.** Growth of syntrophic propionate-oxidizing bacteria with fumarate in the absence of methanogenic bacteria. Appl Environ Microbiol 1993;59(4):1114-9.
50. **Derom CA, Vlietinck RF, Thiery EW, et al.** The East Flanders Prospective Twin Survey (EFPTS). Twin Res Hum Genet 2006;9(6):733-8.
51. **He YL, Ding YF, and Long YQ.** Two cellulolytic Clostridium species: Clostridium cellulosi sp. nov. and Clostridium cellulofermentans sp. nov. Int J Syst Bacteriol 1991;41(2):306-9.
52. **Klieve AV, O'Leary MN, McMillen L, et al.** Ruminococcus bromii, identification and isolation as a dominant community member in the rumen of cattle fed a barley diet. Journal of Applied Microbiology 2007;103(6):2065-2073.
53. **Kovatcheva-Datchary P, Egert M, Maathuis A, et al.** Linking phylogenetic identities of bacteria to starch fermentation in an in vitro model of the large intestine by RNA-based stable isotope probing. Environ Microbiol 2009;11(4):914-926.
54. **Grech-Mora I, Fardeau M-L, Patel BKC, et al.** Isolation and Characterization of Sporobacter termitidis gen. nov., sp. nov., from the Digestive Tract of the Wood-Feeding Termite Nasutitermes lujae. Int J Syst Bacteriol 1996;46(2):512-518.

55. **Pryde SE, Duncan SH, Hold GL, et al.** The microbiology of butyrate formation in the human colon. FEMS Microbiol Lett 2002;217(2):133-9.
56. **Barcenilla A, Pryde SE, Martin JC, et al.** Phylogenetic relationships of butyrate-producing bacteria from the human gut. Appl Environ Microbiol 2000;66(4):1654-61.
57. **Duncan SH, Hold GL, Barcenilla A, et al.** Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. Int J Syst Evol Microbiol 2002;52(Pt 5):1615-20.
58. **Qin J, Li R, Raes J, et al.** A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59-65.
59. **Tap J, Mondot S, Levenez F, et al.** Towards the human intestinal microbiota phylogenetic core. Environ Microbiol 2009;11(10):2574-84.
60. **Claesson MJ, Cusack S, O'Sullivan O, et al.** Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proc Natl Acad Sci U S A 2011;108 Suppl 1:4586-91.
61. **Hamady M and Knight R.** Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 2009;19(7):1141-52.
62. **Salonen A, Salojarvi J, Lahti L, et al.** The adult intestinal core microbiota is determined by analysis depth and health status. Clin Microbiol Infect 2012;18 Suppl 4:16-20.
63. **Van de Merwe JP, Stegeman JH, and Hazenberg MP.** The resident faecal flora is determined by genetic characteristics of the host. Implications for Crohn's disease? Antonie Van Leeuwenhoek 1983;49(2):119-24.
64. **Manichanh C, Rigottier-Gois L, Bonnaud E, et al.** Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006;55(2):205-11.
65. **Mackie RI, Aminov RI, Hu W, et al.** Ecology of uncultivated Oscillospira species in the rumen of cattle, sheep, and reindeer as assessed by microscopy and molecular approaches. Appl Environ Microbiol 2003;69(11):6808-15.
66. **Matthies C, Evers S, Ludwig W, et al.** Anaerovorax odorimutans gen. nov., sp. nov., a putrescine-fermenting, strictly anaerobic bacterium. International journal of systematic and evolutionary microbiology 2000;50 Pt 4:1591-4.
67. **Amos HE, Little CO, and Mitchell GE, Jr.** Proline utilization during cellulose fermentation by rumen microorganisms. J Agric Food Chem 1971;19(1):112-5.



Chapter 2

Host genotype and the effect on microbial communities

Sebastian Tims, Erwin G. Zoetendal,
Willem M. de Vos, Michiel Kleerebezem

Metagenomics of the Human Body, 2011, K.E. Nelson (ed.), p. 15-41 (minor modifications)

Abstract

The microbial ecosystems found along the body surfaces of mammals have provide a variety of complementary metabolic functions to their hosts. It is likely that the mammalian host and its microbiota form a coalition of cells, or a so-called "super-organism", which mutually strives for survival. Unfortunately, the exact interactions between host and microbiota are for the most part unexplored. Our current understanding of host-microbe interactions mostly comes from studies on the gastrointestinal tract microbiota which is the most densely populated microbial ecosystem of the mammalian host. Although mammalian host genes are greatly outnumbered by the total gene pool of their microbiota, there are several indications that host genotype is an important factor affecting the diversity and function of the microbiota. Communication between host cells and microbes likely to dependent on host-immune system related genes and can be therefore be influenced by polymorphisms in these genes. However, there are probably more genes which are important for host-microbe interaction that are not directly related to the immune system. Future studies should focus on the hierarchy in importance of host genotypes with relation to host-microbe interactions. Complicating the studies on host-microbe interactions are environmental factors which can, sometimes drastically, influence both the host and its microbiota. Especially dietary influences should be taken into account while analyzing the interaction between the microbial communities of the gut and the host.

Introduction

Microbial communities inhabit a variety of body surfaces of mammalian hosts. In the human mouth, 10^2–10^3 different species from nine bacterial and a single archaeal phyla have been found [1]. Teeth, cheek, and tongue all have their own specific communities with anaerobic bacteria present at the gum-line and between the teeth. For each of these sites, the selective trait is based on the surface adherence capabilities of the microbes, typically resulting in multispecies-biofilm formation [1]. Microbial diversity and abundance normally decrease further down the gastrointestinal (GI) tract until the stomach. In the esophagus approximately 100 species from six phyla are found, most of which are similar to the species found in the mouth [2]. The stomach is generally regarded derelict for any microbial species except for *Helicobacter pylori* [3]. Nevertheless, 16S rRNA gene surveys have reported up to 128 species from eight phyla in the stomach; however, it seems likely that these findings represent remnants from ingested strains rather than true residents [4]. After the stomach, bacterial populations increase again in the small intestine, ranging from 10^4–10^5 g^{-1} in the duodenum and jejunum to 10^7 g^{-1} in the terminal ileum. In this region, intestinal transit slows down and the microbiota composition changes, favoring the more anaerobic species [3, 5]. Next, along the GI tract is the colon. In the ascending colon, polysaccharide hydrolysis and carbohydrate fermentation support rapid microbial growth, whereas in the transverse and descending colon, amino acid and host-derived glycans (mucin) fermentation occurs, coinciding with a reduction of bacterial growth rate [6]. The fermentations along the entire colon cause a microbial population increase up to 10^{11}–10^{12} cells g^{-1} in feces accompanied by a strong proportional decrease of facultative anaerobes [7]. In the colon, the most numerous species are obligate anaerobes belonging to the phyla Bacteroidetes and Firmicutes [8, 9].

No other body site attains the high bacterial abundance as observed in the colon, although the mouth harbors a taxonomic richness approximating that of the colon. Furthermore, the recent expansion of the human skin ribosomal operon database indicates a diversity level close to that of the GI tract as well [10]. Unfortunately, interactions between host and skin microbiota are for the most part unexplored. Next to skin and GI tract, the urinary tract is being studied on a microbial level as well. The vaginal microbiota is generally assumed to have low diversity and to be dominated by lactobacilli. However, lactobacilli are not dominant vaginal inhabitants in all healthy women [11]. One can conclude that the GI microbiota is the most densely populated microbial ecosystem of the mammalian host and has been subjected to many studies. Hence the remainder of this chapter will mainly focus on the knowledge obtained from GI ecosystems. A variety of functional metabolic activities are generally thought to derive from most, if not all, of the resident communities. Several general processes necessitate the presence of microbial inhabitants in order to function properly, such as the maturation of the immune system, resistance to pathogens, digestion of nutritional components, and the production of essential nutrients. Especially important to the host are microbiota-derived nutrient conversions and contributions that cannot be executed by the host itself, including the degradation of complex polymers in the GI tract that cannot be (completely) digested by the host's enzyme machinery (non-digestible carbohydrates, proteins, and lipids) [12]. Besides the improvement of our nutritional access to

complex nutrients, microorganisms in the GI tract provide a significant supply of essential amino acids [13] and other important compounds such as various vitamins, including vitamin K and several B vitamins [14, 15]. Conversion of several bio-active molecules, such as sex steroid hormones in order to promote their circulation, is another health-related function humans receive from their GI inhabitants [16, 17]. These types of microbial functions complement the metabolic potential of the host, as the host itself does not encode for the required proteins. In addition to the beneficial contributions, the microbial communities can introduce metabolic activities that are detrimental for their host, such as the production of hydrogen sulfide [18, 19] and the potentially tumor-promoting secondary bile acids [20].

It is becoming a generally accepted view that multicellular organisms, especially mammals, should not be considered as autonomously living entities. "Super-organism" is a popular term that better describes mammals for what they really are: a cohort of host cells and microbial cells. This coalition of cells from the different domains of life is striving for the common mutual cause of survival. The microbiota is a commonly used term for the composition of all microbial cells belonging to a super-organism. According to the super-organism concept, metagenomics can be defined as the mammalian host genes combined with the genes of the entire microbiota [21]. Currently, there are no completed metagenomes available for any super-organism. Hence, it remains difficult to establish or estimate the importance of the host genome. It is obvious that many organisms, such as humans, contribute far less genes to their metagenome than their microbial counterparts. In the human GI tract alone there are already ten times more microbial cells present than host cells in the entire human body [22]. Humans are believed to contain approximately 23,000 genes in their genomes [23], whereas current estimations for the GI microbiota unique gene count are up to 9,000,000 [24]. In other words, our human genes are outnumbered by several orders of magnitude by the GI microbiota alone! As more and more body sites are being sampled, this difference can only increase in favor of our microbiota. Such observations underline the limited human genetic input in the whole "super-organism" and pose the question: To what extent our genes matter during our life as a super-organism? Maybe, we as hosts have predominantly lost functions during evolution because our microbes provided them and could execute the corresponding functions more efficiently? Possibly, our evolutionary efficiency is increased by encoding a "limited" gene set, which could be specialized in molecular communication with microbes in order to recruit, nourish, and maintain a microbiota that is able to complement for the essential functions that are lacking in our own genomes? After all, we are still around despite our "limited" genotype.

Interactions in a super-organism

In mammals, a dynamic and complex relationship exists among diet, host phenotype, and the associated GI microbiota. All interactions are dependent on the host genotype, which can be seen as a matrix on which host phenotype and the resident microbiota are projected (Figure 2.1). Diet and transient food organisms are external, yet important, components that complete the complex host–microbiota interactions.

Classical views on many disorder-associated phenotypes do not take into account all underlying factors. For example, the main focus usually lies on diet and host genotype in disorders such as obesity, diabetes, and many cardiovascular diseases. However, the GI microbiota should not be excluded in studies or treatment of these disorders, since changes in the gut communities have been associated with some of these disorders [21, 25-28]. Even though it is hard to determine the causality of observed microbiota deviations with respect to these diseases and disorders, several studies do suggest causal contributions from the gut communities. Microbiota transplants from obese mice to germ-free (GF) littermates induce the obese fat-storage phenotype [21]. Furthermore, already before the onset of type-1 diabetes in genetically predisposed rat models, which are on the same diet, the gut microbiota was different in the rats that eventually developed diabetes compared to those that did not [29]. Moreover, antibiotic treatment of these rats significantly delayed and lowered the incidence of diabetes development [29]. These findings suggest a causal role of the gut microbiota in the development of diabetes.

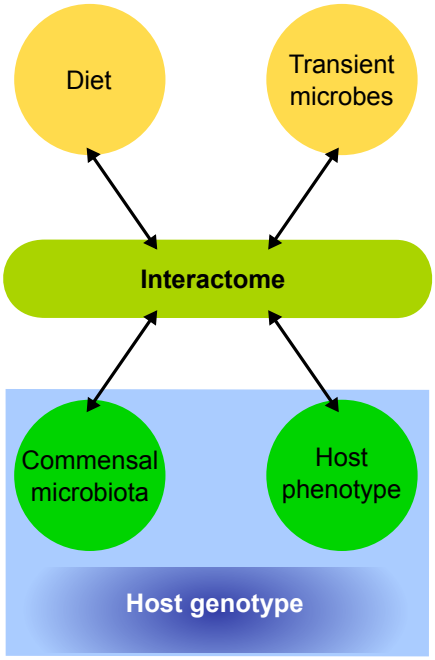


Figure 2.1 Factors involved in host–microbiota–diet interrelations.

Even in the extensively studied GI tract, the interactions between host and microbiota are not yet understood. Most studies to date are restricted to composition analyses. However, more and more insights are emerging. Especially the use of high-throughput technologies to study the diversity and functionality of the GI tract is greatly enhancing the current knowledge level. These technologies use a variety of approaches, such as the revival of culturing methods that are high-throughput [30, 31], metabolite detection [26], phylogenetic microarrays based on 16S rRNA gene sequences [32, 33], and sequencing of 16S rRNA genes as well as sequencing of random microbial DNA [21, 34]. Sequencing of the GI microorganisms has opened up the possibilities for functional metagenomics, which will allow further exploration of the microbial activity patterns.

It is generally accepted that the GI tract is sterile at birth and is swiftly colonized by microbes acquired from maternal and environmental sources. Recently, bacteria have been detected in the fluid in intact amniotic sacs of women in preterm labor [35]. This finding questions the broadly accepted view of postnatal GI tract colonization, since fetuses swallow and "inhale" amniotic fluid continuously, hence exposing their respiratory and GI tracts to everything that resides in it. However, bacteria were only found in 15% of the subjects (n = 166) [35]. Regardless of the precise moment of initial colonization, it remains without doubt that the human GI microbiota evolves over time. The development is especially drastic in the first 2 years of life, followed by

stabilization of the GI community into a microbiota that resembles that of an adult [36]. In adults the fecal microbiota is shown to be highly stable over time within one individual, as well as specific for its host [37]. Interestingly, despite a high variability in the GI species composition between individuals, functional capacity seems to be much more uniform between human adults [38]. Moreover, a significant proportion of microbial phylotypes found in the gut are continuously present during a 10-year timeframe [39]. Therefore it seems likely that adults have, next to some transient guest organisms, a stable individual core of permanent GI tract colonizers [39, 40]. Tap and coworkers have reported 66 microbial phylotypes, which are present in more than 50% of the samples they investigated (n = 17) [41]. Such findings suggest that besides an individual core, a limited number of microbial phylotypes are more prevalently found in people (>50% of the individuals) and appear to represent a common core of the human GI tract microbiota.

Host genotype and microbiota selection

As noted before, the mechanism(s) behind GI tract colonization, succession within the community, and community structure itself are poorly understood. One hypothesis is that colonization at weaning is determined by the primary nutrient foundation supplied by the host [42]. This may be true, but there are several indications that colonization is influenced by the host genotype as well. Mouse studies have revealed that the composition of fecal microbiota is affected by the major histocompatibility complex [43]. Furthermore, different mammalian host species develop a different make-up of their GI microbiota.

Additionally, studies with GF hosts that received inter-species GI microbiota transplants indicate that the hosts might be able to modulate their received microbial lineages toward a composition normally found in their non-GF, conventional status [44]. This can be attributed to obvious variables such as the type of food, the environment the host lives in, the nutritional requirements of the host, or more physiological aspects like host intestinal tract anatomy, body temperature, intestinal peristalsis, and residence time, etc. Intriguing results were obtained from the reciprocal GI microbiota-transplantation of zebra fish and mice, raised under GF conditions [44]. The GF hosts did not have any community legacy, and appeared to retain all intestinal species that were "given" to them during transplantation. Yet they reconstructed, in terms of relative abundance, the gut communities normally associated with conventionally raised animals [44], indicating that a powerful and poorly understood host-mediated mechanism must be in place to coordinate microbial community composition. It seems that zebra fish and mouse host genetics play a prominent role in the natural selection of gut inhabitants, although there are many confounding factors in the differences between these hosts, such as differences in body and environment temperatures, host habitat and activity, bowel anatomy and dimensions, residence time, and dietary intake.

Last but not least, in adult humans, the extent of the variation of the dominant bacteria is associated with the degrees of relatedness between the subjects [45, 46]. This family relatedness, especially with respect to twins, provides perhaps the most profound evidence of host genotypic influences on microbial communities, and will be discussed in more detail in the next sections.

In conclusion, the GI microbiota composition must be dependent on the host genotype, but the exact degree of this dependence remains to be elucidated.

The effect of host genetics on the gut microbiota is most profoundly observed in studies conducted on related individuals. Especially revealing were studies conducted with samples obtained from identical (monozygotic) and/or fraternal (dizygotic) twins. Already in 1983, indications were found that monozygotic twins have more similar fecal microbiota than dizygotic twins [46]. Although these findings were based on cultivation-dependent techniques, which gives an incomplete picture of the GI microbiota [40], this study provided a clear indication of host genetics and its influence on the fecal communities. Two decades later, Zoetendal and coworkers confirmed the significantly higher bacterial profile similarity in monozygotic twins with a culture-independent technique [45]. This study was performed 10 years ago using denaturing gradient gel electrophoresis (DGGE) on fecal bacterial 16S rRNA gene amplicons to assess the bacterial composition similarity. Samples in this study originated from human adults with varying degrees of genetic relatedness (ranging from parents and children, non-twin siblings to twin siblings). This study revealed a positive relationship between the DGGE profile similarity and the genetic relatedness of the subjects (Figure 2.2). Marital partners showed slightly higher similarities than unrelated individuals, but this was not found to be significant [45]. The latter is quite remarkable as marital partners essentially live in the same environment and generally have similar dietary habits. Overall, these results indicate that host-genotype factors indeed have a strong impact on the bacterial community in the adult GI tract.

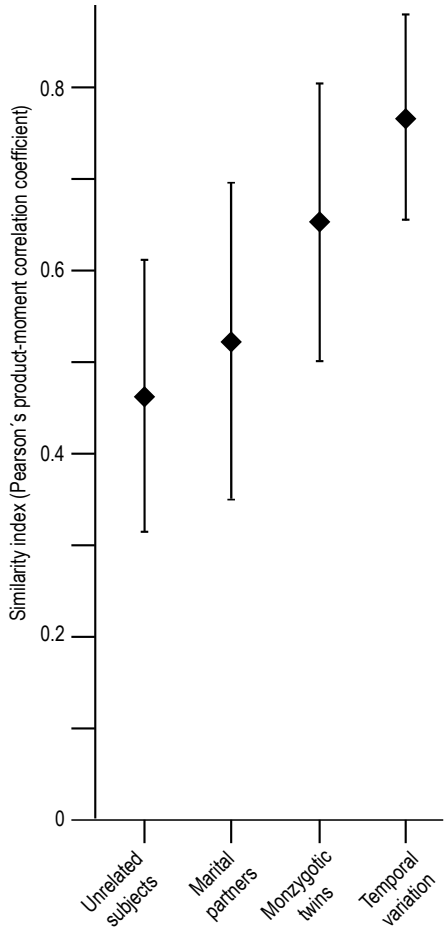


Figure 2.2 Plot of the similarity indices (Pearson's product-moment correlation coefficient) from unrelated subjects, marital partners, monozygotic twins, and temporal variation comparisons. The mean (diamonds) and standard deviation (black bars) are plotted. DGGE profile of the total bacterial community was used to calculate the similarities. Adapted from [45]

In a later study, a slightly different cultivation-independent technique, temporal temperature gradient gel electrophoresis (TTGE), was used to assess the influence of host genetics on the fecal microbiota composition in children [47]. TTGE profile similarity was again the lowest among unrelated children, higher between dizygotic twins, and clearly the highest between monozygotic twins [47].

The high-throughput cultivation-independent Human Intestinal Tract Chip (HITChip), a phylogenetic microarray developed by Rajilic and coworkers [33], was used to re-analyze the five monozygotic twin pair samples from Zoetendal and coworkers [45]. Average similarities between the twins was notably higher than the similarity between random unrelated individuals [39]. Even though only five twin pairs and five unrelated individuals were compared, the observed similarity difference was already borderline significant ($p = 0.067$) [39]. Thus the previous results obtained by DGGE were confirmed by phylogenetic microarray analysis. A recently conducted study on 40 monozygotic twin pairs showed that HITChip profiles were significantly ($p < 0.001$) more similar between the twins than between random unrelated subjects within this cohort (Figure 2.3, adapted from **Chapter 3**). Palmer and coworkers also used a phylogenetic microarray (different from the HITChip) to study GI microbiota development in human infants [32]. They included one dizygotic twin pair and this pair showed a more similar microbiota profile, at any stage of development, compared to the other 12 unrelated children [32].

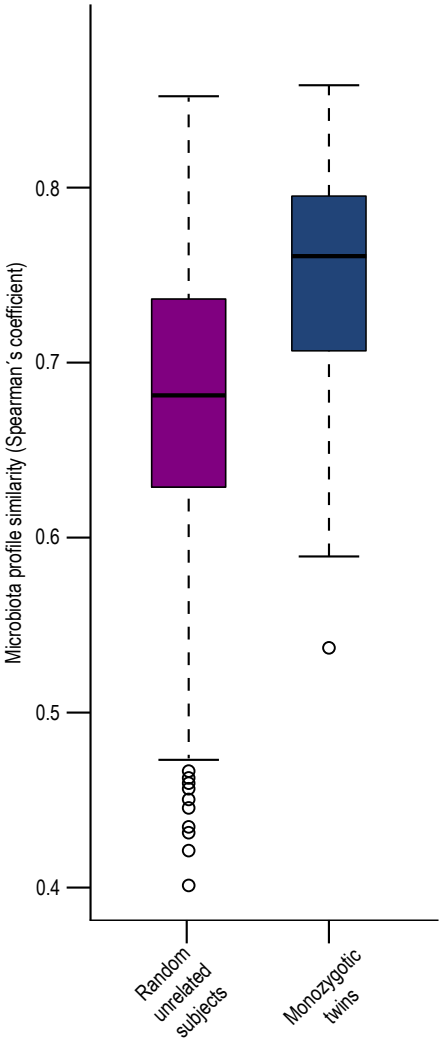


Figure 2.3 Box-whisker plots of the similarity between the total microbiota profiles expressed as Spearman's correlation coefficient. Purple box represents similarities between random unrelated subjects and the blue box represents similarities between monozygotic twins. Similarities were calculated with total microbiota HITChip profiles (adapted from Chapter 3).

Recent developments in sequencing technologies and corresponding reductions of sequencing costs have been of great importance for GI microbiota research. Turnbaugh and coworkers performed pyrosequencing on variable regions of the 16S rDNA on fecal microbial DNA extracts in a cohort of 154 subjects [34]. On average, nearly 4,000 16S V2 region sequences were obtained for all subjects and additionally nearly 25,000 sequences per sample were acquired for 33 subjects. The studied group was composed of 31 monozygotic twin pairs, 23 dizygotic twin pairs, and the mothers of 46 of the twins, and included two samples per individual collected with a 57-day interval. Subjects included in this cohort were differentiated based on concordant leanness or obesity among the twin pairs. Most of the twins (71%) did not live together anymore. Pyrosequencing confirmed the previous observation that each individual has a unique

GI microbiota composition and short-term changes are inferior to the inter-subject variations [34, 37]. Among all of the 154 individuals, no shared 16S rRNA gene-based phylotypes could be identified with an abundance of 0.5% or more. These results are in apparent contradiction with previous suggestions concerning the existence of a shared common core between humans, but different criteria were employed to define such a core community (>50% phylotype prevalence [41] versus 100% prevalence of phylotypes [34]). Nevertheless, subjects from the same family had more similar microbial community structures and shared significantly more phylotypes [34]. Neither the obesity status per individual nor distance between the family members' homes confounded the observed higher similarities for the families [34]. In contrast to earlier findings [47], the similarity between dizygotic twins were not lower than the similarities between the monozygotic twins [34]. Noteworthy is the fact that pyrosequencing still suffers from artifacts such as the formation of chimeric sequences during amplification and/or sequencing errors that are interpreted as distinct phylotypes. When such errors are not effectively removed they obviously influence the results obtained from sequencing-based studies. Perhaps future improvement of sequencing technology (in terms of reduced sequence error rates, advanced data analysis software suites, and/or extended sequence length) combined with increased sequencing depths may resolve the apparent contradictions in the current conclusions. Recently, Claesson and coworkers showed that pyrosequencing at the deepest sequencing-depth currently feasible, i.e., at approximately 400,000 sequences per sample, still does not capture the full microbial richness of GI tract samples [48].

A common approach in the GI tract studies is inferring the possible microbiota functionalities from the microbial lineages present as detected by 16S rRNA gene sequences. Current functional metagenomic study designs are moving toward random sequencing of as much microbial DNA as possible [40]. Through these random sequences, a more direct view on the functional repertoire of the microbiota can be obtained. Therefore, in addition to the 16S rRNA gene sequencing, Turnbaugh and coworkers also analyzed the samples of six families (n = 18) by random shotgun pyrosequencing [34]. In line with the 16S rRNA-based study, the profiles of the functional categories present in the gut communities were more similar between relatives. Interestingly, GI microbiota seemed to have even more similar functional profiles among all subjects despite their sometimes highly distinct microbiota composition profiles [34]. This raises the hypothesis that a core microbiota exists at a functional (and metabolic) level rather than at the level of microbial composition. In line with this hypothesis are results of earlier studies that determined *in situ* concentrations of short-chain fatty acids (SCFAs) in fecal material of different individuals. Different population survey data show the same fecal composition with respect to the ratios of the three main SCFAs acetate, propionate, and butyrate [49]. Dietary changes have been shown to modulate SCFA production and absorption somewhat, but the SCFA ratios are not drastically altered [50]. As an individual has its own unique GI microbiota composition, the fairly constant SCFA ratios are quite remarkable and indicate that the gut microbes perform similar overall functions. Even though a core GI microbiota is likely to exist more on a functional level it must be noted that the species composition, although hypervariable and seemingly chaotic, is not random. Grouping of mammals based on their gut community composition is associated with

their dietary needs, i.e., whether a mammal is a herbivore, carnivore, or omnivore [51]. This could indicate that when roughly similar digestive tasks are required, the GI community tends to be more similar at higher taxonomic levels [51]. However, the latter statement needs to be verified as the GI tract anatomy differs between individual mammalian species belonging to the carnivores, herbivores, and omnivores.

More comprehensive studies following both monozygotic and dizygotic twins from birth to adulthood can provide valuable information on interactions between host genotype and its inhabiting microbiota. In such studies diet should be taken into account, due to its drastic influences.

Dietary influences

Diet is a strongly confounding factor in the ambition to obtain insight in the exact roles of the host genotype and GI microbiota in a super-organism. Not only the GI community is influenced by the dietary influx of microbes and nutrients but also the host phenotype itself, mainly through regulation of gene expression and physiological adaptation. Several genes from a group of nuclear hormone receptors called peroxisome proliferator-activated receptors (PPARs) can be used to illustrate host adaptation to its diet on a transcriptional level. Unsaturated fatty acids are, among others, activating ligands for PPAR-γ [52], and thereby PPAR-γ activities are directly influenced by the diet. PPAR-γ acts as a regulator of a variety of relevant physiological processes, including transcription control of many genes involved in fat-cell differentiation, insulin sensitivity, and lipid homeostasis [53]. The host genotype, in terms of PPAR-γ gene variants, affects the extent to which the host reacts to its diet. For example, a PPAR-γ2 gene polymorphism [54] correlates with the inter-individual variability in serum triacyl–glycerol levels after administration of n–3 fatty acids [55]. This example is just one of the many available in literature, which indicates that although the host contributes with a marginal amount of genes to the metagenome, variations in the host genotype can still have drastic effects for the super-organism. Furthermore, this example indicates the major role of the host's diet as well. This raises the question of how important the diet exactly is.

Obesity in animal models

Obesity has gained popularity with respect to study of the importance of dietary influences on the gut microbiota and the host. The prime cause of obesity is an excessive caloric intake. Such a surplus intake disturbs the normal balance between the amount of energy harvested from the diet and the amount used by the host. This balance, or energy homeostasis, is at least partly defined by the GI microbiota. Studies with GF mice clearly show the microbial impact on host energy homeostasis. GF mice are resistant to obesity development when fed a "Western" (high-fat/high-sugar) diet [56], but GI tract colonization was stimulating weight gain in these mice [57]. Colonization of GF mice leads to an increased release of monosaccharides and SCFAs from complex dietary polysaccharides, enhanced conversion rates of fatty acids toward complex lipids in the liver, and by regulating host genes involved in storage of the converted lipids into adipocytes [58].

However, next to the GI microbiota, host-dependent factors are also required to develop obesity. For instance, mice lacking a functional copy of the *Gpr41* gene, a G-protein-coupled receptor that binds SCFAs, do not readily develop obesity [57]. Without Gpr41, the mice displayed an increased intestinal motility and decreased SCFA absorption [57]. On a side note, another G-protein-coupled receptor called Gpr43 displays a different role in mice [59]. Mouse models of colitis, arthritis, and asthma required stimulation of Gpr43 by SCFAs to counteract their inflammation [59]. Thus, besides metabolic regulation, immune and inflammatory responses are affected by the bacterial SCFA production as well.

Simply having a gut microbiota is not the only microbial factor affecting the energy balance in mice. The composition of the microbial community determines to what extent the microbiota improves energy harvest from food. Both GF mice inoculated with distal gut microbiota from conventionalized obese and lean animals resulted in an increase of bodyweight and total body fat [21]. Yet this increase was found to be significantly larger in the GF mice that had received the microbiota from the obese animals [21]. Correlations between phenotypic variations and attributes of the microbial gut community were reported after studies conducted with lean and genetically obese (*ob/ob*) mice [21, 60] as well as with lean and genetically obese rats [61]. Variations of the members belonging to the bacterial phyla Bacteroidetes, Firmicutes, and Actinobacteria appeared to be associated with leanness or obesity [21] [60, 61]. Especially the Bacteroidetes to Firmicutes ratio was found to be lower in the obese rodents.

Most animal models used to study obesity consist of specific gene knockout mutants, such as those focusing on the role of the *ob* gene in mice that predisposes them to develop obesity. The *ob* gene encodes for the protein hormone leptin, which regulates body weight, metabolism, and reproduction in mammals [62]. Both inactivating mutations in the leptin (*ob*) gene and in its receptor (*db*) gene lead to genetically obese mice [62]. Therefore, the host genotype is important in the development of obesity as well. Another study involving wild-type and *Apoa*-I knockout mice indicates that all three aspects are implicated in the development of obesity and metabolic syndrome (MS) [28]. *Apoa*-I knockout mice have an impaired glucose tolerance and high body-fat levels. Groups of wild-type and of *Apoa*-I knockouts were fed a high-fat diet and normal chow diet for 25 weeks. Diet as well as genetic mutation could explain 57 and 12% of the observed variation found in the GI microbiota communities, respectively [28]. The results of this study indicate a stronger, possibly dominating, role of the diet compared with the genetic variations of the host. Nevertheless, the influence of host genotype is not negligible and should therefore be taken into account in MS studies.

Obesity in humans

Extrapolation of the relations among diet, host genotype, and GI microbiota discovered in animal studies to human obesity seems sensible but has proven to be difficult. Nevertheless, deviating leptin concentrations are associated with obesity in humans as well [63]. Leptin normally suppresses hunger and increases metabolism, and it has been suggested that obese humans are insensitive to leptin. However, although several cases have been described [64], mutations in the human *db* gene are only rarely seen in obese people. However, to accurately assess the importance

of *db* gene variations with respect to obesity development risks, the db mutation frequencies in lean individuals should also investigated.

In contrast to mice studies, research with human subjects provide conflicting results regarding the association of obesity with relative abundances or abundance ratios of specific bacterial phyla [65-67]. However, human studies reported to date employed different molecular techniques and targeted populations of different geographic origin. Duncan and coworkers found no differences in bacterial phyla abundances or abundance ratios, but identified a significantly higher proportion of butyrate producers in the obese subjects [67]. This observation is in agreement with the finding of increased butyrate concentrations in the cecum of *ob/ob* mice as compared to their lean littermates [21]. The SCFA butyrate is mainly produced during carbohydrate fermentation in the colonic lumen, mainly by members of the Firmicutes phylum, especially those belonging to *Clostridium* cluster IV. Luminal butyrate is quickly absorbed by the colon mucosa where it serves as the main energy source for the colonocytes (colonic epithelial cells) [50, 68]. However, the exact physiological effects of butyrate are not fully understood. Different, but related cell-line models can yield direct opposite results regarding the role of butyrate in the modulation of cell proliferation, differentiation, and apoptosis [69, 70]. These conflicting results, commonly referred to as the "butyrate paradox," are extensively reviewed elsewhere [71]. In short, *in vivo* human data are insufficient but most studies support beneficial roles for butyrate, including the restraining of inflammation and carcinogenesis, reinforcement of various components of the mucosal barrier, lowering of colonic oxidative stress, and promotion of satiation [71]. Overall, it is obvious that the production of butyrate by the GI microbiota has a major influence on colonic mucosa. Thereby the differential abundance levels of butyrate-producing microbes, as reported by Duncan and coworkers, seem relevant with respect to human health and may be associated with energy homeostasis and obesity risk.

Diet: Transient or permanent effects?

One may conclude from the animal experiments and the observations in human volunteers that the interactive factors, constituted by dietary intake and gut microbial ecology, are of major importance for the host's well-being. This raises the question whether dietary effects should be seen as transient or can also generate permanent effects. The microbiota transplant approach in mice revealed that efficient energy harvesting traits are transferable by the GI microbiota [21]. In continuation of this approach, C57BL/6 J mice were conventionalized in such a way that all animals inherited similar gut microbiota [72]. Also in these mice, the change from a chow diet (low-fat/high-fiber) to the "Western" diet (high-fat/high-sugar) resulted in an increased weight gain [72]. In the diet-induced obese mice the relative abundances of the Firmicutes were higher, whereas those of the Bacteroidetes were lower compared with their lean status [72]. These findings are in agreement with previous results obtained with genetically obese (*ob/ob*) mice [60]. However, the changes in the Firmicutes phylum were not division wide but appeared to be mainly restricted to an increased abundance of the *Mollicutes* class [72]. Apparently, these diet-induced changes invoked an adaptation of the microbiota to the quality and quantity of the available nutrients, and these diet-induced microbiota adaptations are apparently reversible [72].

Follow-up experiments in mice receiving human gut microbiota transplants confirmed that the diet-induced changes on the GI communities in these so-called humanized mouse models are reversible as well [73]. Interestingly, obese and lean associated microbial communities could be maintained by diet alone. Maintenance of community structure and diversity was even achieved across several generations of mice following initial transplantation [73]. This again illustrates the prominent influence of the diet. Whether the diet is able to overrule the host genotype by permanent alterations of the GI community remains to be explored. In rats and humans, dietary "metabolic imprinting" through epigenetic modifications on the host genotype seems likely [74-76]. However, the reversibility of the gut ecology in the inter-host as well as intra-host species microbiota transplants described above seem to rule out the possibility of imprinting through the GI community [72, 73]. Although not permanent, it remains a fact that diet has a major and rapid impact on the microbiota.

Host–microbiota co-evolution: Selective geographic pressure?

Environment can be easily overlooked while studying GI tract microbiota, but is quite important as it determines physiological as well as microbial influences, e.g., through availability of food, food consumption habits, temperature, and humidity. Many environmental aspects are geographically confined, which raises the question if geography and its associated factors, such as climate, availability of food, and composition of the diet, could be an important aspect in host–microbe interactions. Several studies indicate an intimate co-evolution of humans and the gastric pathogen *Helicobacter pylori* [77, 78]. *Helicobacter pylori* appears to have spread from east Africa 58,000 years ago along with its human host. This finding implies that geography can also influence the microbial community, although modern commuting could blur the extent of such geographical impact. Naturally, the exact course and speed of this "blurring" depends on the magnitude of the evolutionary developed differences between the ethnic groups involved.

Many studies try to minimize the drastic effects of diet on the microbiota. Dietary habits usually depend on the geographical location of the subjects. Hence, not many studies have been performed on GI microbiota composition across country boundaries. One of the first studies across several countries was performed by Lay and coworkers and conducted on 91 subjects from five Northern European countries (France, Denmark, Germany, the Netherlands, and the United Kingdom) who consumed a non-restricted Western European diet [79]. However, the identified bacterial proportion of the gut microbiota did not significantly differ in composition when grouping the samples according origin, gender, or age [79]. Mueller and coworkers conducted a study that included as many as 230 healthy subjects from the more distant European countries such as France, Germany, Italy, and Sweden [80]. Several differences were found regarding country, age, gender, and combinations thereof [80]. Without much effort, dietary justifications can be found for most of the observed phylogenetic differences. An interesting example is the relative abundance of *Faecalibacterium prausnitzii*. Strict vegetarians appear to have no detectable amounts of *F. prausnitzii* [81]. The authors suggest that the highest levels of *F. prausnitzii* and related species was in the Swedish subjects and may be related to a high level of fish and meat consumption, which is a known dietary tradition of the Swedish population [80]. However, at

the time of this study the Swedes, Italians and French consumed the same amount of animal products (World Resources Institute - EarthTrends Environmental [http://earthtrends.wri.org/]), hence other dietary influences are probably implicated as well. The European studies seem to indicate that differences in GI microbiota composition increase with the distance between the geographic origins of the subjects. At a larger intercontinental distance, comparable studies were performed by American [9, 82] and Chinese [83] research groups. Although at phylum level, the Chinese and American subjects exhibited comparable phylogenetic GI tract compositions, principal coordinate analysis showed clear differences at species-level composition (Figure 2.4) [83]. Importantly, these findings reflect the differences in nuclear magnetic resonance based metabolic urine phenotypes found between large groups of Chinese and American subjects. Many of the differential urine metabolites do not have a mammalian origin but are derived from microbial sources [84]. Thus, it can be hypothesized that there is a co-variation between gut microbiota structure and the host metabolic phenotype. Urine metabolite profiles associated with microbial products are able to discriminate Japanese living in Japan and Japanese living in America [85]. Therefore, the differences found between the Americans and Chinese could be more dependent on diet and/or geography than on host genotype. Interestingly, the Chinese volunteers (Figure 2.4) all belonged to the same family and had more similar GI communities when compared among each other than the unrelated subjects within the American cohort. Hence, this observation does associate genetic relationship with GI microbiota composition and therefore points at a role for the host genotype in gut community development.

Whether geographic pressure has left its mark on the co-evolution of men and GI microbes is impossible to tell from the studies described in the previous paragraphs, without the knowledge on the influence of the diet and host genetics. Geography could be important with respect to the availability or prevalence of environmental or dietary microbial lineages for host colonization. Although the host (genotype) is dependent on the microbes it receives, it might be able to put selective pressure on sub-populations [44]. However, currently, no colonization restrictions have been observed in people who migrated from their traditional home countries into other geographic regions.

Intriguingly, the variation in colon cancer prevalence among Afro-American and native African populations appears to be in agreement with co-evolutionary relationships between human and GI microbiota caused by geographic and environmental conditions. Americans of African origin have the highest risk of all American subpopulations to develop colon cancer, whereas native Africans rarely suffer from this type of cancer [86]. The biggest difference between these two populations is probably their dietary habits, which prominently influences their GI microbiota communities [65, 72, 73]. Initial cultivation-dependent analyses confirmed differences in GI communities between native Africans and Afro-Americans [86], although these methods provide an incomplete impression of the microbiota. Newborn Afro-Americans are not exposed to the (relative) abundant levels of many microorganisms as they would in the native African region. Thus, no major colonization occurs of microbes that normally dominantly reside among the child's kin group, which essentially is a settled population established through generations-long consistent geographic and lifestyle factors. This means that the emigration process that

commonly coincides with changes in dietary habits seems to disrupt the co-evolved mutualism between the host genotype and its GI microbiota, resulting in the increased colon cancer risk in African-Americans. In conclusion, the functioning of the human super-organism appears to be

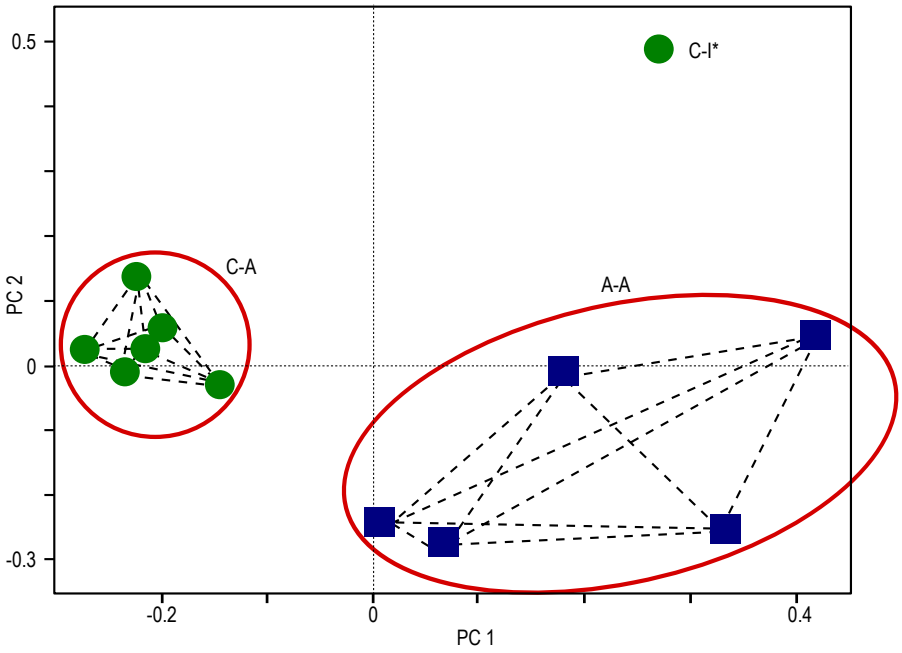


Figure 2.4 Principal component (PC) analysis on the species-level composition of the gut microbiota of Chinese family members (green circles) and American volunteers (blue squares). UniFrac metrics were used to generate the principal coordinate scores plot. The percentage of variation described by PC 1 is 19.8% and by PC 2 is 13.5%. Red circles are drawn around the group of adult Chinese (C-A) and around the American subjects (A-A). Intersample differences within the adult groups are visualized with dashed lines. C-I* indicates the Chinese baby 1.5 years of age, which was left out the C-A group due to its still developing gut microbiota. Figure adapted from [83] (American microbiota data from [9, 82]).

affected by currently fading barriers in human and environmental ecology caused by urbanization, global traveling, and emigration. These fading barriers probably coincide with a reduction in the number of microbe encounters, which could lead in humans to an underdeveloped immune system according to the hygiene hypothesis [87]. Such disruptions of long-term co-evolved interactions between man and microbe might partially explain the observed increase of chronic and degenerative disease frequencies in industrialized countries [86, 87].

Host-genotype polymorphisms

The genetic make-up of humans is very similar, and small genetic polymorphisms form an important aspect of genetic variation among human beings. Evidence is accumulating that these genetic variations are important determinants for the interactions with host-associated microbiota. Intuitively, the highly polymorphic immune-related genes are eminent candidates to define the interaction between host and microbe. The mucosal epithelial barrier has traditionally been considered to prevent contact between microbiota and underlying cells, including immune cells. Any contact was thought to provoke immune reactions that would even eradicate the

commensal organisms from the host. However, current knowledge clearly establishes frequent and essential communications between immune system and microbiota [88, 89]. Immune tolerance is promoted by local immune modulations that only inhibit further host tissue penetration of the commensal microbes [89-91]. This "peaceful situation" is normally maintained despite the massive presence of bacterial molecules that are capable of activating the host's bacterial molecular pattern recognition receptor and their cognate immune regulation cascades [88]. Polymorphisms in any of the immune-related genes involved in immune tolerance are prominent candidate determinants of the bacterial selection by the host.

Host–microbiota communication: Innate immune system

Important bacterial molecular pattern recognition receptors capable of initiating innate immune responses are the toll-like receptors. Toll-like receptor-4 (TLR-4) can recognize lipopolysaccharide (LPS), a cell wall component of Gram-negative bacteria [92, 93]. Mutations of the *TLR-4* gene result in a weakened immune response to LPS in mice [93]. Moreover, *TLR-4* sequence variants in humans correlated with a reduced response to inhaled endotoxins [94]. For Gram-positive-produced lipoproteins and lipoteichoic acids TLR-2 seems to be the main mammalian receptor [95]. Moreover, TLR-2 either alone or in heterodimer form with other TLRs can recognize more than a single ligand [96]. Naturally occurring mutations in the human *TLR-2* gene were shown to have diminished response to lipoproteins harvested from the Gram-positive bacteria *Borrelia burdorferi* and *Treponema pallidum* [95]. Although the polymorphisms in *TLR-4* and *TLR-2* are medically relevant, no studies have been reported, which determine their impact on human (or other mammalian organisms) associated microbiota. However, it is reasonable to suspect gene variants of important components of the innate immune system to be involved in the host–microbiota interactions. From mice studies, it is known that a genetically disabled innate immune system is associated with a collitogenic murine gut community [97]. Unfortunately, in these types of studies it is unclear whether the colitis is caused by a change in community structure or by the defective defense system of the murine hosts.

Other immune system components have been shown to be involved with the host–microbiota crosstalk. Besides TLRs, the innate immune system also depends heavily on the nucleotide-binding oligomerization domain (NOD) receptors [98]. An overview on the current knowledge of TLRs, NOD receptors, and other innate immune system receptors is given elsewhere [96]. NOD2 variants clearly show a relation between host genotype and gut microbiota composition, which is associated with increased risk for Crohn's disease (CD) [99, 100]. Normally, NOD2 binds the muramyl dipeptides of bacterial peptidoglycan [101], but certain polymorphisms in *NOD2* can result in failure of muramyl dipeptide detection. Such a failure results in a lack of tolerance development for commensal bacteria and dietary antigens and consequently leads to "inappropriate" immune response against them [102]. Considering these results, it is not surprising that CD patients were found to have lower diversity and diminished levels of normally abundant bacteria [103]. Especially the phylogenetic group *Clostridium leptum* in the Firmicutes phylum seems to be reduced in patients with CD. Several molecular techniques have indicated that *F. prausnitzii*, which is a member of the *C. leptum* group, is depleted in

the mucosa-associated communities [103-105]. Importantly, secreted metabolites of *F. prausnitzii* have been shown to exert anti-inflammatory effects *in vitro* [106]. In addition, *F. prausnitzii* is associated with an *in vivo* reduction of pro-inflammatory cytokine synthesis and increase of anti-inflammatory cytokine production in the colon. The observations concerning *F. prausnitzii* combined with the association of CD and *NOD2* mutations illustrate a potential link among host genotype, phenotype, and microbiota composition. *NOD2* mutant genotypes are likely to promote a negative selection of *F. prausnitzii* in the colon, which in turn can lead to the development of CD. Nevertheless, it is possible that other variations in the host genotype can allow the selection of other microbes similar to anti-inflammatory influences as *F. prausnitzii*.

Host–microbe interactions are not restricted to direct interactions between bacterial ligands and receptors of the innate immune system. Other proteins that should be considered are, for instance, further down the signaling pathway of TLRs or are involved in other cellular processes supporting immune responses, such as cell movement and restructuring. The gene *MEFV* encodes the protein pyrin, which is involved in innate immune response regulation, but has currently no definite function assignment [107]. This gene has no direct contact with microbes as it has been found in the cytoskeleton, but mutations of *MEFV* can lead to Mediterranean fever, a hereditary auto-inflammatory disorder [108]. In addition, patients with this disease have lower bacterial diversity and prominent population shifts in the Bacteroidetes, Firmicutes and Proteobacterium phyla during periods of active disease, whereas when the disease is in remission the bacterial gut community is more similar to normal microbial composition but still atypical [108]. Therefore, even polymorphisms in genes encoding proteins that are not supposed to be in direct contact with the gut microbes can influence host–microbe interactions. Recently, in healthy human mucosa, gene expression patterns have been found, which correlate with the development of immune tolerance for the organism *Lactobacillus plantarum* [109]. Even for one organism, these patterns involve many genes and therefore may be affected by many potential polymorphisms, which again exemplifies the complex nature of host–microbe interactions.

Host–microbiota communication: Non-immune-related mechanisms

Variations in genes not directly involved in immune system pathways can exert influence on the microbiota composition and functioning as well. Changing the conditions for the GI microbiota through modulation of host-derived resources or available attachment site in the mucosal layer seems of considerable importance. A nice illustration can be given by the abundant commensal species in the human and murine gut *Bacteroides thetaiotaomicron*. This organism matches its demand for fucose, a growth substrate, by upregulating fucosylated glycan production of epithelial cells in mice whenever pentose sugars are scarce [42]. Other bacteria might (ab)use this epithelial fucose synthesis regulation of *B. thetaiotaomicron*, as such this mechanism is important for multiple species in the GI microbiota [110]. Extrapolating this finding to the human situation, one can imagine the involvement of the fucosyltransferase enzyme polymorphisms, which are determinant for human blood types [111]. This type of gene has many variants since human glycoproteins are likely to be continuously evolving through natural selection provided by

commensal and pathogenic microorganisms [112]. Notably, some correlations between blood type and gut community variations indeed have been reported in the past [46, 113].

Another example of potential host-genotype-dependent interactions with gut microbes is mucin encoding (*MUC*) genes. These *MUC* genes encode protein backbones for mature mucin molecules, which are heavily glycsylated by O-linked oligosaccharides on the threonine, proline, and/or serine repeats [114]. The terminal oligosaccharides of mucin molecules contain sulfate and/or O-acetyl-substituted sialic acids. Mucins can both promote and prevent bacterial cell adhesion, depending on the exact structures of their O-glycan chains [115]. Changes in mucin composition have been associated with inflammatory bowel diseases [116]. Next to microbial mucin degradation, these changes could be due to genotype-related issues, such as polymorphisms in *MUC* genes, variations in *MUC* mRNA or protein levels, and variable posttranslational modification changes (i.e., the extent of glycosylation and sulfation) [116].

Many more host genes can be found that are involved in the functioning of the GI tract and thereby polymorphisms in such genes could have an impact on the microbial communities. Mutants in host enzymes responsible for nutrient breakdown and/or absorption could potentially influence the microbiota by altered nutrient composition and availability in the different regions of the intestine. For example, enterocytes in the small intestine can absorb glucose through active and passive glucose transporters [117]. Variants of these host-encoded transporters could cause alterations in the rate of carbohydrate absorption and thereby modulate the carbon source availability for the resident microbes, which may favor different microbial communities. More complex host–microbiota interactions are mediated by bile acids, which, next to their digestive functions (i.e., solubilization of lipids and lipid-soluble vitamins to enhance their absorption), have a role in maintaining the intestinal barrier [118]. Mammalian-microbiota co-metabolism result in the so-called secondary bile acids, which exert biologically important effects on both host and microbiota constituents [118]. Variations in level or composition of bile may have prominent effects on microbial communities and may correspond with specific consequences in mucosal cell biology. Primary bile acids are synthesized in the liver by a cascade of enzymes, providing many possibilities for gene variants that influence bile composition and corresponding host–microbe interactions.

Direct interactions of host genome and microbiota

Of all the possible interactions taking place in or around a host (Figure 2.1), modulation of the host genotype by the microbiota seems extraordinary. Nevertheless, when 223 genes in the rough draft of the human genome were found to potentially have a bacterial origin, horizontal gene transfer (HGT) from bacteria to humans has been suggested [119]. This could indicate that bacteria can manipulate their host, likely for their own benefit. However, HGT between human and bacteria is a difficult process because the genes should be stably integrated into the host DNA of germ line cells, to which bacteria normally do not have physical access. Furthermore, in 2001, Salzberg and coworkers carefully reexamined protein sequences of human, four other eukaryotes, and all completed prokaryote genomes at that time [120]. They only found about 40 human genes to be possible candidates for HGT from bacteria to humans. Therefore, HGT

between bacteria and humans remains doubtful since alternative, more plausible biological and technical explanations may be responsible for the few shared genes that are observed. One such biological factor is the high probability that the analyzed species had lost several genes from the common eukaryotic ancestor gene pool. Furthermore, nucleotide substitution rates can vary between genes within one genome as well as between similar genes in different organisms (i.e., evolutionary rate variation [121]). Therefore, in HGT analyses, evolutionary relatedness cannot be based on sequence similarity alone, indicating an important technical limitation in the currently available studies on this topic [122]. Furthermore, only five eukaryotic genomes were available at the time of analysis, three of which belong to the animal lineage (*Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*) [119, 120]. Hence, the total eukaryotic diversity was poorly represented. By contrast, the available prokaryotic genomes at that time embody a much broader evolutionary diversity [123]. This limited sample size of eukaryotic genomes is yet another technical problem confounding the HGT from bacteria to human. Concluding, it seems unlikely that bacteria have permanently manipulated their human hosts through HGT.

It seems more feasible that microbes sometimes pick genes up from their hosts. For example, the possible transfer of genes encoding serpins, which are protease inhibitors involved in the regulation of many physiological processes [124]. Although serpins are found in all three domains of life, which indicate that they could originate from a common ancient, serpins are found in relatively few prokaryotes. The latter would imply that serpins are not essential for survival or that they may have been acquired by prokaryotes as the result of HGT. Even though the serpin of *Bifidobacterium longum* is distantly related to eukaryotic serpins, Ivanov and coworkers showed that it exerts inhibitory functions through an identical mechanism. Another clearer example is the presence of *nptA* gene, a sodium/phosphate co-transporter, in *Vibrio cholerae*. This gene is also present in animals but seems to be absent in any other bacterial species [125]. Furthermore, *V. cholerae* has been shown to exhibit activity similar to that of its animal homologs [125]. It is likely that this transporter facilitates *V. cholerae* in the GI tract and consequently could be involved in the pathogenicity of this microbe. However, in both the serpin and the nptA case, it remains difficult to prove that the genes are not derived from an ancestral gene instead of being transferred from a (mammalian) host species.

Mitochondria provide off course evidence that indeed bacterial DNA resides in the mammalian genotype. Although these eukaryotic organelles originate from the endosymbiosis of an α-Proteobacterium ancestor, they show no indications that they were introduced in order to manipulate the host for their own benefit. By contrast, recent findings show that it was the eukaryotic host that took control and manipulated the bacterial endosymbiont for its own benefit [126]. During the transformation of bacterium to organelle, many bacterial genes not involved in energy conversion were lost or replaced by genes originating from the eukaryotic host [126].

Conclusions

Humans and other mammalian hosts provide only a minor quantity of genes to their super-organism metagenome. Yet these genes are essential and decisive in defining the final host–microbiota interactions. Disturbances in the host genotype can lead to malfunctioning of the super-organism, i.e., all kinds of metabolic disorders, immune diseases, and other disorders. Human genotypes consist of a huge amount of variables, many of which could be of importance for host–microbiota interactions. Not only genes directly related to the immune system should be considered in future studies. For instance, *Escherichia coli* has been implicated to, via quorum sensing, cross-communicate with the host epinephrine signaling pathway [127]. Although this is coming from a pathogenic species, one cannot exclude this type of non-immune-system-related communication between the host and its commensals. Another form of communication is through metabolites, such as SCFAs or (secondary) bile acids [59, 118]. Hence, many different mechanisms constitute the overall host–microbe interactions pallet. Basically, the hierarchy in importance of human genotypes in relation to host–microbe interactions is unknown. Currently, the majority of predictions hint at immune-related factors, i.e., gene polymorphisms. However, genes involved in metabolic functions and their control or those involved in biosynthetic pathways, such as mucus production and modification, or bile metabolism are likely to be important modulators as well.

New human genotyping efforts using extensive volunteer cohorts, combined with in-depth microbiota profiling, provide a possibility to mine for all factors underlying the relation between human host and its microbial communities. Furthermore, comprehensive studies following both monozygotic and dizygotic twins from birth to adulthood will provide vital information to assess the relative contribution of host genotype to the GI microbiota composition. These studies will be most successful when they acquire additional metadata, such as dietary habits, and actual short-term nutrient intake, and preferably also include intergenerational analysis of the subjects' families. Such multivariant analyses will be essential to dissect influences of dietary, environmental, and host-genotype factors.

Nevertheless, diet will probably always be an obscuring factor due to its dramatic, but apparently reversible effects on the microbiota. Future studies could benefit from the consistent use of family members, different ethnic groups, or both. Difficult studies, from an ethical point of view, in which the subjects are isolated for longer periods of time and under strict dietary regimes, might provide better insight in the human–diet–microbiota relationship. Regardless of the chosen study types, it will be a "life-changing" experience to finally fully understand both dietary and host genotype influences involved in shaping and interacting with the intestinal microbiota. Such knowledge may enable the definition of dietary regimes that provide prophylactic and therapeutic possibilities for a variety of disorders and/or diseases, provided that a causal relationship underlies the observed diet and microbiota correlation with these disorders or diseases. Specific dietary design might be attempted to correct deviating microbiota compositions and/or activity associated with specific diseases toward a more "healthy microbiota." The rapidly developing field of (functional) metagenomics may allow us in the near future to actually come

to the accurate description of what can be considered a "healthy microbiota," which could then be employed as a biomarker in diagnosis and treatment of diseases and/or disorders.

"He who does not know food, how can he understand the diseases of men?" – Hippocrates (460–357 B.C.)

Acknowledgments

ST and EGZ received funding from the European Community's Seventh Framework Program (FP7/2007-2013): MetaHIT, grant agreement HEALTH-F4-2007-201052, coordinated by S. Dusko Ehrlich (Institut National de la Recherche Agronomique, France).

References

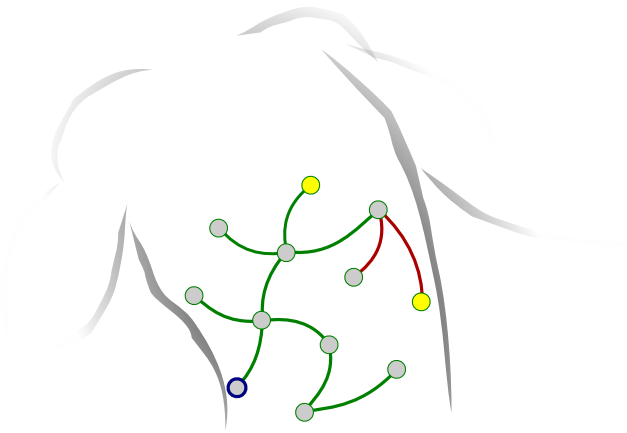
1. **Aas JA, Paster BJ, Stokes LN, et al.** Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 2005;43(11):5721-32.
2. **Pei Z, Bini EJ, Yang L, et al.** Bacterial biota in the human distal esophagus. Proc Natl Acad Sci U S A 2004;101(12):4250-5.
3. **Finegold SM.** Normal Indigenous Intestinal Flora. in Human Intestinal Microflora in Health and Disease. D.J. Hentges. Editor. 1983. Academic Press: New York. p. 3-31.
4. **Pavoine S, Dufour AB, and Chessel D.** From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. J Theor Biol 2004;228(4):523-37.
5. **Hayashi H, Takahashi R, Nishi T, et al.** Molecular analysis of jejunal, ileal, caecal and recto-sigmoidal human colonic microbiota using 16S rRNA gene libraries and terminal restriction fragment length polymorphism. J. Med. Microbiol. 2005;54(11):1093-1101.
6. **Cummings JH and Macfarlane GT.** The control and consequences of bacterial fermentation in the human colon. J. Appl. Bacteriol. 1991;7(6):443-459.
7. **Marteau P, Pochart P, Doré J, et al.** Comparative study of bacterial groups within the human cecal and fecal microbiota. Appl. Environ. Microbiol. 2001;67(10):4939-4942.
8. **Bäckhed F, Ley RE, Sonnenburg JL, et al.** Host-bacterial mutualism in the human intestine. Science 2005;307(5717):1915-1920.
9. **Eckburg PB, Bik EM, Bernstein CN, et al.** Diversity of the human intestinal microbial flora. Science 2005;308(5728):1635-1638.
10. **Gao Z, Tseng CH, Strober BE, et al.** Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. PLoS One 2008;3(7):e2719.
11. **Fredricks DN, Fiedler TL, and Marrazzo JM.** Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med 2005;353(18):1899-911.
12. **Cummings JH and Macfarlane GT.** Colonic microflora: Nutrition and health. Nutrition 1997;13(5):476-478.
13. **Metges CC.** Contribution of microbial amino acids to amino acid homeostasis of the host. J. Nutr. 2000;130(7):1857S-1864.
14. **Ramotar K, Conly JM, Chubb H, et al.** Production of menaquinones by intestinal anaerobes. J. Infect. Dis. 1984;150(2):213-218.
15. **Albert MJ, Mathan VI, and Baker SJ.** Vitamin B12 synthesis by human small intestinal bacteria. Nature 1980;238(5749):781-782.
16. **Begley M, Gahan CGM, and Hill C.** The interaction between bacteria and bile. FEMS Microbiol. Rev. 2005;29(4):625-651.
17. **Adlercreutz H, Pulkkinen MO, Hamalainen EK, et al.** Studies on the role of intestinal bacteria in metabolism of synthetic and natural steroid hormones. J. Steroid. Biochem. 1984;20(1):217-229.
18. **Attene-Ramos MS, Wagner ED, Plewa MJ, et al.** Evidence that hydrogen sulfide is a genotoxic agent. Mol. Cancer Res. 2006;4(1):9-14.
19. **Schicho R, Krueger D, Zeller F, et al.** Hydrogen sulfide is a novel prosecretory neuromodulator in the Guinea-pig and human colon. Gastroenterology 2006;131(5):1542.
20. **Summerton J, Goeting N, Trotter GA, et al.** Effect of deoxycholic acid on the tumour incidence, distribution, and receptor status of colorectal cancer in the rat model. Digestion 1985;31(2-3):77-81.

21. **Turnbaugh PJ, Ley RE, Mahowald MA, et al.** An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 2006;444(7122):1027-31.
22. **Savage DC.** Microbial ecology of the gastrointestinal tract. Ann. Rev. Microbiol. 1977;31:107-133.
23. **Wei C and Brent MR.** Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 2006;7:327.
24. **Yang X, Xie L, Li Y, et al.** More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body. PLoS One 2009;4(6):e6074.
25. **Cani PD, Neyrinck AM, Fava F, et al.** Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. Diabetologia 2007;50(11):2374-83.
26. **Holmes E and Nicholson JK.** Human metabolic phenotyping and metabolome wide association studies. Ernst Schering Found Symp Proc 2007;(4):227-49.
27. **Wen L, Ley RE, Volchkov PY, et al.** Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature 2008;455(7216):1109-13.
28. **Zhang C, Zhang M, Wang S, et al.** Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. ISME J 2009;4(2):232-241.
29. **Brugman S, Klatter FA, Visser JT, et al.** Antibiotic treatment partially protects against type 1 diabetes in the Bio-Breeding diabetes-prone rat. Is the gut flora involved in the development of type 1 diabetes? Diabetologia 2006;49(9):2105-8.
30. **Zengler K, Walcher M, Clark G, et al.** High-throughput cultivation of microorganisms using microcapsules. Methods Enzymol 2005;397:124-30.
31. **Ingham CJ, Sprenkels A, Bomer J, et al.** The micro-Petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms. Proc Natl Acad Sci U S A 2007;104(46):18217-22.
32. **Palmer C, Bik EM, DiGiulio DB, et al.** Development of the human infant intestinal microbiota. PLoS Biol 2007;5(7):e177.
33. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
34. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-484.
35. **DiGiulio DB, Romero R, Amogan HP, et al.** Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: a molecular and culture-based investigation. PLoS One 2008;3(8):e3056.
36. **Conway PL.** Microbial ecology of the human large intestine. in Human Colonic Bacteria. Role in Nutrition, Physiology, and Pathology. G.R. Gibson and G.T. Macfarlane. Editors. 1995. CRC Press: Boca Raton, FL. p. 1-24.
37. **Zoetendal EG, Akkermans AD, and de Vos WM.** Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. Appl. Environ. Microbiol. 1998;64(10):3854-3859.
38. **Kurokawa K, Itoh T, Kuwahara T, et al.** Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. 2007;14(4):169-181.
39. **Rajilić-Stojanović M.** Diversity of the human gastrointestinal microbiota: novel perspectives from high throughput analyses. in Laboratory of Microbiology. 2007. Wageningen University: Wageningen. p. 216.
40. **Zoetendal EG, Rajilić-Stojanović M, and de Vos WM.** High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota Gut 2008;57:1605-1615.
41. **Tap J, Mondot S, Levenez F, et al.** Towards the human intestinal microbiota phylogenetic core. Environ Microbiol 2009;11(10):2574-84.
42. **Hooper LV, Xu J, Falk PG, et al.** A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. Proc. Natl. Acad. Sci. USA 1999;96(17):9833-9838.
43. **Toivanen P, Vaahtovuo J, and Eerola E.** Influence of major histocompatibility complex on bacterial composition of fecal flora. Infect Immun 2001;69(4):2372-7.
44. **Rawls JF, Mahowald MA, Ley RE, et al.** Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell 2006;127(2):423-33.
45. **Zoetendal EG, Akkermans ADL, Akkermans-van Vliet WM, et al.** The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. Microbial Ecology in Health and Disease 2001;13(3):129-134.
46. **Van de Merwe JP, Stegeman JH, and Hazenberg MP.** The resident faecal flora is determined by genetic characteristics of the host. Implications for Crohn's disease? Antonie Van Leeuwenhoek 1983;49(2):119-24.
47. **Stewart JA, Chadwick VS, and Murray A.** Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. J Med Microbiol 2005;54(12):1239-1242.

48. **Claesson MJ, O'Sullivan O, Wang Q, et al.** Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One 2009;4(8):e6669.
49. **Topping DL and Clifton PM.** Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. Physiol Rev 2001;81(3):1031-64.
50. **Cummings JH.** Short chain fatty acids in the human colon. Gut 1981;22(9):763-79.
51. **Ley RE, Hamady M, Lozupone C, et al.** Evolution of Mammals and Their Gut Microbes. Science 2008;320(5883):1647-1651.
52. **Kliewer SA, Sundseth SS, Jones SA, et al.** Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors Î± andâ€‰Î³. Proceedings of the National Academy of Sciences of the United States of America 1997;94(9):4318-4323.
53. **Debril M-B, Renaud J-P, Fajas L, et al.** The pleiotropic functions of peroxisome proliferator-activated receptor ? Journal of Molecular Medicine 2001;79(1):30-47.
54. **Yen C-J, Beamer BA, Negri C, et al.** Molecular Scanning of the Human Peroxisome Proliferator Activated Receptor [gamma] (hPPAR[gamma]) Gene in Diabetic Caucasians: Identification of a Pro12Ala PPAR[gamma]2 Missense Mutation. Biochemical and Biophysical Research Communications 1997;241(2):270-274.
55. **Lindi V, Schwab U, Louheranta A, et al.** Impact of the Pro12Ala polymorphism of the PPAR-gamma2 gene on serum triacylglycerol response to n-3 fatty acid supplementation. Mol Genet Metab 2003;79(1):52-60.
56. **Backhed F, Manchester JK, Semenkovich CF, et al.** Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U S A 2007;104(3):979-84.
57. **Samuel BS, Shaito A, Motoike T, et al.** Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. Proc Natl Acad Sci U S A 2008;105(43):16767-72.
58. **Backhed F, Ding H, Wang T, et al.** The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 2004;101(44):15718-23.
59. **Maslowski KM, Vieira AT, Ng A, et al.** Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature 2009;461(7268):1282-6.
60. **Ley RE, Backhed F, Turnbaugh P, et al.** Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 2005;102(31):11070-5.
61. **Waldram A, Holmes E, Wang Y, et al.** Top-down systems biology modeling of host metabotype-microbiome associations in obese rodents. J Proteome Res 2009;8(5):2361-75.
62. **Friedman JM and Halaas JL.** Leptin and the regulation of body weight in mammals. Nature 1998;395(6704):763-70.
63. **Considine RV, Sinha MK, Heiman ML, et al.** Serum immunoreactive-leptin concentrations in normal-weight and obese humans. N Engl J Med 1996;334(5):292-5.
64. **Clement K, Vaisse C, Lahlou N, et al.** A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. Nature 1998;392(6674):398-401.
65. **Ley RE, Turnbaugh PJ, Klein S, et al.** Microbial ecology: Human gut microbes associated with obesity. Nature 2006;444:1022-1023.
66. **Schwiertz A, Taras D, Schafer K, et al.** Microbiota and SCFA in Lean and Overweight Healthy Subjects. Obesity (Silver Spring) 2009.
67. **Duncan SH, Lobley GE, Holtrop G, et al.** Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes 2008;32(11):1720-1724.
68. **Roediger WEW.** Role of anaerobic bacteria in the metabolic welfare of the colonic mucosa in man. Gut 1980;21(9):793-798.
69. **Hague A, Singh B, and Paraskeva C.** Butyrate acts as a survival factor for colonic epithelial cells: further fuel for the in vivo versus in vitro debate. Gastroenterology 1997;112(3):1036-40.
70. **Medina V, Afonso JJ, Alvarez-Arguelles H, et al.** Sodium butyrate inhibits carcinoma development in a 1,2-dimethylhydrazine-induced rat colon cancer. JPEN J Parenter Enteral Nutr 1998;22(1):14-7.
71. **Hamer HM, Jonkers D, Venema K, et al.** Review article: the role of butyrate on colonic function. Aliment. Pharmacol. Ther. 2008;27(2):104-119.
72. **Turnbaugh PJ, Backhed F, Fulton L, et al.** Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe 2008;3(4):213-23.
73. **Turnbaugh PJ, Ridaura VK, Faith JJ, et al.** The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci. Transl. Med. 2009;1(6):6ra14.
74. **Bateson P, Barker D, Clutton-Brock T, et al.** Developmental plasticity and human health. Nature 2004;430(6998):419-21.
75. **Godfrey KM and Barker DJ.** Fetal programming and adult health. Public Health Nutr 2001;4(2B):611-24.

76. **Lillycrop KA, Phillips ES, Jackson AA, et al.** Dietary protein restriction of pregnant rats induces and folic acid supplementation prevents epigenetic modification of hepatic gene expression in the offspring. J Nutr 2005;135(6):1382-6.
77. **Linz B, Balloux F, Moodley Y, et al.** An African origin for the intimate association between humans and Helicobacter pylori. Nature 2007;445(7130):915-8.
78. **Moodley Y, Linz B, Yamaoka Y, et al.** The peopling of the Pacific from a bacterial perspective. Science 2009;323(5913):527-30.
79. **Lay C, Rigottier-Gois L, Holmstrom K, et al.** Colonic microbiota signatures across five northern European countries. Appl Environ Microbiol 2005;71(7):4153-5.
80. **Mueller S, Saunier K, Hanisch C, et al.** Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. Appl. Environ. Microbiol. 2006;72(2):1027-1033.
81. **Hayashi H, Sakamoto M, and Benno Y.** Fecal microbial diversity in a strict vegetarian as determined by molcular analysis and cultivation. Microbiol. Immunol. 2002;46(12):819-831.
82. **Gill SR, Pop M, Deboy RT, et al.** Metagenomic analysis of the human distal gut microbiome. Science 2006;312(5778):1355-9.
83. **Li M, Wang B, Zhang M, et al.** Symbiotic gut microbes modulate human metabolic phenotypes. Proc. Natl. Acad. Sci. U S A. 2008;105(6):2117-2122.
84. **Dumas ME, Maibaum EC, Teague C, et al.** Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. Anal Chem 2006;78(7):2199-208.
85. **Holmes E, Loo RL, Stamler J, et al.** Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 2008;453(7193):396-400.
86. **O'Keefe SJ, Chung D, Mahmoud N, et al.** Why do African Americans get more colon cancer than Native Africans? J Nutr 2007;137(1 Suppl):175S-182S.
87. **Guarner F, Bourdet-Sicard R, Brandtzaeg P, et al.** Mechanisms of disease: the hygiene hypothesis revisited. Nat Clin Pract Gastroenterol Hepatol 2006;3(5):275-84.
88. **Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, et al.** Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell 2004;118(2):229-241.
89. **MacDonald TT and Monteleone G.** Immunity, inflammation, and allergy in the gut. Science 2005;307(5717):1920-1925.
90. **Cebra JJ.** Influences of microbiota on intestinal immune system development. Am. J. Clin. Nutr. 1999;69(5):1046S-1051.
91. **Macpherson AJ, Geuking MB, and McCoy KD.** Immune responses that adapt the intestinal mucosa to commensal intestinal bacteria. Immunology 2005;115(2):153-62.
92. **Hoshino K, Takeuchi O, Kawai T, et al.** Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. J Immunol 1999;162(7):3749-52.
93. **Poltorak A, He X, Smirnova I, et al.** Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. Science 1998;282(5396):2085-8.
94. **Arbour NC, Lorenz E, Schutte BC, et al.** TLR4 mutations are associated with endotoxin hyporesponsiveness in humans. Nat Genet 2000;25(2):187-91.
95. **Lorenz E, Mira JP, Cornish KL, et al.** A novel polymorphism in the toll-like receptor 2 gene and its potential association with staphylococcal infection. Infect Immun 2000;68(11):6398-401.
96. **Opitz B, Eitel J, Meixenberger K, et al.** Role of Toll-like receptors, NOD-like receptors and RIG-I-like receptors in endothelial cells and systemic infections. Thromb Haemost 2009;102(6):1103-9.
97. **Garrett WS, Lord GM, Punit S, et al.** Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. Cell 2007;131(1):33-45.
98. **Inohara N, Ogura Y, Chen FF, et al.** Human Nod1 confers responsiveness to bacterial lipopolysaccharides. J Biol Chem 2001;276(4):2551-4.
99. **Hugot JP, Chamaillard M, Zouali H, et al.** Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 2001;411(6837):599-603.
100. **Ogura Y, Bonen DK, Inohara N, et al.** A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 2001;411(6837):603-6.
101. **Girardin SE, Boneca IG, Viala J, et al.** Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. J Biol Chem 2003;278(11):8869-72.
102. **Landers CJ, Cohavy O, Misra R, et al.** Selected loss of tolerance evidenced by Crohn's disease-associated immune responses to auto- and microbial antigens. Gastroenterology 2002;123(3):689-99.

103. **Manichanh C, Rigottier-Gois L, Bonnaud E, et al.** Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006;55(2):205-11.
104. **Frank DN, St Amand AL, Feldman RA, et al.** Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc. Natl. Acad. Sci. U S A. 2007;104(34):13780-13785.
105. **Martinez-Medina M, Aldeguer X, Gonzalez-Huix F, et al.** Abnormal microbiota composition in the ileocolonic mucosa of Crohn's disease patients as revealed by polymerase chain reaction-denaturing gradient gel electrophoresis. Inflammatory Bowel Diseases 2006;12(12):1136-1145.
106. **Sokol H, Pigneur Bnd, Watterlot L, et al.** Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proceedings of the National Academy of Sciences 2008;105(43):16731-16736.
107. **Ting JP, Kastner DL, and Hoffman HM.** CATERPILLERs, pyrin and hereditary immunological disorders. Nat Rev Immunol 2006;6(3):183-95.
108. **Khachatryan ZA, Ktsoyan ZA, Manukyan GP, et al.** Predominant role of host genetics in controlling the composition of gut microbiota. PLoS One 2008;3(8):e3064.
109. **van Baarlen P, Troost FJ, van Hemert S, et al.** Differential NF-kappaB pathways induction by Lactobacillus plantarum in the duodenum of healthy humans correlating with immune tolerance. Proc Natl Acad Sci U S A 2009;106(7):2371-6.
110. **Hooper LV and Gordon JI.** Commensal host-bacterial relationships in the gut. Science 2001;292(5519):1115-1118.
111. **Becker DJ and Lowe JB.** Fucose: biosynthesis and biological function in mammals. Glycobiology 2003;13(7):41R-53R.
112. **Hooper LV and Gordon JI.** Glycans as legislators of host-microbial interactions: spanning the spectrum from symbiosis to pathogenicity. Glycobiology 2001;11(2):1R-10R.
113. **Hoskins LC.** Mucin degradation in the human gastrointestinal tract and its significance to enteric microbial ecology. European Journal of Gastroenterology & Hepatology 1993;5(4):205-213.
114. **Herrmann A, Davies JR, Lindell G, et al.** Studies on the "insoluble" glycoprotein complex from human colon. Identification of reduction-insensitive MUC2 oligomers and C-terminal cleavage. J Biol Chem 1999;274(22):15828-36.
115. **Hollingsworth MA and Swanson BJ.** Mucins in cancer: protection and control of the cell surface. Nat Rev Cancer 2004;4(1):45-60.
116. **Morita H, Kettlewell MG, Jewell DP, et al.** Glycosylation and sulphation of colonic mucus glycoproteins in patients with ulcerative colitis and in healthy subjects. Gut 1993;34(7):926-32.
117. **Los EL, Wolters H, Stellaard F, et al.** Intestinal capacity to digest and absorb carbohydrates is maintained in a rat model of cholestasis. Am J Physiol Gastrointest Liver Physiol 2007;293(3):G615-22.
118. **Martin F-PJ, Dumas M-E, Wang Y, et al.** A top-down systems biology view of microbiome-mammalian metabolic interactions in a mouse model. Mol Syst Biol 2007;3.
119. **Lander E, Linton L, Birren B, et al.** Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.
120. **Salzberg SL, White O, Peterson J, et al.** Microbial genes in the human genome: lateral transfer or gene loss? Science 2001;292(5523):1903-6.
121. **Li W-H.** Molecular Evolution 1997. Sunderland, MA: Sinauer Associates.
122. **Eisen JA.** Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 1998;8(3):163-7.
123. **Nelson KE, Paulsen IT, Heidelberg JF, et al.** Status of genome projects for nonpathogenic bacteria and archaea. Nat Biotechnol 2000;18(10):1049-54.
124. **Ivanov D, Emonet C, Foata F, et al.** A serpin from the gut bacterium Bifidobacterium longum inhibits eukaryotic elastase-like serine proteases. J. Biol. Chem. 2006;281(25):17246-17252.
125. **Lebens M, Lundquist P, Soderlund L, et al.** The nptA gene of Vibrio cholerae encodes a functional sodium-dependent phosphate cotransporter homologous to the type II cotransporters of eukaryotes. J Bacteriol 2002;184(16):4466-74.
126. **Gabaldón T and Huynen MA.** From Endosymbiont to Host-Controlled Organelle: The Hijacking of Mitochondrial Protein Synthesis and Metabolism. PLoS Comput Biol 2007;3(11):e219.
127. **Sperandio V, Torres AG, Jarvis B, et al.** Bacteria-host communication: the language of hormones. Proc Natl Acad Sci U S A 2003;100(15):8951-6.



Chapter 4

Multispecies probiotic intervention induces microbial network changes in line with symptom reduction in patients with irritable bowel syndrome

Sebastian Tims, Mirjana Rajilić–Stojanović, Kajsa Kajander, Riina A. Kekkonen, Erwin G. Zoetendal, Michiel Kleerebezem, Willem M. de Vos

Submitted

Abstract

Irritable bowel syndrome (IBS) is a common, usually life-long, intestinal disorder. Although IBS etiology is largely unknown, some successes of probiotic interventions have been reported. The aim of this study was to decipher the microbiota response to a multi-species probiotic treatment that led to the improvement of IBS symptoms. Fecal samples from 62 IBS patients were taken before, during, and after a double-blind placebo controlled trial. Microbiota analysis was performed by the HITChip phylogenetic microarray. The microbiota dynamics were assessed and changes in the microbiota were correlated with IBS symptom scores. Finally, the development of the bacterial co-occurrence networks were analyzed in both treatment arms. A stability increase of the total microbiota occurred in the probiotic-consuming subjects (p=0.012), whereas stability in the placebo-receiving group did not change. No consistent changes in relative abundance of the microbial groups were observed other than groups representing the probiotic strains. The patients on the probiotic intervention could be divided on responders and non-responders and, remarkably, the non-responders had almost 2-fold lower levels of bacteria related to *Faecalibacterium prausnitzii* than the responders. Moreover, a restructuring of the microbial co-occurrence networks was observed in the responders, which could explain the probiotic-mediated stabilization of the microbial community. Deep phylogenetic analysis revealed that the microbiota composition went through temporal stabilization and shifts in bacterial connectivity networks occurred along with IBS symptom improvement during the probiotic intervention.

Introduction

Irritable bowel syndrome (IBS) is a common and chronic intestinal disorder. Worldwide, IBS prevalence among the adult and adolescent population is about 10-20% [1]. Patients with IBS suffer from abdominal discomfort or pain associated with defecation, bloating, discomfort, and disturbed defecation habits [1]. Most IBS patients experience fluctuations in severity of the symptoms, sometimes affecting social and personal life to an extent similar to chronic diseases like diabetes [2,3]. Although the IBS etiology is still largely unknown, it can be regarded as a multi-factorial disease which involves lifestyle [4], psychological stress [5], host genetics [6], but also the intestinal microbiota [7-9] and the gut-brain axis [10]. Most of the earlier studies on the relationship between IBS and the intestinal microbiota were limited in power and analysis depth [11]. More recent studies employing high throughput DNA technologies have revealed an altered intestinal microbiota of IBS patients compared to healthy subjects, both in adults [12-15] and children [16,17]. Although many studies have revealed specific differences in microbiota from (subgroups of) IBS patients and healthy subjects, general consensus with respect to composition is still lacking due to the heterogeneity of IBS and the variation in approaches that were used to study the microbiota composition [18]. A common observation is an enrichment of certain Firmicutes groups, mainly of species related to *Ruminococcus gnavus* [15,19,20], in conjunction with reduced abundance of some *Bacteroides* groups in IBS patients [13,15,21]. Therefore, treatments of IBS are shifting towards modulation of the intestinal microbiota through dietary therapies with FODMAPs [22], prebiotics, administration of probiotics, and even fecal transplantations [23].

In this study we investigated the microbiota dynamics in a successful double-blind placebo controlled trial that showed IBS symptom score decrease and the microbiota stabilization in a subset of analyzed samples upon consuming multi-species probiotic diary drink containing *Lactobacillus rhamnosus* strains GG and Lc705, *Bifidobacterium animalis* BB12 and *Propionibacterium freudenreichii* subsp. shermanii JS [24]. The participants consisted of a Finnish IBS cohort of 62 patients whose microbiota prior to intervention showed distinct composition compared to healthy controls [15]. Phylogenetic microarray (HITChip [25]) profiling of the fecal microbiota showed a significantly increased microbiota stability after consumption of the probiotic, extending earlier observations with a smaller subset of subjects [24]. In addition, several microbial groups were found to develop new associations during the trial specifically in the probiotic-consuming patients.

Materials & methods

Participants

The study group consisted of 62 primary care IBS patients recruited from the Helsinki area in Finland for a follow-up clinical trial [24]. All IBS patients fulfilled the Rome II criteria [26] and were investigated with colonoscopy or barium enema of the colon in the 5 years prior to the start of the trial. The average age of the 62 IBS patients (57 females, 5 males) was 49 years (range, 22–66 years). Detailed subject inclusion criteria were reported earlier [24]. Written informed

consent, approved by Ethics Committees of the Pirkanmaa Hospital District (Finland), was provided by all participants [24].

Study design

The probiotic intervention study was conducted and described previously [24]. In short, the trial was randomized, double-blind, placebo-controlled covering a 5-month period that was preceded by a three week run-in period and followed by a three week follow-up period. Fecal samples were collected for 62 subjects at baseline (sample A; run-in period), halfway through (sample B; variable) and just after the end of the trial (sample C; follow-up period). The probiotic treatment arm consumed daily 120 mL of a milk-based drink (80% lactose-free milk and 20% fruit juice) containing approximately 1×107 colony-forming units (CFU)/mL of each *Lactobacillus rhamnosus* GG (LGG®), *Lactobacillus rhamnosus* Lc705, *Propionibacterium freudenreichii* subsp. shermanii JS, and *Bifidobacterium animalis* subsp. *lactis* Bb12 (Valio Ltd, Helsinki, Finland) while the control arm received a placebo acidified drink lacking these probiotic strains, but otherwise similar to the probiotic drink.

IBS diaries

The IBS symptoms and bowel habits were surveyed by means of a patient diary as reported earlier [24]. A total IBS score was composed of the scores for the following symptoms: abdominal pain, distension, flatulence, and rumbling. Both the total score as well as the individual symptom scores were correlated to the microbial data gathered in this study.

Fecal samples collection, DNA extraction and phylogenetic microarray analysis

Fecal samples were collected and DNA was isolated as reported previously [24]. Analyses by the phylogenetic microarray, the HITChip, were performed as described earlier [15,25]. Total microbiota assessments were performed on normalized signal values of the 3,699 unique HITChip probes. These probe values were used to calculate Simpson's Diversity index for each sample and Spearman's correlation coefficient between different samples to assess microbiota similarity. Microbiota stability for each subject was analyzed by subtracting the similarity coefficient between samples A and B from the similarity coefficient between samples B and C. HITChip probes are assigned to different phylogenetic levels: order-like 16S rRNA gene sequence groups; genus-like 16S rRNA gene sequence groups (sequence similarity >90%); and species-like 16S rRNA gene sequence groups (sequence similarity >98%) [25]. For each sample relative abundances were calculated for the genus-like groups by summing all signal values of the probes targeting these phylogenetic groups and subsequently dividing by the total of all probe signals. All comparisons were pair-wise and significance was assessed with dependent 2-group Wilcoxon signed rank tests, unless noted otherwise. For all statistical tests that were performed on multiple parameters the false discovery rate was estimated by calculating q-values. Results with both p-values <0.05 and q-values <0.05 were regarded as significant. In addition, all analyses were repeated on the set of samples from the subjects that responded to the treatment. Similar to the definition recently used

by Chumpitazi and co-workers [27], we defined subjects that showed >50% reduction in any of the symptom scores as responders.

In order to link microbiota to IBS symptoms, Spearman's correlation coefficient (ϱ) was calculated between genus-like group relative abundances and IBS symptom scores. Network analysis was performed by calculating ϱ between the genus-like groups of all samples per treatment arm and per time point separately. In all correlation analyses only the correlations with $|\varrho|>0.5$ (and $q<0.05$) were accepted as "real" correlations. Specific correlations induced during the trial (for details see SI) were visualized with the 'Gephi Graph Visualization and Manipulation' software [28].

Results

Composition and stability of the fecal microbiota during the trial

Detailed and reproducible analysis of the fecal microbiota by phylogenetic microarray was performed to obtain a high-resolution microbiota profile of each sample of the 62 subjects, taken before, during or after the intervention (samples A, B and C, respectively). Unsupervised clustering (Figure S4.1) showed that the 186 samples clustered exclusively according to subject, confirming the strong individuality of the intestinal microbiota and the limited impact of the probiotic intervention [29-31]. Intra-individual similarities (alpha-diversity; ϱ=0.851 ± 0.065 SD) were significantly higher compared to the inter-individual similarities (beta-diversity; ϱ=0.699 ± 0.065 SD; p=1e-97, two-sample Student's T-test with unequal variance between groups). No microbiota differences were detected between the baseline samples of both treatment groups and

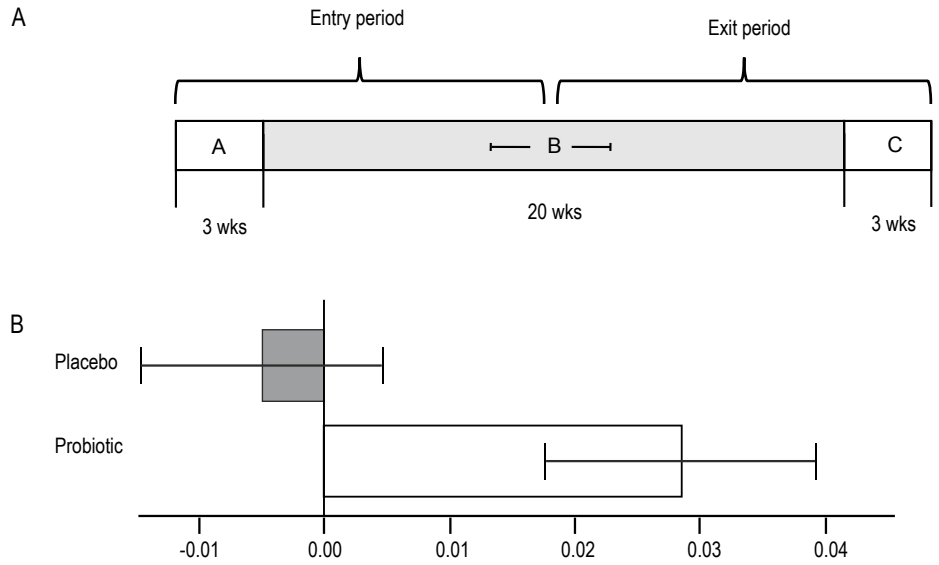


Figure 4.1 Trial set-up and microbiota stability. A) Probiotic clinical trial set-up. B) Intestinal microbiota stability of subjects consuming the placebo and probiotic drinks. Microbiota stability was visualized by subtracting the similarity (correlation coefficient) of the entry period A-B from the similarity of the exit period B-C. The increase in stability occurring within the probiotic treatment arm was significant (p=0.012), while the change in stability within the placebo treatment arm was not significant (p=0.612).

no consistent changes in diversity were observed during the intervention period for either the treatment or the placebo group (Figure S4.2).

The entry (from sample A to B) and exit (from point B to C) periods were used to estimate the microbiota stability throughout the intervention. Time between sampling points was not related to the similarity between the samples (Figure S4.3) and therefore allowed this microbiota stability analysis (Figure 4.1). Subjects from the probiotic group showed significantly higher microbiota similarity between the exit period samples compared the entry period samples (p=0.012), whereas for the subjects from the placebo group showed no differences between these periods (p=0.612), indicating a significantly increased temporal stability of the fecal microbiota following the probiotic intervention.

To evaluate whether the increased microbiota stability in the probiotic group could be assigned to detectable changes of specific microbial groups, the relative abundance was determined for the 129 16S rRNA-based genus-like phylogenetic groups that are targeted by the HITChip microarray [25]. However, none of those groups showed any consistent changes during the intervention except for those targeting the species in the probiotic product (Table 4.1). Only the bifidobacteria did not appear change throughout the intervention, but when zooming-in on the species-like level we observed a strong (35-fold) increase of the signals of belonging to *B.animalis* subsp. *lactis* Bb12 (Table S4.1).

Table 4.1 Genus-like groups with significant changes throughout the trial in the probiotic consuming subjects. Significant changes (q < 0.05) are indicated in bold. The Lactobacillus plantarum et rel. group includes the probes for the probiotic Lactobacillus rhamnosus strains. strains.

Genus-like group	Treatment arm	Relative abundance at baseline (%)	Median change entry period (%)	Median change exit period (%)
Lactobacillus plantarum et rel.	Placebo	0.136	+2.6	-5.5
	Probiotic	0.149	+15.5	-15.9
Propionibacterium	Placebo	0.035	+1.7	-5.6
	Probiotic	0.037	+12.7	-14.8

As reported previously, patients treated with the probiotic showed an improved intestinal health [20] and the vast majority (30 out of 35) showed a more than 50% reduction in any of the symptom scores and these were identified as responders. Only a small group (n=5) were found to be non-responders. None of the results described above was influenced by the removal of the samples from the non-responders. Diversity, richness and microbiota profiles of the non-responders were all well within the ranges observed in the responders. However, one noticeable difference in microbiota between the non-responders and the responders in the probiotic treatment arm was observed. The relative abundance of bacteria related to *Faecalibacterium prausnitzii* at baseline was 1.75-fold lower for the non-responders (p=0.081; Figure S4.4). However, when categorizing the first quartile (Q1) of the measured *F.prausnitzii* abundances as "low", non-responders have a significantly larger proportion of subjects with low levels of *F.prausnitzii* et rel. (Chi-square test p=0.014).

Table 4.2 Genus-like group significantly correlating to IBS symptom scores.

Genus-like group	IBS symptom score	Time point	Treatment arm	Spearman's correlation coefficient (ρ)	q-value
Klebisiella pneumoniae et rel.	Pain	A	Placebo	0.534	0.039
Lachnospira pectinoschiza et rel.	Pain	A	Placebo	-0.529	0.042
Clostridium cellulosi et rel.	Total score	B	Placebo	0.573	0.020
Subdoligranulum variable at rel.	Distension	C	Probiotic	-0.567	0.005

Correlation between changes in IBS symptom scores and changes in quantitative microbiota data

We observed previously that the IBS patients receiving the probiotic showed on average an improvement in the total IBS score of 11 points compared to that of the placebo group [24]. Using sample A as a baseline measurement, the present cohort (i.e. all patients with a complete set of samples, n=62) showed the same decrease in total IBS score in the probiotic treatment arm (Figure S4.5). Compared to the total IBS score, the symptom scores distension, flatulence, pain, and rumbling showed similar trends, i.e. improvements symptom score of 1.5 to 3 points more compared to the placebo group (Figure S4.5), similar to the previous observation [24].

To determine which microbiota fluctuations were associated with IBS symptom changes the genus-like group abundances were correlated to the symptom scores. Four genus-like groups (with bacteria related to *Klebisiella pneumoniae* in a positive way and *Lachnospira pectinoschiza*, *Clostridium cellulosi* and *Subdoligranulum variable* in a negative way) could be directly correlated to IBS symptom scores at different time points for each treatment arm (Table 4.2).

Ecological networks of microbial groups

The effect of the probiotic intervention on the ecological connections within the GI microbiota was investigated by calculating the correlations between the microbial groups on all three time points for both probiotic and placebo treatment arms separately. These correlations represent (positive) co-occurrences and (negative) anti-occurrences between the microbial groups and shall, from now on, be referred to as "connections". A total of 49 genus-like groups showed consistent positive connections (ϱ>0.5) at all three time points in both placebo and probiotic treatment arms. For 38 genus-like groups these consistent connections were captured in four separate network modules (I to IV; Figure 4.2). No consistent negative connections were detected (ϱ< -0.5 in all time points, for both treatment arms). Besides the consistent connections, a large amount of variable connections were observed between the different time points and treatment arms. Interestingly, no connections appeared to develop during the trial that were unique for the placebo-receiving subject group. In contrast, several connections emerged during the trial that were unique for the probiotic-receiving subject group (Figure 4.2). These unique connections, referred to as "probiotic-induced connections", encompassed 34 genus-like groups of which half belong to the consistent network modules I, II, and III. All three genus-like groups that represent the probiotic strains used in this intervention are among the probiotic induced connections.

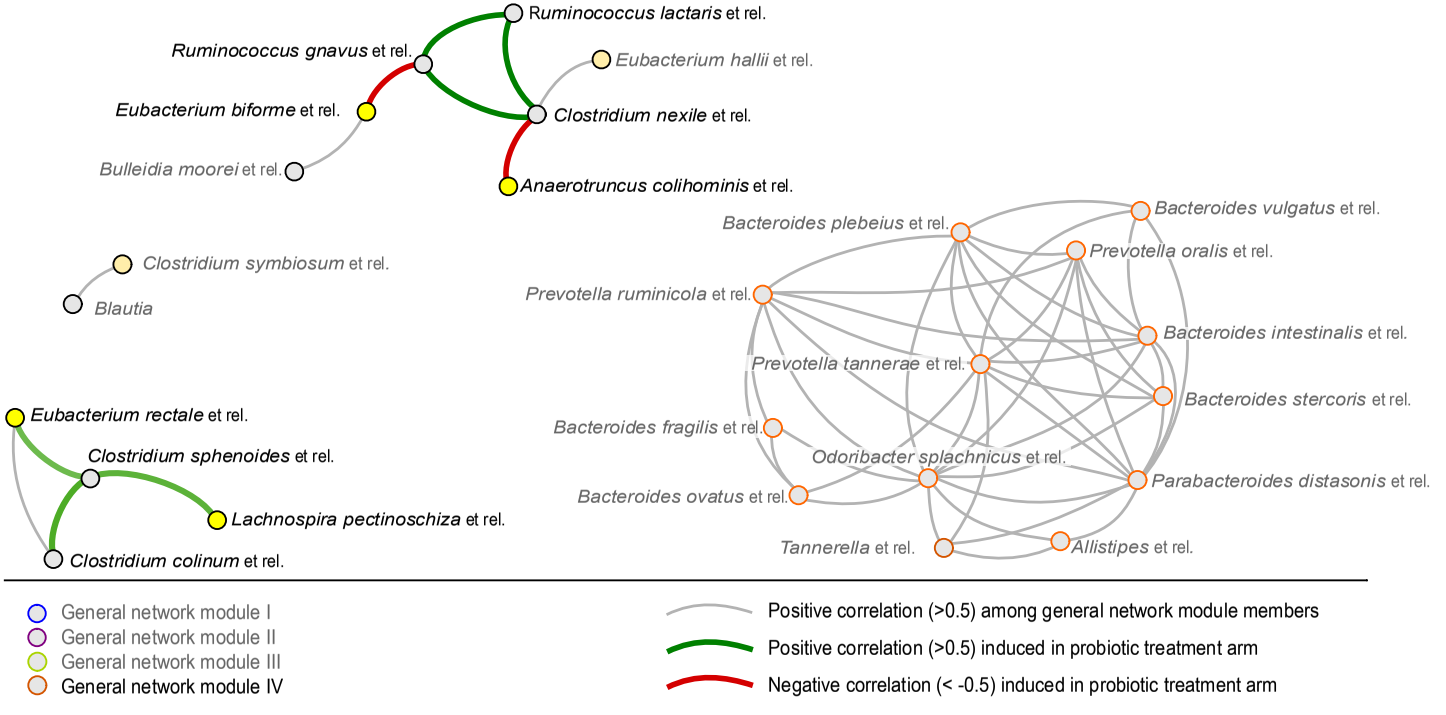


Figure 4.2 Consistent network modules of genus-like groups and probiotic treatment induced connections. Consistent connections of genus-like groups across all time points and both treatment arms of the trial are colored grey. All connections shown have correlations of ρ>0.5 or ρ<-0.5; q<0.05). Nodes outlines are colored by network module. Nodes with yellow fill indicate genus-like groups that cover known butyrate producing isolates. Genus-like groups that represent strains of the used multi-species probiotic are marked with blue font.

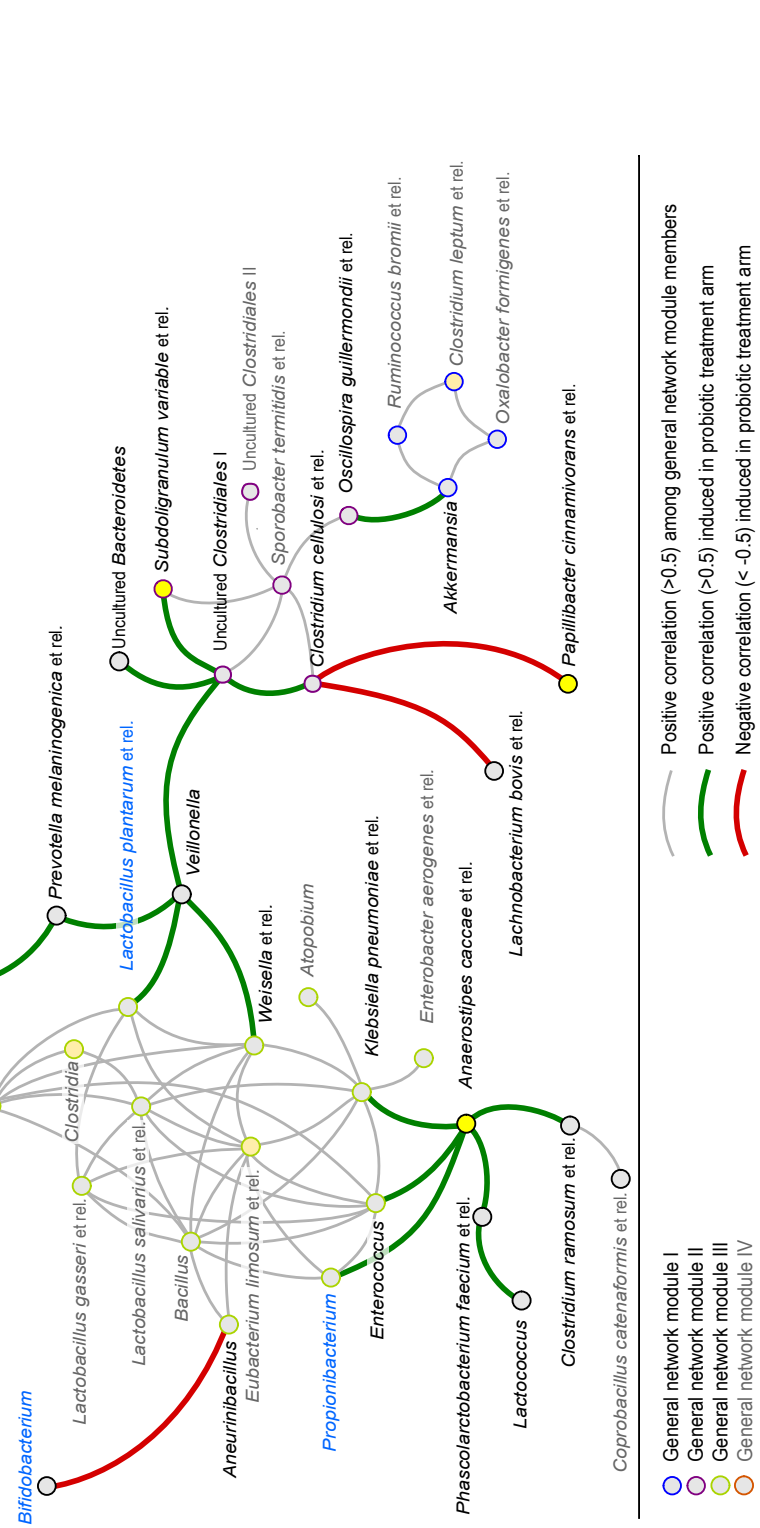


Figure 4.2 (continued) Consistent network modules of genus-like groups and probiotic treatment induced connections. Consistent connections of genus-like groups across all time points and both treatment arms of the trial are colored grey. All connections shown have correlations of $\rho>0.5$ or $\rho<-0.5$; $q<0.05$). Nodes outlines are colored by network module. Nodes with yellow fill indicate genus-like groups that cover known butyrate producing isolates. Genus-like groups that represent strains of the used multi-species probiotic are marked with blue font.

Careful inspection of the networks revealed that the probiotic treatment appeared to elicit four types of microbial network consequences: 1) extension of existing network modules to encompass additional genus-like groups; 2) strengthening existing network modules by increasing connections of members intrinsic to the network module; 3) connecting existing network modules; 4) formation of novel network modules. Network module II extended with (negative) connections to bacteria related to *Lachnobacterium bovis* and *Papillibacter cinnamivorans*, strengthened internal connections between its members, as well as aligning with module I (through alignment of *Akkermansia* and the relatives of *Oscillospira guillermondii*) and III (through the addition of *Veillonella* to the network). Network module III extended its connections with the bacteria related to *Anaerostipes caccae*, *Bifidobacterium*, *Eubacterium siraeum*, *Lactococcus*, *Phascolarctobacterium faecium*, *Prevotella melaninogenica*, and *Veillonella*. New network modules emerged in the intervention group as well: the consistent connection between the relatives of *Clostridium colinum* and *Eubacterium rectale* got extended with *Clostridium sphenoides* and *Lachnospira pectinoschiza* (Figure 4.2); the relatives of *Anaerotruncus colihominis*, *Clostridium nexile*, *Eubacterium biforme*, *Ruminococcus gnavus*, and *Ruminococcus lactaris* formed a new dynamic network. These results show that the probiotic treatment elicited an extensive restructuring of the ecological environment, which included altered connectivity patterns for some microbial groups of which the abundance correlated with disease symptom scores.

Discussion

The increasing awareness of the potential role of the intestinal microbiota in the etiology of IBS opens avenues for treatments of the disease that focus on modulating the microbiota [7-11]. Here we report on the microbial changes observed during the administration of a multi-species probiotic. Analysis of the fecal microbiota composition of 62 IBS patients before, during and after a randomized double-blind placebo controlled clinical trial provided new insights into the mechanisms underlying microbiota modulation, including the rearrangement of ecosystem-network connectivity of bacterial groups to form a more stable ecological environment, which may have contributed to the observed improvements in IBS symptoms.

While not all probiotic interventions in IBS patients have been equally successful [8], two independent double-blind and controlled trials have reported significant improvement of symptoms following consumption of the multispecies probiotic mixture consisting of *L.rhamnosus* strains GG and Lc705, *P.freudenreichi* PAJ and either *B.breve* Bb99 or *B.animalis* BB12 in capsule or milk-based drink [24,32]. Remarkably, in both studies the placebo group did not show a significant change apart from a slight initial decrease in symptoms that was ascribed to placebo effects. Therefore it is tempting to speculate that consumption of the probiotic bacteria species had an impact on the intestinal microbiota composition and that this impact was relevant for the observed symptom improvement. Although the present study could not account for direct interactions between the probiotic strains and the host, e.g. mucosal interactions in the small intestine, it does give a novel insight into the microbiota changes induced by probiotic consumption that were associated with changes in IBS symptoms. Notably, only in the probiotic group did the microbiota stability significantly increase, indicating that treatment dependent

changes in the microbiota did occur and persisted during the trial, reinforcing the previous report on a subset of 20 patients from our cohort [24].

No consistent changes were observed in the relative abundances of microbial groups in the feces of the subjects consuming the probiotic other than the species present in the probiotic supplement (Table 4.1). In concordance overall richness and diversity were not affected by the probiotic treatment either. Given that the intestinal microbiota is extremely complex and stable ecosystem, that is able to resist harsh environmental challenges (such as dietary changes or an antibiotic treatment course) for periods measured in years [33] this result does not come as a surprise. Since other studies have reported similar results in probiotic trials [8,34] we can conclude that impact of the multi-species probiotic on the intestinal microbiota is subtle in nature.

Direct correlations between microbial abundances and symptom scores linked several microbial groups to the total IBS symptom score, as well as those for the specific distension and pain symptoms (Table 4.2). These correlations implicated bacteria related to *Klebisiella pneumoniae* (positive), *Lachnospira pectinoschiza*, *Clostridium cellulosi* and *Subdoligranulum variable* (negative) with IBS symptoms, across treatment arms and time points. Hence, the bacteria-symptom associations were variable over time, even in the placebo treatment arm. Therefore, the impact of the multi-species probiotic is probably dependent on the ecological context (at baseline), suggesting that IBS subject stratification could be improved by taking microbiota composition into account. Jeffrey and co-workers have already shown a microbiota clustering independent of Rome II IBS sub-typing, which was predicted accurately by determining the Firmicutes to Bacteroidetes ratio (F:B) [13]. Our cohort was previously reported to be characterized by a higher F:B ratio in IBS patient compared to healthy individuals [15], which is a common observation in IBS cohorts [18-20]. Nevertheless, we did not observe correlations between F:B ratios and symptom scores. Another way to stratify the studied subjects is by classifying their response to the treatment. In probiotic trials the comparison of the baseline microbiota of non-responders versus responders could reveal factors influencing the effectiveness of the treatment. No significant difference in microbiota characteristics between the non-responders and responders were found here. Yet, it is interesting that the largest baseline difference between responders and non-responders was in a specific Firmicutes group: the relatives of *F.prausnitzii* were present in a 1.75-fold lower abundance in the non-responders (p=0.081). Moreover, among the non-responders the proportion of subjects with low amounts of *F.prausnitzii* was significantly higher (Chi square test, p=0.014). *F.prausnitzii* is a predominant butyrate producer in the human intestinal tract, suggested to have anti-inflammatory properties [35]. Low amounts of *F.prausnitzii* appear to be a common denominator for patients with (active) IBD (as reviewed in [36]), and have a negative correlation with symptom score at baseline sample of our IBS cohort [15]. Since both IBS and IBD are disorders that can exhibit similar symptoms, we could hypothesize that the non-responders are more akin to IBD patients, which appears to render the studied multi-species probiotic less effective.

Underlining the impact of the ecological context for IBS symptoms are the observed differences between microbial co-occurrence networks and their development throughout that were exclusively elicited in the probiotic treatment arm of the trial. This observation strengthens

the suggestion that administration of the multi-species probiotic modulates the intestinal microbiota, especially while considering that all microbial groups that represent the probiotic strains, i.e. *Bifidobacterium*, *L.plantarum* and *Propionibacterium*, are part of the probiotic induced connections (Figure 4.2). When elaborating on the potential consequences of the probiotic induced connections, we should keep the intrinsic microbial-ecosystem context in mind, i.e. the consistent network connections present in the cohort irrespective of time point or treatment. Although the consistent network module IV stays rather inert, network modules I, II, and III encounter one or more type(s) of consequences elicited by the probiotic treatment.

The probiotic consumption strengthened the internal connections of module II and aligned modules II and I, resulting in a network structure which resembles one previously described in a healthy twin cohort [37] (**Chapter 3**). This network was hypothesized to encompass primary fiber degraders whose fermentation products could be used by secondary fermenter butyrate producers. Interestingly one member of this network, i.e. the Uncultured *Clostridiales* I group, has recently been identified to occur in a bimodal abundance pattern among 1,000 western adults and was related to IBS as well as severe obesity [38]. Bacteria related to *Subdoligranulum variable* account for another consistent member of network module III in this cohort but *S.variable* was not seen in the healthy twin cohort network. However, isolates of *S.variable* can be regarded secondary fermenter butyrate producers [39], which further implicates modules I and II in butyrate production potential. Moreover, the extensions of network module III introduced new microbial members of which cultured isolates have various short chain fatty acid (SCFA) production profiles [40-44], resulting in a network module seemingly without a clear preference for any particular SCFA. Previous studies have associated abnormal fecal SCFA content, both high and low, with IBS [45], abdominal symptoms [46] or diarrhea [47]. Absorption of SCFAs by epithelial cells greatly influences water and electrolyte balance [48], indirectly impacting gut transit. Rat and canine models have even shown direct stimulation of colonic transit and motility by SCFA mixtures via different signaling pathways [49,50]. Although these previously reported results do not provide clear insights for IBS symptom mechanisms, they do underline that the quantities and ratios of the different SCFAs can have important consequences for gut functioning. Indeed in diarrhea-predominant IBS patients, administration of enteric-coated sodium butyrate tablets rapidly reduced fecal water content and furthermore improved abdominal pain, bloating and flatulence after 30 days of treatment [51]. Therefore, the observed ecosystem-structure rearrangements surrounding network module I, II and III are potentially able to affect the SCFA profile in the colon and thereby colon functioning.

Veillonella, previously described to be a typical small intestinal inhabitant [52,53], connects to both network modules II and III. Next to *Veillonella*, two more microbial groups that may originate from the small intestine appear in the probiotic induced connections: *Enterococcus* spp. and, the here sometimes pain-associated, *Klebisiella pneumoniae*. All three groups have been found among the most common isolates from patients with small intestinal bacterial overgrowth (SIBO) [54]. Studies addressing the involvement of SIBO in IBS have generated controversial result (for a review see [55]). However, a mild increase of bacteria that are among the dominant members of the small intestinal microbiota was commonly detected in IBS patients in comparison

with healthy individuals [56]. On the host side, *in vivo* data has shown that microbiota can modulate gene expression in the mucosa, most notably in the small intestine [57]. In a small human cohort a six week intervention with one of the strains used in this trial, *L.rhamnosus* GG, altered the immune and mucosal related transcriptional networks profiles in the duodenum [58]. If we can hypothesize that the multi-species probiotic in this cohort affected the physiological conditions in the small intestine similarly, such changes would likely have impacted the flow through of small intestinal commensals into the colon. Therefore, our results suggest that typical small intestinal inhabitants tie into IBS symptom mechanisms either directly by the restructured community or indirectly by reflecting changes in the function of the small intestinal tissue.

The formation of new network modules by the probiotic treatment is an intriguing aspect of the present work that may shed light on the mechanisms in IBS. Since the functions of many intestinal bacteria are not yet know, an attempt to explain all connections would be highly speculative. The largest, most complex new probiotic induced network module were found to include two groups of *Ruminococcus* spp. i.e. *R.gnavus* and *R.lactaris*, which are frequently identified in relation to health and disease phenotypes and appear to have associations with IBS [59]. For example, increased abundance levels of *R.gnavus* have been associated with IBS [15,19,20] and IBD [60,61], which may be related to the mucolytic activity of this species. Mucosal barrier integrity might explain the observed negative connection of these bacterial groups and *Clostridium nexile* group to the relatives of *Eubacterium biforme* and *Anaerotruncus colihominis*. Relatives of both *E.biforme* and *A.colihominis* have been shown to be able to produce butyrate [62,63]. One of the physiological effects attributed to butyrate is the increase of the intestinal barrier by stimulation of mucin glycoprotein production [64] and/or strengthening of tight junctions [65]. In a recent study that correlated mucosal gene expression and the intestinal microbiota of IBS and healthy subjects [66], it was found that increased cell junction and permeability was associated with healthy microbiota that included SCFA producers *S.variable* and *Prevotella melaninogenica*, both members of the probiotic-induced network (Figure 4.2).

We can conclude that extensive restructuring of the ecological environment, unique to the probiotic treatment arm, was observed. Our data strongly suggest that microbiota management influences IBS symptom scores, supporting the role of the microbiota in IBS. Dietary modulation, through probiotic consumption, is capable of eliciting seemingly subtle changes in microbial abundances and yet extensive ecological context dependent microbiota restructuring, of which the exact underlying mechanisms remain to be deciphered. Moreover, we could stratify probiotic responders and non-responders based on the abundance of *F.prausnitzii*. Microbiota-based stratification, microbiota stability improvements and probiotic-induced microbiota network connectivity changes that may impact SCFA production, are all credible targets in dietary strategies aimed to relieve IBS symptoms and provide an improved quality of life for IBS patients.

Acknowledgements

This work was partially funded by the unrestricted Spinoza Award of the Netherlands Organization of Scientific Research awarded to WMdV.

Supplementary information

Supplementary materials & methods

Co-occurrence network analysis

Spearman's correlation coefficient (ϱ) was calculated between the genus-like groups of all samples per treatment arm and per time point separately. In all correlation analyses only the correlations with $|\varrho| > 0.5$ (and $q < 0.05$) were accepted as "real" correlations. To identify which correlations were specifically induced in either treatment arm the following rule was applied: the correlation must be "real" in sample C ($|\varrho| > 0.5$ and $q < 0.05$) of the considered treatment arm and $|\varrho| < 0.3$ in sample A of the same treatment arm and all samples of the other treatment arm. For all correlation analysis only the bacterial groups that were present at >0.01% abundance in >50% of the subjects in the respective data sets were used. Highly related genus-like groups that share the majority of their responsive probes were discarded from the network analysis. A combination of the R package 'network' [67] and the 'Gephi Graph Visualisation and Manipulation' software [28] was used to visualize the significant correlations.

Supplementary figures

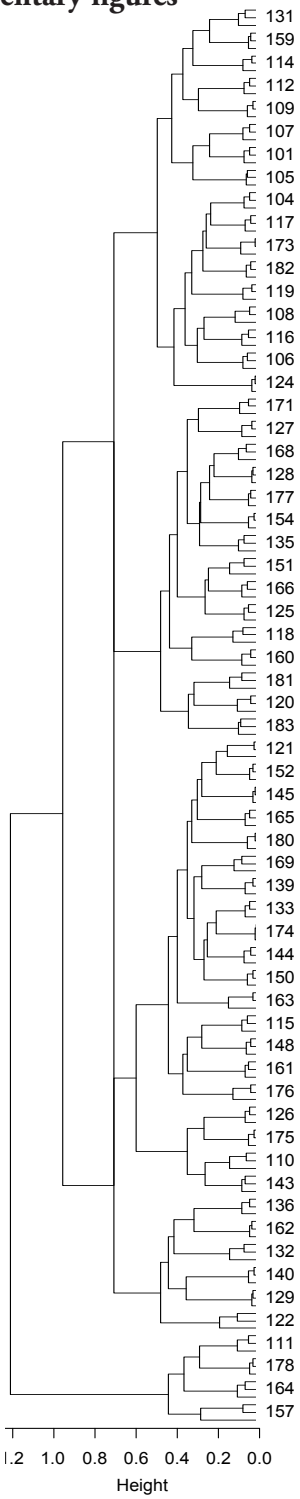


Figure S4.1 unsupervised hierarchical clustering (based on Spearman's correlation coefficient) shows the total of 186 samples to cluster exclusively according to subject. Only subject number is indicated (time points not marked).

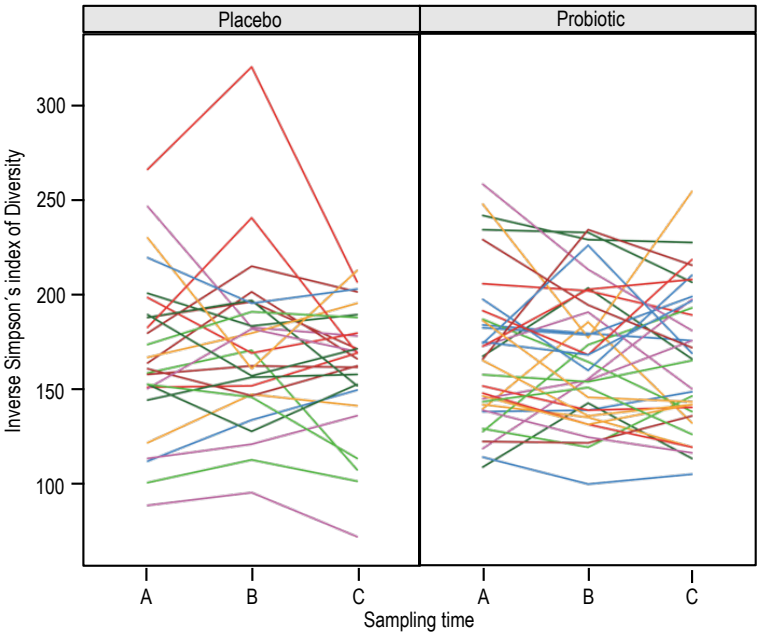


Figure S4.2 Sample total microbiota diversity development in time.

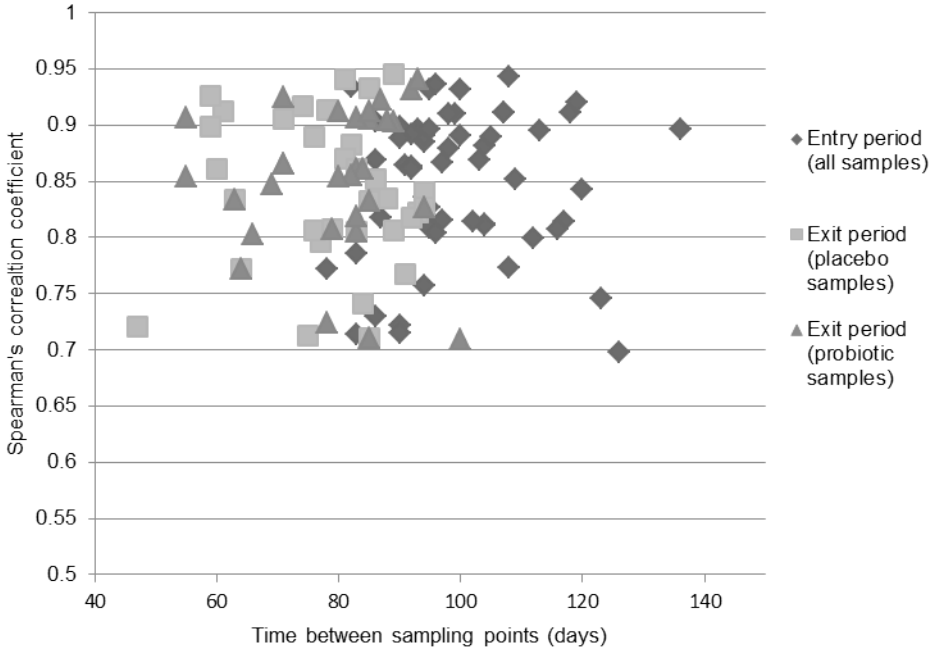


Figure S4.3 Sample microbiota similarity versus sampling time. Regression analysis showed no association between Spearman correlation coefficient and time between sampling in the inception period (p = 0.725) or in the settled period (all samples p = 0.548, placebo samples p = 0.804; probiotic samples p = 0.499).

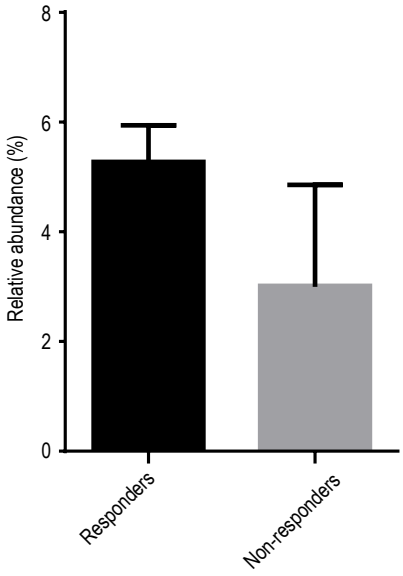


Figure S4.4 Faecalibacterium prausnitzii genus-like group relative abundance in responders and non-responders.

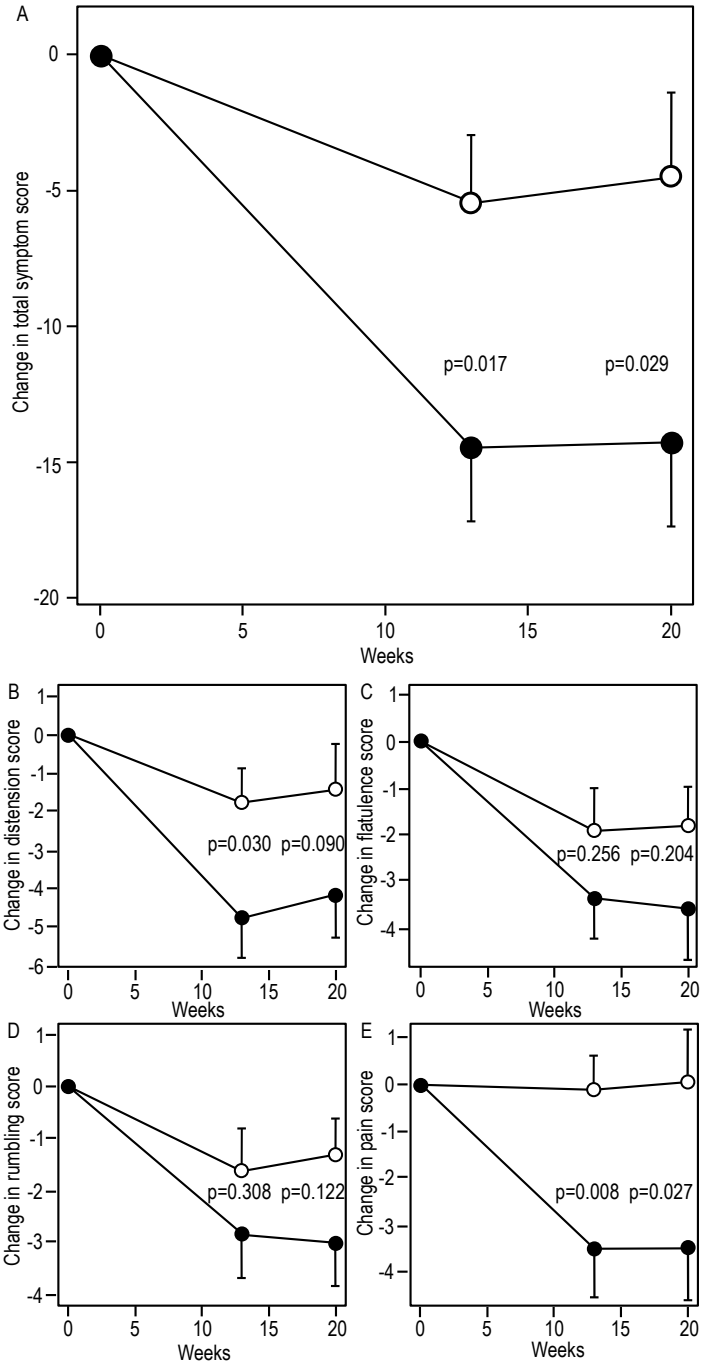


Figure S4.5 IBS symptom score changes throughout the trial (normalized for baseline scores).

Supplementary tables

Table S4.1 Signal intensity changes within genus-like group which harbour probes targeting probiotic strains.

Genus-like group	Species-like group	Change signal intensity (%)				Probiotic strain coverage
		Placebo		Probiotic		
		period A-B	period B-C	period A-B	period B-C	
Lactobacillus plantarum et rel.	*Lactobacillus antri*	4.6	-6.6	0.7	-0.8	
	Lactobacillus brevis	0.5	-2.4	-0.1	-0.3	
	Lactobacillus buchneri	1.6	-3.0	0.1	-1.7	
	Lactobacillus casei	0.0	5.1	47.9	-51.4	*Lactobacillus rhamnosus* GG (LGG®); *Lactobacillus rhamnosus* Lc705
	Lactobacillus fermentum	6.0	-7.0	2.0	-3.8	
	Lactobacillus gastricus	6.8	-8.8	0.7	2.5	
	Lactobacillus mucosae	7.0	-8.1	2.03	4.1	
	Lactobacillus oris	5.4	-8.5	1.0	0.5	
	Lactobacillus paracasei	1.5	-3.8	3.6	-7.7	
	Lactobacillus plantarum	7.6	-10.4	-3.8	0.01	
	Lactobacillus reuteri	3.9	-0.8	0.01	-1.9	
	Lactobacillus rhamnosus	-0.1	9.5	85.6	-89.7	*Lactobacillus rhamnosus* GG (LGG®); *Lactobacillus rhamnosus* Lc705
	Lactobacillus vaginalis	3.7	-3.9	-0.01	-1.9	
	Lactobacillus vaginalis KC19	4.7	-4.6	-0.02	-1.3	
	uncultured bacterium cadhufec111c10	0.8	-2.3	-0.3	-2.0	
	uncultured *Lactobacillus* sp. LabB103	-3.0	14.8	136.7	-132.4	*Lactobacillus rhamnosus* GG (LGG®); *Lactobacillus rhamnosus* Lc705
	uncultured Pediococcus sp. NS1A12	-1.0	-1.2	-3.0	0.3	
Propionibacterium	*Propionibacterium acnes*	1.2	-1.9	-0.1	-1.9	
	Propionibacterium freudenreichii	6.6	-3.5	27.2	-27.6	*Propionibacterium freudenreichii* subsp. shermanii JS
	Propionibacterium granulosum	0.5	-1.4	1.9	-3.8	

Table S4.1 (continued) Signal intensity changes within genus-like group which harbour probes targeting probiotic strains.

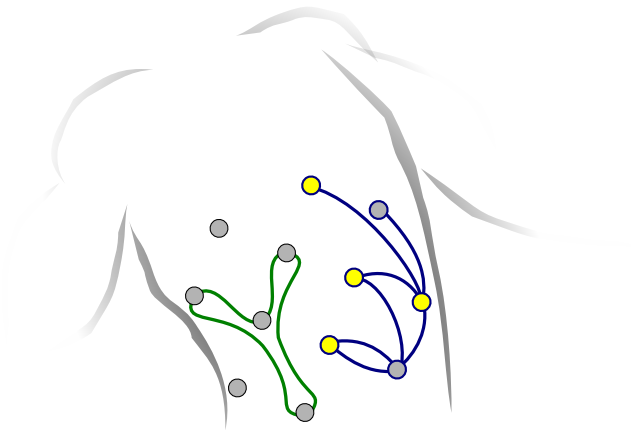
Genus-like group	Species-like group	Change signal intensity (%)				Probiotic strain coverage
		Placebo		Probiotic		
		period A-B	period B-C	period A-B	period B-C	
Bifidobacterium	*Bifidobacterium adolescentis*	-15.0	-3.1	2.7	11.2	
	Bifidobacterium angulatum	-11.3	-3.0	-4.6	9.8	
	Bifidobacterium animalis	-10.3	18.1	35.4	-39.4	*Bifidobacterium animalis* subsp. lactis Bb12
	Bifidobacterium bifidum	-29.9	7.7	-2.06	12.0	
	Bifidobacterium breve	-29.3	4.5	6.6	8.9	
	Bifidobacterium catenulatum	-19.5	4.0	3.9	11.9	
	Bifidobacterium dentium	-18.3	2.4	4.5	6.7	
	Bifidobacterium gallicum	-8.7	4.2	6.6	-2.2	
	Bifidobacterium infantis	-25.3	2.6	4.3	6.9	
	Bifidobacterium longum	-32.0	6.0	-0.2	11.0	
	Bifidobacterium pseudocatenulatum	-37.2	9.2	-0.7	13.3	
	Bifidobacterium pseudolongum	-36.1	9.6	-2.3	14.5	
	Bifidobacterium thermophilum	-23.2	2.9	3.8	13.0	
	uncultured bacterium Adhufec069rbh	-13.5	-4.0	4.2	10.4	
	Uncultured bacterium clone Eldhufec085	-34.3	-0.5	9.4	19.2	
	Uncultured bacterium clone Eldhufec088	-15.3	-1.8	4.8	10.1	
	uncultured *Bifidobacterium* sp. 13D	-15.3	-1.8	4.8	10.1	
	uncultured *Bifidobacterium* sp. 15D	-22.6	1.6	10.6	7.3	

References

1. **Longstreth GF, Thompson WG, Chey WD, et al.** Functional bowel disorders. Gastroenterology 2006;130(5):1480-91.
2. **Hungin APS, Whorwell PJ, Tack J, et al.** The prevalence, patterns and impact of irritable bowel syndrome: an international survey of 40 000 subjects. Alimentary Pharmacology & Therapeutics 2003;17(5):643-650.
3. **Thompson WG, Heaton KW, Smyth GT, et al.** Irritable bowel syndrome in general practice: prevalence, characteristics, and referral. Gut 2000;46(1):78-82.
4. **Gwee KA.** Irritable bowel syndrome in developing countries--a disorder of civilization or colonization? Neurogastroenterol Motil 2005;17(3):317-24.
5. **Drossman DA, Camilleri M, Mayer EA, et al.** AGA technical review on irritable bowel syndrome. Gastroenterology 2002;123(6):2108-2131.
6. **Bengtson MB, Ronning T, Vatn MH, et al.** Irritable bowel syndrome in twins: genes and environment. Gut 2006;55(12):1754-1759.
7. **Pimentel M, Lembo A, Chey WD, et al.** Rifaximin therapy for patients with irritable bowel syndrome without constipation. N Engl J Med 2011;364(1):22-32.
8. **Spiller R.** Review article: probiotics and prebiotics in irritable bowel syndrome. Aliment Pharmacol Ther 2008;28(4):385-96.
9. **Zoetendal EG, Rajilic-Stojanovic M, and de Vos WM.** High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. Gut 2008;57(11):1605-15.
10. **Collins SM and Bercik P.** The relationship between intestinal microbiota and the central nervous system in normal gastrointestinal function and disease. Gastroenterology 2009;136(6):2003-14.
11. **Salonen A, de Vos WM, and Palva A.** Gastrointestinal microbiota in irritable bowel syndrome: present state and perspectives. Microbiology 2010;156(Pt 11):3205-15.
12. **Carroll IM, Ringel-Kulka T, Keku TO, et al.** Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. Am J Physiol Gastrointest Liver Physiol 2011;301(5):G799-807.
13. **Jeffery IB, O'Toole PW, Ohman L, et al.** An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. Gut 2012;61(7):997-1006.
14. **Noor SO, Ridgway K, Scovell L, et al.** Ulcerative colitis and irritable bowel patients exhibit distinct abnormalities of the gut microbiota. BMC Gastroenterol 2010;10:134.
15. **Rajilic-Stojanovic M, Biagi E, Heilig HG, et al.** Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology 2011;141(5):1792-801.
16. **Rigsbee L, Agans R, Shankar V, et al.** Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable bowel syndrome. Am J Gastroenterol 2012;107(11):1740-51.
17. **Saulnier DM, Riehle K, Mistretta TA, et al.** Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology 2011;141(5):1782-91.
18. **Simren M, Barbara G, Flint HJ, et al.** Intestinal microbiota in functional bowel disorders: a Rome foundation report. Gut 2013;62(1):159-76.
19. **Lyra A, Rinttila T, Nikkila J, et al.** Diarrhoea-predominant irritable bowel syndrome distinguishable by 16S rRNA gene phylotype quantification. World J Gastroenterol 2009;15(47):5936-45.
20. **Malinen E, Rinttila T, Kajander K, et al.** Analysis of the fecal microbiota of irritable bowel syndrome patients and healthy controls with real-time PCR. Am J Gastroenterol 2005;100(2):373-82.
21. **Krogius-Kurikka L, Lyra A, Malinen E, et al.** Microbial community analysis reveals high level phylogenetic alterations in the overall gastrointestinal microbiota of diarrhoea-predominant irritable bowel syndrome sufferers. BMC Gastroenterol 2009;9:95.
22. **Pedersen N, Andersen NN, Végh Z, et al.** Ehealth: Low FODMAP diet vs Lactobacillus rhamnosus GG in irritable bowel syndrome. World Journal of Gastroenterology : WJG 2014;20(43):16215-16226.
23. **Vos WM.** Fame and future of faecal transplantations — developing next-generation therapies with synthetic microbiomes. Microbial Biotechnology 2013;6(4):316-325.
24. **Kajander K, Myllyluoma E, Rajilic-Stojanovic M, et al.** Clinical trial: multispecies probiotic supplementation alleviates the symptoms of irritable bowel syndrome and stabilizes intestinal microbiota. Aliment Pharmacol Ther 2008;27(1):48-57.
25. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
26. **Thompson WG, Longstreth GF, Drossman DA, et al.** Functional bowel disorders and functional abdominal pain. Gut 1999;45 Suppl 2:II43-7.

27. **Chumpitazi BP, Hollister EB, Oezguen N, et al.** Gut microbiota influences low fermentable substrate diet efficacy in children with irritable bowel syndrome. Gut Microbes 2014;5(2):165-75.
28. **Bastian M, Heymann S, and Jacomy M.** Gephi: An Open Source Software for Exploring and Manipulating Networks. in International AAAI Conference on Weblogs and Social Media. 2009: San Jose, California.
29. **Stewart JA, Chadwick VS, and Murray A.** Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. J Med Microbiol 2005;54(Pt 12):1239-42.
30. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-4.
31. **Zoetendal EG, Ben-Amor K, Akkermans AD, et al.** DNA isolation protocols affect the detection limit of PCR approaches of bacteria in samples from the human gastrointestinal tract. Syst Appl Microbiol 2001;24(3):405-10.
32. **Kajander K, Hatakka K, Poussa T, et al.** A probiotic mixture alleviates symptoms in irritable bowel syndrome patients: a controlled 6-month intervention. Alimentary Pharmacology & Therapeutics 2005;22(5):387-394.
33. **Rajilic-Stojanovic M, Heilig HG, Tims S, et al.** Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 2012.
34. **Lahti L, Salonen A, Kekkonen RA, et al.** Associations between the human intestinal microbiota, Lactobacillus rhamnosus GG and serum lipids indicated by integrated analysis of high-throughput profiling data. PeerJ 2013;1:e32.
35. **Sokol H, Pigneur Bnd, Watterlot L, et al.** Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proceedings of the National Academy of Sciences 2008;105(43):16731-16736.
36. **Manichanh C, Borruel N, Casellas F, et al.** The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol 2012;9(10):599-608.
37. **Tims S, Derom C, Jonkers DM, et al.** Microbiota conservation and BMI signatures in adult monozygotic twins. ISME J 2013;7(4):707-17.
38. **Lahti L, Salojärvi J, Salonen A, et al.** Tipping elements in the human intestinal ecosystem. Nat Commun 2014;5.
39. **Holmstrom K, Collins MD, Moller T, et al.** Subdoligranulum variabile gen. nov., sp. nov. from human feces. Anaerobe 2004;10(3):197-203.
40. **Schwiertz A, Hold GL, Duncan SH, et al.** Anaerostipes caccae gen. nov., sp. nov., a new saccharolytic, acetate-utilising, butyrate-producing bacterium from human faeces. Syst Appl Microbiol 2002;25(1):46-51.
41. **Del Dot T, Osawa R, and Stackebrandt E.** Phascolarctobacterium faecium gen. nov, spec. nov., a Novel Taxon of the Sporomusa Group of Bacteria. Systematic and Applied Microbiology 1993;16(3):380-384.
42. **Foubert EL and Douglas HC.** Studies on the Anaerobic Micrococci: II. The Fermentation of Lactate by Micrococcus lactilyticus. J Bacteriol 1948;56(1):35-6.
43. **Williams RAD, Bowden GH, Hardie JM, et al.** Biochemical Properties of Bacteroides melaninogenicus Subspecies. International Journal of Systematic Bacteriology 1975;25(3):298-300.
44. **Vaughan EE, Heilig HGHJ, Ben-Amor K, et al.** Diversity, vitality and activities of intestinal lactic acid bacteria and bifidobacteria assessed by molecular approaches. FEMS Microbiology Reviews 2005;29(3):477-490.
45. **Mortensen PB, Andersen JR, Arffmann S, et al.** Short-chain fatty acids and the irritable bowel syndrome: the effect of wheat bran. Scand. J. Gastroenterol. 1987;22(2):185-192.
46. **Treem WR, N.** A, Kastoff G, et al. Fecal short-chain fatty acids in patients with diarrhea-predominant irritable bowel syndrome: in vitro studies of carbohydrate fermentation. J. Pediatr. Gastroenterol. Nutr. 1996;23(3):280-286.
47. **Tana C, Umesaki Y, Imaoka A, et al.** Altered profiles of intestinal microbiota and organic acids may be the origin of symptoms in irritable bowel syndrome. Neurogastroenterol. Motil. 2010;22(5):512-519.
48. **Canani RB, Costanzo MD, Leone L, et al.** Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol 2011;17(12):1519-28.
49. **Fukumoto S, Tatewaki M, Yamada T, et al.** Short-chain fatty acids stimulate colonic transit via intraluminal 5-HT release in rats. Am J Physiol Regul Integr Comp Physiol 2003;284(5):R1269-76.
50. **Kamath PS, Hoepfner MT, and Phillips SF.** Short-chain fatty acids stimulate motility of the canine ileum. Am J Physiol 1987;253(4 Pt 1):G427-33.
51. **Scarpellini E, Lauritano EC, Lupascu A, et al.** Efficacy of butyrate in the treatment of diarrhoea-predominant irritable bowel syndrome. Digestive and Liver Disease Supplements 2007;1(1):19-22.
52. **Booijink CC, Zoetendal EG, Kleerebezem M, et al.** Microbial communities in the human small intestine: coupling diversity to metagenomics. Future Microbiol 2007;2(3):285-95.
53. **van den Bogert B, de Vos WM, Zoetendal EG, et al.** Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. Appl. Environ. Microbiol. 2011:Epub ahead of print.
54. **Pyleris E, Giamarellos-Bourboulis EJ, Tzivras D, et al.** The prevalence of overgrowth by aerobic bacteria in the small intestine by small bowel culture: relationship with irritable bowel syndrome. Dig Dis Sci 2012;57(5):1321-9.

55. **Ohman L and Simren M.** Intestinal microbiota and its role in irritable bowel syndrome (IBS). Curr Gastroenterol Rep 2013;15(5):323.
56. **Posserud I, Stotzer PO, Bjornsson ES, et al.** Small intestinal bacterial overgrowth in patients with irritable bowel syndrome. Gut 2007;56(6):802-8.
57. **Larsson E, Tremaroli V, Lee YS, et al.** Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. Gut 2012;61(8):1124-31.
58. **van Baarlen P, Troost F, van der Meer C, et al.** Human mucosal in vivo transcriptome responses to three lactobacilli indicate how probiotics may modulate human cellular pathways. Proc Natl Acad Sci U S A 2011;108 Suppl 1:4562-9.
59. **Salonen A, de Vos WM, and Palva A.** Gastrointestinal microbiota in irritable bowel syndrome: present state and perspectives. Microbiology 2010;156(11):3205-3215.
60. **Png CW, Linden SK, Gilshenan KS, et al.** Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. Am J Gastroenterol 2010;105(11):2420-2428.
61. **Joossens M, Huys G, Cnockaert M, et al.** Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut 2011;60(5):631-637.
62. **Lawson PA, Song Y, Liu C, et al.** Anaerotruncus colihominis gen. nov., sp. nov., from human faeces. Int. J. Syst. Appl. Microbiol. 2004;54(2):413-417.
63. **Holdeman LV and Moore WEC.** New Genus, Coprococcus, twelve new species, and emended description of four previously described species of bacteria from human feces. Int. J. Syst. Bacteriol. 1974;24(2):260-277.
64. **Finnie IA, Dwarakanath AD, Taylor BA, et al.** Colonic mucin synthesis is increased by sodium butyrate. Gut 1995;36(1):93-9.
65. **Ma X, Fan PX, Li LS, et al.** Butyrate promotes the recovering of intestinal wound healing through its positive effect on the tight junctions. J Anim Sci 2012;90 Suppl 4:266-8.
66. **Jalanka-Tuovinen J, Salojärvi J, Salonen A, et al.** Faecal microbiota composition and host–microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. Gut 2014;63(11):1737-1745.



Chapter 5

Fructo-oligosaccharides induce a shift from Bacteroidetes to *Bifidobacterium* and restructure gut microbiota in healthy subjects

Sebastian Tims, Johannes Snel, Sandra J.M. ten Bruggencate, Saskia van Schalkwijk, Harro Timmerman, Jos Boekhorst, Erwin G. Zoetendal, Willem M. de Vos, Michiel Kleerebezem

Submitted

Abstract

Studies on the microbiota effects of fibers and other non-digestible oligosaccharides have relied on methods that often fail to reveal the full extent of their impact on the bacterial community in the intestine. The goal of this study was to gain community wide information on the impact on the fecal microbiota of healthy human subjects of fructo-oligosaccharides (FOS) that are widely used as fibers. Fecal samples from a previously published double-blind, placebo-controlled crossover study design with twenty-eight healthy men were used for microbiota analyses. Subjects consumed either 20g of FOS or placebo per day for fourteen days. Fecal microbiota diversity, dynamics and composition in samples obtained on the last day of both supplement periods was determined by phylogenetic microarray analysis and complemented by group-specific analysis using pyrosequencing and qPCR analyses. FOS increased the abundance of bacteria within the *Bifidobacterium* genus. This enrichment came at the expense of the bacteria within the Bacteroidetes phylum. Although the butyryl-CoA:acetate CoA transferase gene content in the fecal material decreased, FOS consumption increased both positive and negative interactions between genera representing lactate-utilizing, and butyrate-producing members of the ecosystem, indicating a shift in the cooperative butyrate production pathway that is active in the ecosystem.

Introduction

It has become increasingly recognized that the gastrointestinal (GI) microbiota plays a critical role in human health, affecting nutrient utilization and adsorption by the host, the development and maturation of the immune system, and resistance to infections [1]. The best recognized strategy by which the composition and metabolic activity of the intestinal microbiota can be modulated is via the dietary introduction of fibers and other prebiotics [2]. A prebiotic has been defined as selectively fermented ingredients that result in specific changes in the composition and/or activity of the gastrointestinal microbiota, thus conferring benefit(s) upon host health [3]. To have these functions, the used prebiotic should be resistant to gastric acidity, hydrolysis by mammalian enzymes, and to GI absorption. In addition, they should be fermented by the intestinal microbiota and selectively stimulate the growth and/or activity of those intestinal bacteria that contribute to health [4]. Several non-digestible oligosaccharides, including inulin, resistant starch, galacto-oligosaccharides (GOS) and fructo-oligosaccharides (FOS) are added as a prebiotic to a variety of food products.

FOS are naturally found in varying concentrations in a variety of foods such as wheat, bananas, asparagus, and garlic. Moreover, they are supplemented as prebiotics to several products, such as various dairy products and infant formulas. FOS are composed of linear chains of fructose moieties, linked by ß(2→1) bonds and often contain a glucose moiety at their reducing end. The number of fructose moieties varies in different FOS preparations, and ranges from 3 to 10. FOS are not hydrolyzed by human small intestinal glycosidases and hence are considered to reach the colon intact, where they are subject to microbial degradation and serve as a carbon source for fermentation [5].

Intervention studies have convincingly shown that FOS supplementation can promote growth of bifidobacteria in the intestinal tract of infants and adults [6-9]. However, over 1,000 species have now been described in the human gut [10] and the impact of FOS on other members of the intestinal microbiota other than bifidobacteria is generally less well understood, especially for those bacteria that cannot be readily cultured in the laboratory. The latter is exemplified by the notion that the majority of the studies addressing the modulation of the intestinal microbiota by FOS have relied on culture-based enumeration methods [6-9, 11] that failed to detect the majority of microbial species present in the human gut. Additionally, the studies that employed culture-independent molecular methods to unravel the effects of FOS used quantitative real time (qRT)-PCR [12] or fluorescent *in situ* hybridization [13], which are restricted to selected bacterial groups of interest and fail to reveal effects on the microbiota as a whole. More recent studies focused on FOS that is present in different prebiotic or fiber mixtures, for example combined with inulin [14], or have investigated potential microbial genes involved in the breakdown of different prebiotics such as FOS [15]. The aim of this study was to determine the FOS-induced shifts on overall microbiota composition by high throughput phylogenetic microarray analysis using the Human Intestinal Tract Chip (HITChip) [16]. The microbiota composition information was complemented and validated by barcoded-pyrosequencing and qRT-PCR analysis of specific groups. Moreover, since the health benefit of FOS has been proposed to be attributable to butyrate production we also assessed the butyrate production potential of the microbiota by

evaluating the abundances of butyrate producing microbial groups and their ecosystem network associations, as well as the abundance of the butyryl-CoA:acetate CoA transferase genes within the community.

Materials & methods

Design intervention study

Fecal samples obtained during a previously published double-blind, placebo-controlled crossover FOS supplementation study that encompassed two supplement periods (FOS and placebo) of two weeks separated by a washout period of two weeks were used [17]. Participants (n=28) were 27.7 (±1.7 SEM) years and had a mean BMI of 23.2 (±0.5 SEM) kg/m^2. Participants had no history of GI disease, surgical operations of either the small or large intestine, and no lactose intolerance. Prebiotics, antibiotics, immunosuppressive drugs, anti-diarrhea drugs, laxatives, or probiotics were not used in the 3 months prior to the study.

Throughout the study, participants were instructed to maintain their usual pattern of physical activity and maintain their habitual diet but to abstain from all dairy products, foods with high calcium content and foods containing prebiotics or large amounts of non-digestible fibers and probiotics. Participants consumed either lemonade with 20 g of FOS (purity 93%, Raftilose P95, Orafti, Tienen, Belgium) per day in the FOS supplementation period or 6 g of sucrose per day in the placebo period. The sucrose in the placebo lemonade ensured equal taste, viscosity and color of both lemonades. The daily dose of lemonade was divided into 3 portions, which had to be consumed in the morning, afternoon and evening. On the last day in both supplement periods, 24 h fecal samples were collected for microbiological analyses. Fecal samples were frozen on dry ice immediately after defecation and stored at -40°C. Afterwards, samples were weighed, freeze-dried and subsequently ground to obtain homogeneous powdered samples.

Fecal DNA extraction

Total DNA from fecal material was extracted using a repeated bead beating protocol, as reported before [18]. DNA yield was quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). DNA concentration was adjusted to 10 ng/µl and was used as a template for PCR amplification. Sequence polymorphism analyses of mitochondrial DNA were used to verify the origin of the fecal samples [19].

Phylogenetic microarray analysis

DNA amplification, RNA (reverse-)transcription and HITChip microarray hybridization were performed as described previously [20]. Bacterial profiles were obtained by extracting the normalized signal values of all 3,699 unique HITChip probes. These probes are assigned to several phylogenetic levels of which two were used here: the order-like 16S rRNA gene sequence group level and genus-like 16S rRNA gene sequence group level (sequence similarity >90%). Relative abundances were calculated for the order-like and genus-like groups by summing all probe signals targeting a group and dividing this by the sum of all probe signals. The cumulative

relative abundance of genus-like groups representing characterized isolates known to be able to butyrate production (Table S5.2) in the microbial ecosystem of the intestinal tract was estimated by summing the probes signals representing bacterial isolates known to produce butyrate and compare this to the total signal of all probes. Normalized signal values of all HITChip probes were used to calculate the inverse Simpson's index of Diversity, richness and evenness (Pielou's evenness index) for each sample and the Spearman's correlation coefficient between samples. Ward's minimum variance method was used for the construction of hierarchical clusters of the total microbiota probe profiles with a distance matrix based on the Spearman's correlation coefficients between the samples (complete observation correlations).

Because of the cross-over design, pair-wise dependent two-group Wilcoxon signed rank tests were used, combined with FDR correction [21], to examine whether statistically significant differences existed between the placebo and FOS supplementation period ($p<0.05$ and $q<0.05$). Co-(and anti)occurrence networks were built using Spearman's correlation coefficient (ϱ) calculated on the relative abundances of the genus-like groups that were present at >0.1% abundance in more than half of the subjects. A permutation test was performed to test whether the observed ϱ values were significantly different from zero with 10,000 permutations. These ϱ values, or co-(and anti)occurrences, from the network analysis were regarded as connectivity measures for ecosystem interactions. Assessing the connectivity differences between the two treatment period networks (i.e. differences between the taxon pair correlations of each treatment period) was used to show modulations in ecosystem interactions, which could suggest possible metabolic interplay between the involved members. To test the significance of the difference between the taxon pair correlations of each treatment period another permutation test was performed in which samples were randomly assigned samples to the placebo and FOS periods (with 10,000 permutations), and subsequently it was determined how often taxon pairs were formed with larger difference as observed in the real data. To exclude significant differences between weak correlations (e.g. with $\varrho=0.3$ in one treatment period and $\varrho=-0.3$ in the other treatment period), from the analysis a correlation cut-off of $\varrho>0.5$ or $\varrho<-0.5$ (in at least one of the treatment periods) was employed. The R package 'network' [22] and the 'Gephi Graph Visualisation and Manipulation' software [23] were used to visualize the co-(and anti)occurrences.

Pyrosequencing of 16S rRNA amplicons of the *Bifidobacterium* genus

Bifidobacterium specific primers were used to amplify part of the 16S rRNA gene which was subsequently sequenced and phylogenetically reconstructed down to the species level. Phylogenetic coverage and entropy of a set of literature derived primers targeting the V2-V5 region of 620 full length 16S rRNA sequences from *Bifidobacterium* spp. was calculated to identify primer pairs targeting the majority of gastrointestinal derived bifidobacteria and providing high phylogenetic resolution. In addition, primer combinations were tested against a selection of *Bifidobacterium* isolates. Two primers pairs, Bif164f - Bif662r [24] and 338f [25] - Bif662r, were selected and validated on two mock communities, consisting of four and ten *Bifidobacterium* species, and eight fecal samples, with and without spiking of the two mock communities, subjected to 454 sequencing and analyzed with a workflow based on QIIME v1.2 [26] using settings as

recommended in the QIIME 1.2 tutorial, with the following exceptions: reads were filtered for chimeric sequences using Chimera Slayer [27] and OTU clustering was performed with settings as recommended [28] using an identity threshold of 99%. Taxonomic assignment up to the species level and evaluation of ability of the amplicon to differentiate between different species of *Bifidobacterium* was done as described previously [29]. Primer pair 338f and Bif662r was most efficient in identification of the different *Bifidobacterium* species (data not shown) and therefore selected for the current cohort. For the preparation of the amplicon pool for pyrosequencing, the following primers were applied: forward primer, 5'- CCA TCT CAT CCC TGC GTG TCT CCG ACT AG – NNNNNN – ACT CCT ACG GGA GGC AGC AG-3' (454 Life Sciences primer A – sample-specific six-base barcode tag – primer 338F); reverse primer 5'-CCT ATC CCC TGT GTG CCT TGG CAG TCT CAG – CCA CCG TTA CAC CGG GAA-3' (454 Life Sciences primer B – primer Bif662). PCR amplification mixture contained: 1 µL faecal DNA, 1 µL bar-coded forward primer, 15 µL master mix (1 µL KOD Hot Start DNA Polymerase (1 U/µL; Novagen, Madison, WI, USA), 5 µL KOD-buffer (10×), 3 µL $MgSO_4$ (25 mM), 5 µL dNTP mix (2 mM each), 1 µL (10 µM) of reverse primer) and 33 µL sterile water (total volume 50 µL). PCR conditions were: 95°C for 2 minutes followed by 35 cycles of 95°C for 20 s, 55°C for 10 s, and 70°C for 15 s. The PCR product was subsequently purified using the MSB Spin PCRapace kit (Invitek) and the concentration was checked with a Nanodrop ND-1000 spectrophotometer. A composite sample for pyrosequencing was prepared by pooling 200 ng of these purified PCR products of each sample. The pooled sample was electrophoresed on a 1% agarose gel and the approximately 400 bp band was excised and extracted from the agarose gel with the MinElute Gel Extraction kit (Qiagen, Venlo, The Netherlands) and submitted for pyrosequencing on the 454 Life Sciences GS-FLX platform using Titanium sequencing chemistry (GATC-Biotech, Germany). The significance of putative differences in relative abundance of specific taxa between sample groups was calculated using the Mann-Whitney U test as implemented in SciPy [30]. Principal component analysis (PCA) and partial regularized discriminant analysis (RDA) on OTU abundance (99% clustering) of the fecal bifidobacterial population was performed using CANOCO5.0 (Microcomputer Power, USA) according to the manufacturer's instructions [31]. For the RDA Statistical significance was assessed by MCPP with 499 random permutations under the full model. Variance attributable to the variable intervention period (placebo or FOS supplementation) was removed by covariate analyses.

Quantitative PCR analysis

Real-time quantitative PCR was used to simultaneously quantify phylogenetically distinct butyrate-producing bacteria, by targeting the butyryl-CoA:acetate CoA transferase encoding genes, which has previously been described as an approach to quantify the abundance of this butyrate production pathway in the intestinal ecosystem [32] (primers and PCR conditions see SI materials and methods). Previously, real-time quantitative PCR was used to quantify the *Bifidobacterium* genus within the samples of this cohort [17]. In the present study, we expand this with quantitative PCR analyses to simultaneously quantify three genera making up a large proportion of the Bacteroidetes phylum: *Bacteroides*, *Prevotella*, and *Porphyromonas* (together

called the BPP-group). The PCR was based on the detection of the 16S rRNA genes from the BPP-group as was described previously [33] (primers and PCR conditions see SI materials & methods). Pair-wise dependent two-group Wilcoxon signed rank tests were performed to identify statistically significant differences between the placebo and FOS supplementation period ($p<0.05$).

Results

Microbiota profiles

Phylogenetic profiling of the fecal microbiota by HITChip analysis was performed to obtain a high-resolution, high-sensitive, and complete microbiota profile of each individual following the FOS and placebo dietary interventions. The most abundant phyla consisted of Firmicutes, followed by Bacteroidetes and Actinobacteria. For the total microbiota there was no significant difference in diversity, evenness or richness between the FOS and control sample groups. To obtain insight in the overall similarity of the microbiota between all the study subjects, HITChip profiles were hierarchically clustered (Figure 5.1). This analysis revealed a clear, subject-based clustering, confirming previous findings that indicate that the largest difference in the microbiota profiles is related to inter-individual differences [34-36]. This is clearly illustrated by the mean inter-individual Spearman correlation coefficient (ϱ) of 0.710 (±0.057 SEM), which is substantially lower than the intra-individual ϱ of 0.859 (±0.051 SEM). However, the intra-individual microbiota similarity of subject 22 was strikingly low (ϱ=0.721) indicating a profound effect of the intervention or another unknown confounding factor, especially since these results were consistently obtained in independent repetitions of the analyses, and sample- and subject-identities were confirmed by sequence polymorphism analysis of amplified human mitochondrial DNA [19] (Table S5.1).

Effect of FOS on fecal microbial communities

Comparison of microbiota profiles from feces obtained after FOS and placebo consumption revealed that FOS consumption resulted in a significant increase in Actinobacteria compared to the placebo (Figure 5.2A). This change can be exclusively attributed to an increase of the *Bifidobacterium* group (Figure 5.2B and Table 5.1). In addition, the diversity within the Actinobacteria group was significantly decreased by FOS consumption, supporting that FOS stimulated the growth of *Bifidobacterium* specifically (Figure 5.2C). No other order-like groups showed significant difference in diversity between consumption periods (data not shown). To investigate whether the increase in *Bifidobacterium* was specifically stimulating the expansion of certain members of this genus, a group-targeted barcoded pyrosequencing approach was employed. The majority of the bifidobacterial OTUs (99% clustering) were closely related to *B.longum* and OTUs related to *B.adolescentis*, *B.bifidum*, and *B.pseudocatenulatum* were also detected. No stimulation of specific bifidobacterial OTUs was observed between the sample groups of either consumption period (Figure 5.S1). Furthermore, principal component analysis based on OTU abundance was performed, where variance attributable to the variable intervention period was

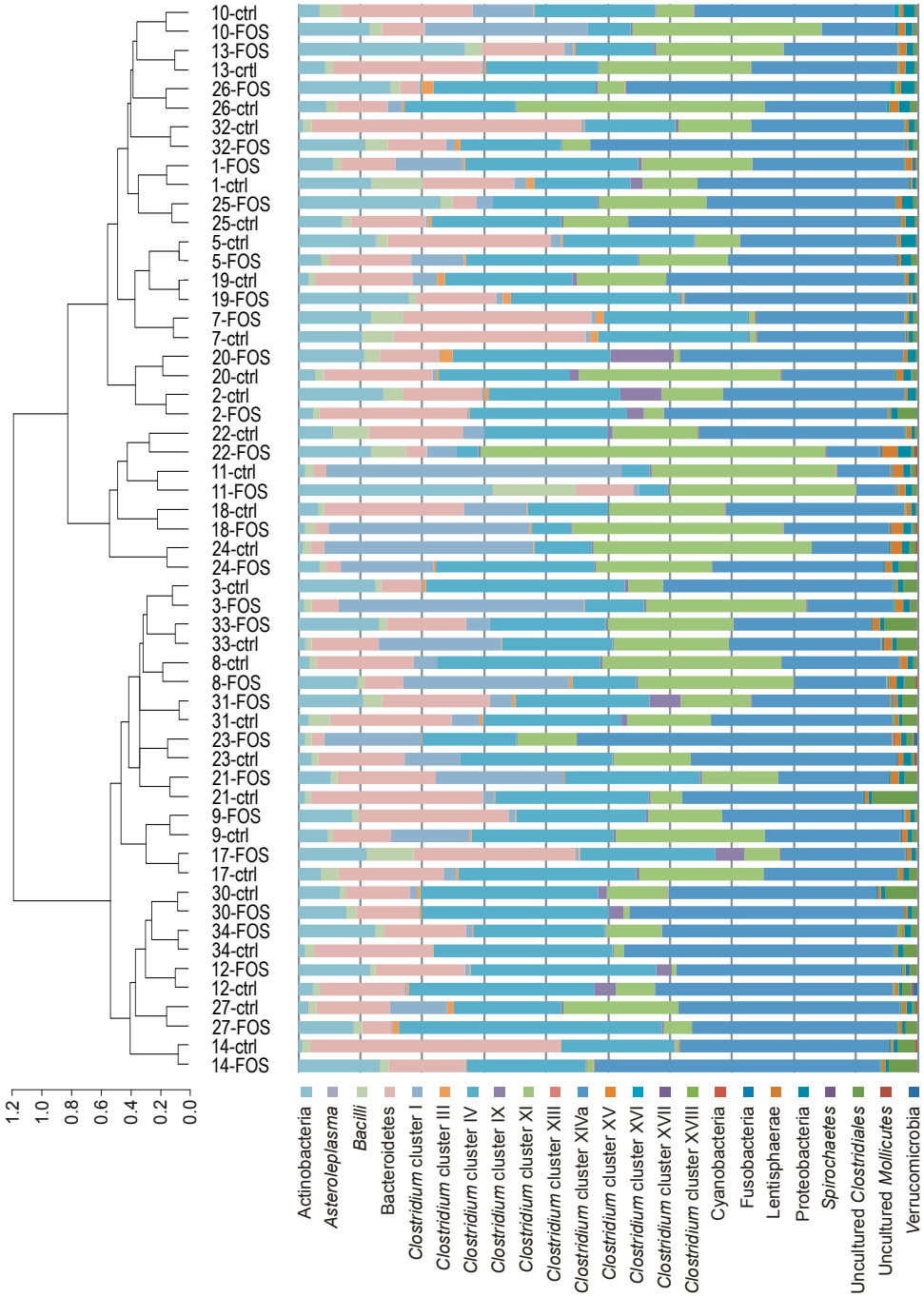


Figure 5.1 Abundance of phylum-like bacterial taxa in fecal samples of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Subject numbers are appended with: "-FOS" for the samples from the FOS consumption period, and "-ctrl" for the placebo consumption period.

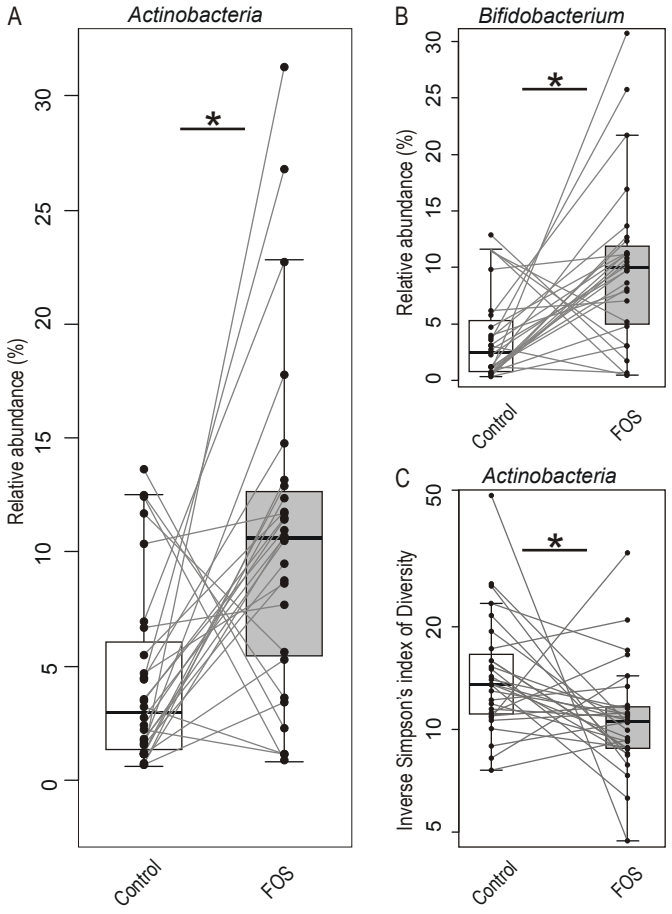


Figure 5.2 Actinobacteria differences in feces of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of FOS and placebo consumption periods. Relative abundance significantly increased in samples from the FOS consumption period for Actinobacteria (A; p = 0.002, q = 0.011), mainly due to the genus-like Bifidobacterium group (B; p = 0.002, q = 0.011). Diversity of Actinobacteria significantly decreased in samples from the FOS consumption period (C; p = 0.011; q = 0.034). Asterisk marks significantly different results.

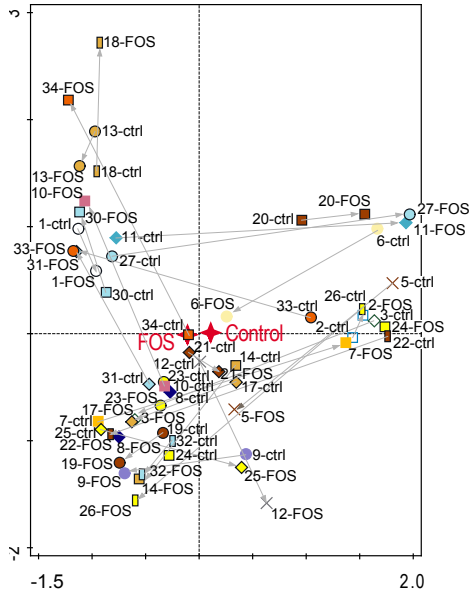


Figure 5.3 Principal component analysis based on OTU abundance (99% clustering) of the fecal bifidobacterial population of healthy individuals fed either placebo or inulin in a randomized, crossover study. Variance attributable to the variable intervention period was removed by covariate analyses. The supplementary variables subject and treatment accounted for 55.8% of the observed variance in microbiota composition of which individuality accounted for the highest percentage (54.4%).

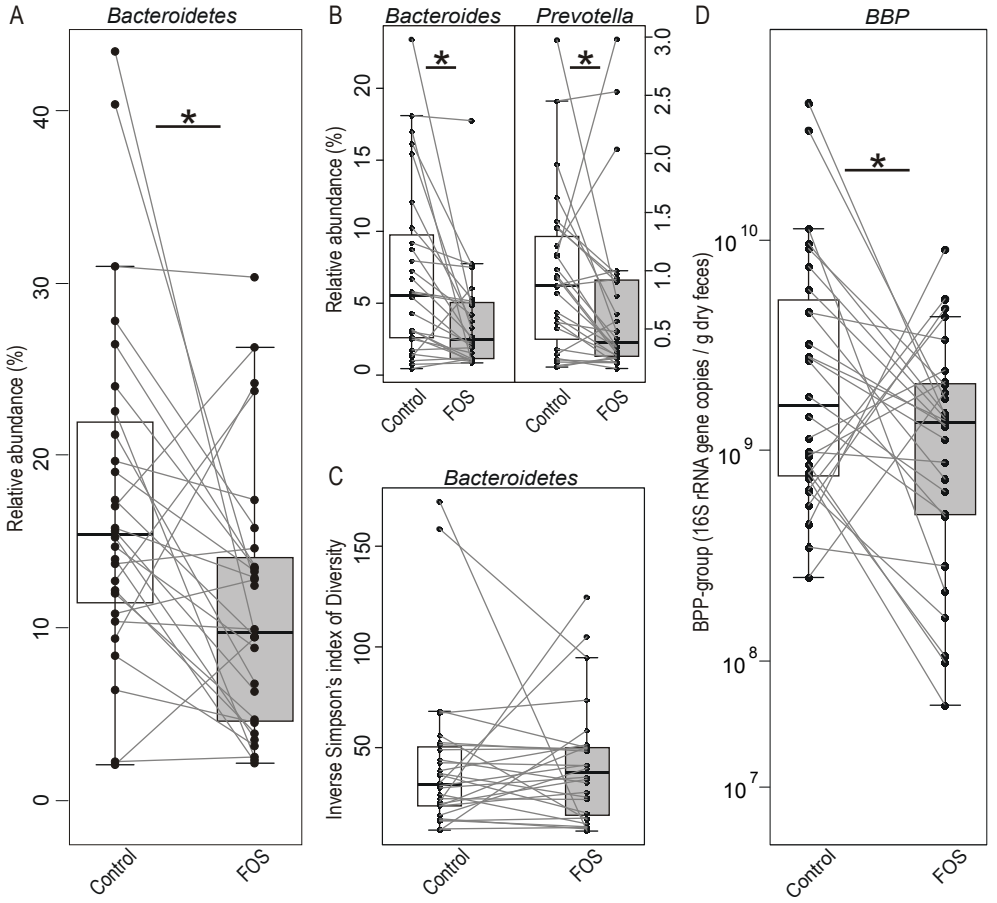


Figure 5.4 Bacteroidetes differences in feces of 28 healthy subjects determined with the Human Intestinal Tract (HIT) Chip and real-time quantitative PCR. Samples were taken at the end of the FOS and placebo consumption periods. Relative abundance significantly decreased in samples from the FOS consumption period for Bacteroidetes (A; p = 0.005, q = 0.013), which is mainly due to *Bacteroides* and *Prevotella* genus-like groups (more details see Table 5.1). For *Bacteroides* the combined relative abundances of the relatives of *Bacteroides fragilis, B.intestinalis, B.ovatus, B.plebeius, B.splachnicus, B.stercoris, B.uniformis, B.vulgatus*, and Uncultured Bacteroidetes are shown (B; p = 3.66e-05, q = 0.001). For *Prevotella* the combined relative abundances of the relatives of *Prevotella oralis, P. ruminicola*, and *P. tannerae* are shown (B; p = 0.013, q = 0.039). Diversity of Bacteroidetes was not significantly different in samples from the FOS consumption period (C; p = 0.920, q = 0.335), indicating a nearly equal decrease of all Bacteroidetes members. 16S rRNA gene copies / g of feces significantly decreased in samples from the FOS consumption period for the Bacteroides-Prevotella-Porphyromonas (BPP)-group (D; p = 0.029). Asterisk marks significantly different results.

removed by covariate analyses (Figure 5.3). The supplementary variables subject and treatment accounted for 55.8% of the observed variance in microbiota composition of which individuality accounted for the highest percentage (54.4%). Centroids for the treatment groups are adjacent, suggesting no clear impact of FOS intervention on bifidobacterial OTU distribution. This was confirmed by partial RDA (MCPP, p=0.606, after removal variance attributed by intervention period and subject). Individuality was the only significant signal (partial RDA: MCPP, p=0.01).

Table 5.1 Actinobacteria and Bacteroidetes subgroup differences in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of the FOS and placebo consumption periods. Significant changes ($p < 0.05$, $q < 0.05$) are indicated in bold.

		Control group relative abundances		FOS group relative abundances		p-	q-
Phylum	**Genus-like group**	**Mean %**	**(± Std.err)**	**Mean %**	**(± Std.err)**	**value**	**value**
Actinobacteria	*Actinomycetaceae*	0.004	(±0.000)	0.005	(±0.000)	0.138	0.139
	Atopobium	0.017	(±0.001)	0.020	(±0.001)	0.053	0.114
	Bifidobacterium	3.810	(±0.757)	10.028	(±1.360)	**0.002**	**0.011**
	Collinsella	0.449	(±0.062)	0.481	(±0.056)	0.186	0.171
	Corynebacterium	0.013	(±0.001)	0.016	(±0.001)	0.050	0.114
	Eggerthella lenta et rel.	0.073	(±0.008)	0.080	(±0.008)	0.236	0.197
	Micrococcaceae	0.009	(±0.001)	0.011	(±0.001)	0.202	0.183
	Propionibacterium	0.026	(±0.002)	0.039	(±0.012)	0.695	0.373
Bacteroidetes	***Allistipes et rel.***	1.748	(±0.429)	0.919	(±0.164)	**<0.001**	**0.002**
	Bacteroides fragilis et rel.	0.710	(±0.171)	0.342	(±0.062)	**0.001**	**0.010**
	Bacteroides intestinalis et rel.	0.357	(±0.066)	0.207	(±0.052)	**0.007**	**0.029**
	Bacteroides ovatus et rel.	0.868	(±0.168)	0.497	(±0.100)	**0.001**	**0.010**
	Bacteroides plebeius et rel.	1.239	(±0.241)	0.661	(±0.151)	**<0.001**	**0.003**
	Bacteroides splachnicus et rel.	1.102	(±0.239)	0.591	(±0.101)	**<0.001**	**0.002**
	Bacteroides stercoris et rel.	0.689	(±0.152)	0.356	(±0.075)	**0.002**	**0.011**
	Bacteroides uniformis et rel.	0.722	(±0.140)	0.393	(±0.083)	**0.016**	0.051
	Bacteroides vulgatus et rel.	1.260	(±0.262)	0.573	(±0.139)	**<0.001**	**<0.001**
	Parabacteroides distasonis et rel.	1.453	(±0.226)	0.772	(±0.129)	**<0.001**	**0.004**
	Prevotella melaninogenica et rel.	4.628	(±1.020)	4.324	(±1.076)	0.537	0.348
	Prevotella oralis et rel.	0.455	(±0.081)	0.313	(±0.060)	**0.007**	**0.029**
	Prevotella ruminicola et rel.	0.350	(±0.057)	0.235	(±0.049)	**0.007**	**0.029**
	Prevotella tannerae et rel.	0.905	(±0.130)	0.632	(±0.136)	**0.009**	**0.034**
	Tannerella et rel.	0.714	(±0.154)	0.436	(±0.073)	**<0.001**	**0.002**
	Uncultured Bacteroidetes	0.118	(±0.050)	0.053	(±0.019)	**0.002**	**0.013**

The expansion of the *Bifidobacterium* genus-like group appeared to come at the expense of one order-like bacterial group, the Bacteroidetes (Figure 5.4). Contrary to the attribution of the expansion of the Actinobacteria group to a single genus-like group (*Bifidobacterium*), the decrease of the Bacteroidetes relative abundance was generally quite evenly distributed over its subgroups (Table 5.1).

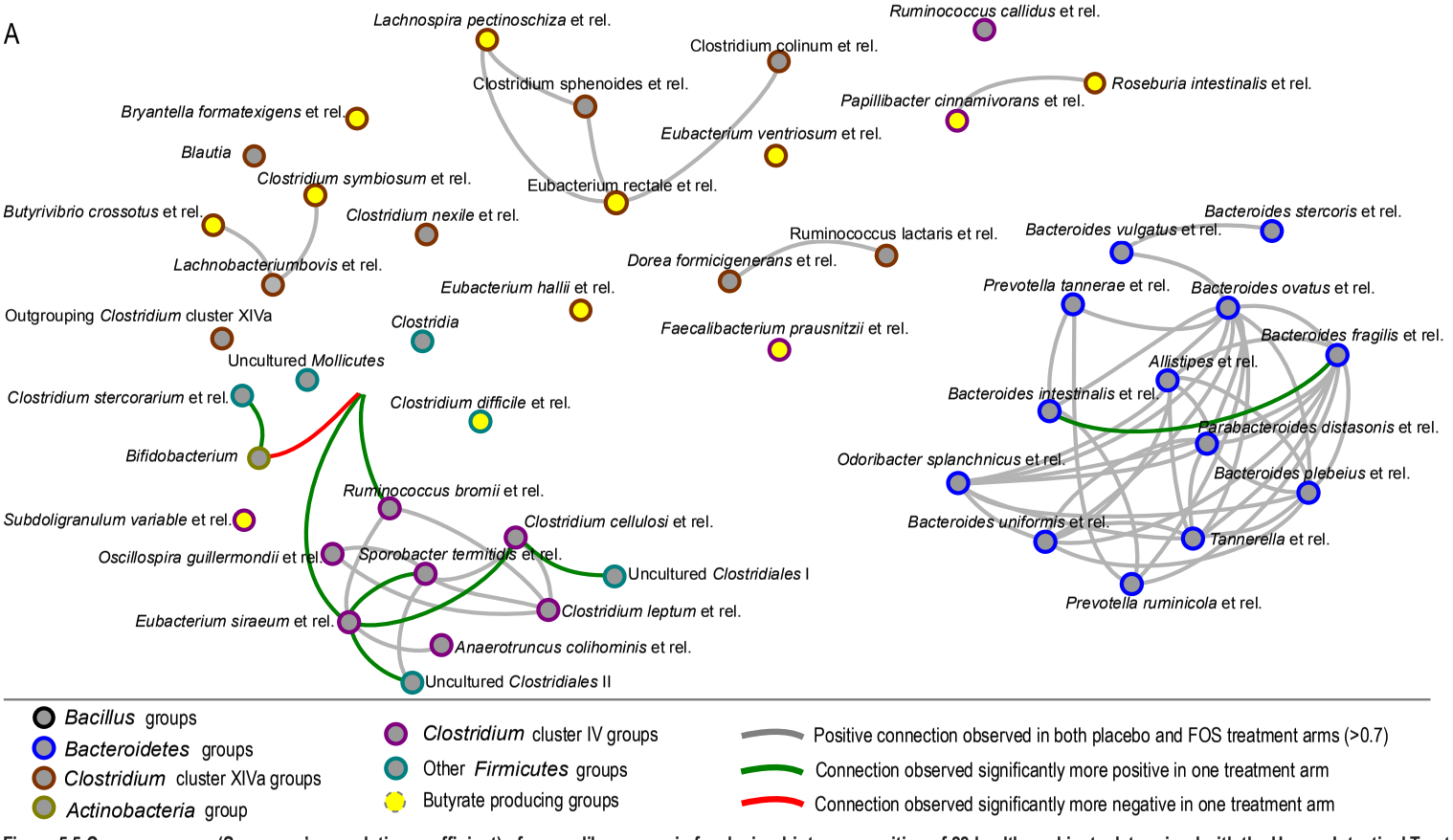


Figure 5.5 Co-occurrences (Spearman's correlation coefficient) of genus-like groups in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Yellow nodes indicate groups with known butyrate producers. Grey lines represent strong co-occurrences between the microbial groups that were found in both treatment periods ($p > 0.7$). A) Green lines represent the co-occurrences that are significantly higher ($p < 0.05$; permutation test) between the microbial groups in the samples from the placebo treatment period. Red lines represent anti-occurrences that are significantly lower ($p < 0.05$; permutation test) between the microbial groups in the samples from the placebo treatment period.

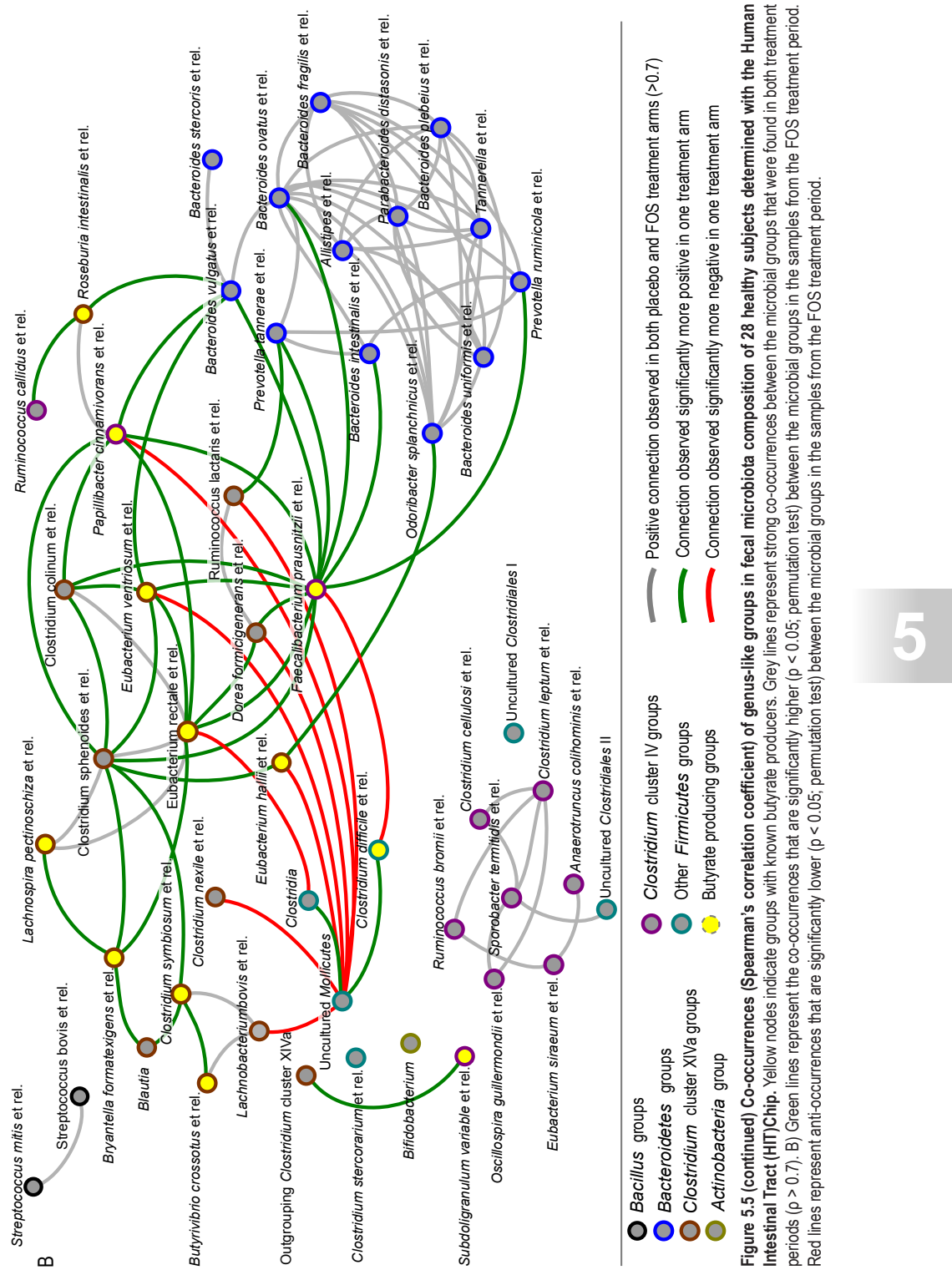


Figure 5.5 (continued) Co-occurrences (Spearman's correlation coefficient) of genus-like groups in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Yellow nodes indicate groups with known butyrate producers. Grey lines represent strong co-occurrences between the microbial groups that were found in both treatment periods ($p > 0.7$). B) Green lines represent the co-occurrences that are significantly higher ($p < 0.05$; permutation test) between the microbial groups in the samples from the FOS treatment period. Red lines represent anti-occurrences that are significantly lower ($p < 0.05$; permutation test) between the microbial groups in the samples from the FOS treatment period.

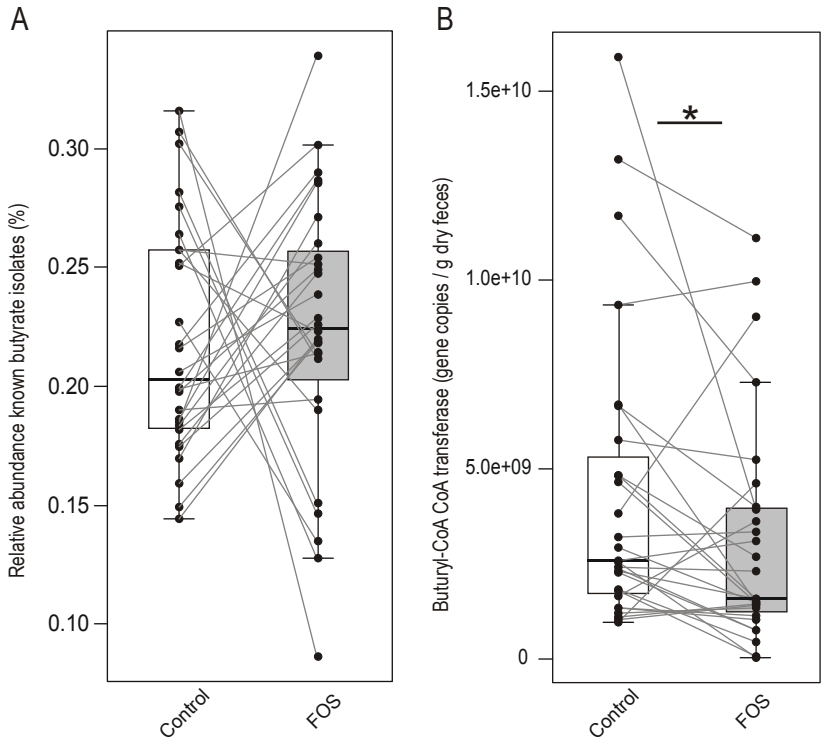


Figure 5.6 Differences in fecal microbiota composition of 28 healthy subjects determined with Human Intestinal Tract (HIT)Chip and real-time quantitative PCR. Samples were taken at the end of the FOS and placebo consumption periods. Relative abundance of the combined signal of all HITChip probes targeting known butyrate producing isolates did not significantly change (A; p = 0.630). Butyryl-CoA:acetate CoA transferase gene copies / g of feces significantly decreased in samples from the FOS consumption period (B; p = 0.022). Asterisk marks significantly different results.

Effect of FOS consumption on fecal microbial interactions

Co-occurrence analysis was performed to determine the effect of FOS consumption on the ecological connections within the fecal microbiota. Consistent connectivity was detected among the members of the Bacteroidetes phylum as well as among a group of *Clostridium* cluster IV members (together with uncultured *Clostridiales* taxa) that resembles a network module seen before in a healthy twin cohort ([37]; **Chapter 3**; Figure 5.5). However, the fecal microbiota obtained from subjects in the control period showed more connections with and among this previously identified *Clostridium* cluster IV module, than that derived from subjects consuming FOS (Figure 5.5A). A higher connectivity can be an indication for a more extensive metabolic relationship between these bacterial groups within the ecosystem, revealing an alignment of the functional capacities of these bacterial groups under the conditions tested.

Outside the *Clostridium* cluster IV module the connectivity differences between the samples from placebo and FOS consumption periods were reversed. Relative to the microbiota obtained from subjects on the placebo diet, those obtained from the FOS consumption period showed more, and significantly stronger connections, both positive and negative, mainly among the *Clostridium* cluster XIVa members or linking the cluster XIVa members to other *Clostridium*

clusters and members of the Bacteroidetes phylum (Figure 5.5B). The broad phylogenetic distribution of co- and anti-occurrence connections in the microbiota from the FOS period, included an increase in connections with genus-like groups that are among the known butyrate-producing members of the intestinal community (12 predicted butyrate-producing groups gained additional connections), as compared to the networks identified in the microbiota from the placebo period (no butyrate-producing groups gained additional connections). This observation could imply that FOS consumption elicits changes in the butyrate production capacity of the microbiota community.

Effect of FOS on microbiota butyryl-coenzyme A (CoA) CoA transferase abundance

To assess whether the butyrate production potential of the microbiota was affected by FOS consumption, all bacterial groups with known butyrate-producing representatives on the HITChip were further analyzed. The cumulative relative abundance of bacterial groups harboring known butyrate-producing isolates showed no consistent change between the samples from the placebo and FOS consumption periods (p=0.630; Figure 5.6A). To verify more directly whether FOS consumption affected the butyrate production potential of the microbiota, a previously described degenerate qPCR approach was employed that generically quantifies the abundance of butyryl-coenzyme A (CoA) CoA transferase encoding (EC2.8.3.8) genes within the microbiota, irrespective of their phylogenetic origin, thereby providing an estimate of the butyrate-producing capacity of the system [32]. FOS consumption led to a 1.4 fold decrease in the presence of the butyrate-producing enzyme butyryl-CoA transferase (p=0.022; Figure 5.6B). However, since the butyryl-CoA transferase reflects only one of several known pathways for butyrate production, this gene quantification may not fully reflect the total butyrate production capacity (see also discussion).

Discussion

In this study we describe the impact of FOS consumption on the total fecal microbiota community composition using culture-independent technologies. Our cohort, which is comprised of healthy young males, displays typical GI microbiota community compositions found human volunteers, which are dominated by Bacteroidetes, Firmicutes, and Actinobacteria [38-40]. Furthermore, hierarchical clustering of the microbiota profiles shows that the GI microbiota is host-specific as expected from earlier studies [37, 41-44] (Figure 5.1), and that the impact of the dietary intervention is substantially smaller than the intrinsic difference between the individuals. Nevertheless, some subjects' microbiota responded much stronger to FOS consumption (e.g. subject 22) compared to the average response determined in the group. Intriguingly, none of the available metadata of subject 22 appeared to explain the enhanced response to the FOS diet [17], which could imply the FOS effect-size may to a certain extent depend on the 'baseline' composition of the subject's microbiota. Importantly, besides the exception of subject 22, we concluded that FOS supplementation elicits no drastic microbiota-wide modulations and generally induces specific changes.

The definition of prebiotics defines these compounds to be 'selectively fermented' and able to induce changes in the GI microbiota that are 'specific' [3]. Previous studies employing classical cultivation enumerations and/or 16S rRNA gene-targeted FISH indeed found a specific increase in *Bifidobacterium* species upon consumption of inulin and FOS (as reviewed by Kolida and Gibson [45]). The present study employs a phylogenetic microarray, which provides a cultivation-independent, high resolution, and deep analysis of the overall GI microbiota, which is congruent with pyrosequencing analyses [46, 47] and next-generation metagenome sequencing [48]. Even with the broad phylogenetic spectrum analyzed by the HITChip, this study underpins that the microbiota changes elicited during FOS consumption are remarkably restricted to a small number of bacterial groups and most prominently affect the members of the *Bifidobacterium* genus. These findings verify previously reported qPCR analyses performed within the same cohort study [17] and provides clear support for the 'specificity' of FOS as a prebiotic. Moreover, these findings confirm a previous study [45] that also reported that the only bacteria consistently increased in abundance in response to FOS consumption were members of the *Bifidobacterium* genus (Figure 5.1; Table 5.1). This response to FOS consumption was dependent on *Bifidobacterium* abundance at baseline, with higher fold increase of abundance in subjects with lower baseline abundance levels of this bacterial group (ϱ=-0.57, p=7.33e-5 ANOVA; Figure S5.2), which corroborates previously reported findings [7, 49]. The qPCR determinations of specific microbial groups that were reported on this cohort previously [17], not only highlighted an increase in *Bifidobacterium* abundance, but also a significant increase in the bacteria associated with the genus *Lactobacillus*. The results presented here tend to confirm this slight increase in the *Lactobacillus* groups represented in the HITChip probe sets, which include bacteria related to *Lactobacillus gasseri*, *Lactobacillus plantarum*, and *Lactobacillus salivarius*, although the increase of these *Lactobacillus* groups as detected by the HITChip was not significant (Table S5.2).

While a significant abundance expansion of the members of the genus *Bifidobacterium* was observed, the majority of the Bacteroidetes phylum members significantly decreased in abundance. Previously reported factors that were associated with FOS consumption could potentially contribute to the reduced abundance of Bacteroidetes: pH decrease, likely caused by the bifidobacteria increase, combined with increased lactate concentrations and/or the reported mucin production increase [17]. Although the previously reported pH decrease was not significant, it has been shown that Bacteroidetes are sensitive to lowering of pH [50], especially in combination with the presence of substantial amounts of organic acids like lactate, implying that local accumulation of lactate and pH lowering may contribute to the Bacteroidetes decline. An overall *Bifidobacterium* increase combined with a general Bacteroidetes decrease has also been reported in recent studies on the impact of the consumption of different prebiotics on the GI microbiota that employed different analytical technologies. For example, the impact of GOS consumption was explored with barcoded pyrosequencing [51], whereas a panel of qPCR approaches and FISH were employed to determine the (long-term) impact of inulin consumption [52, 53]. The congruency of the results presented in these studies implies that these microbiota changes represent a quite consistent effect of non-digestible oligosaccharides. Nevertheless, individual studies also generated some variable results. Firstly, a long-chain inulin

(degree of polymerization: 50 – 103) intervention reported on a significant increase in the levels of *Atopobium* [53], which is not observed in the present study and may relate to the substantial difference in chain length of the prebiotic used, but may also relate to differences in the dietary regime employed during the study, such as calcium content which was restricted in the present study. Secondly, some previous studies have reported increases of the relative abundance of *Faecalibacterium prausnitzii* using oligosaccharide intakes regimes of 5.0 gram GOS [51] or 10.0 inulin [52]. The present study employed a substantially higher daily dose of oligosaccharides (20 gram) and failed to detect an increase in the population abundance of *F. prausnitzii*, which is in agreement with results reported for a inulin supplementation study employing daily dosages of 7.7 and 15.4 gram [54]. Hence, the commonly observed increase of *Bifidobacterium* and decline of Bacteroidetes phylum members appears to be the most robust finding of these studies.

Some studies indicated that specific lineages or species of *Bifidobacterium*, especially *B.adolescentis*, are responding to the consumption of GOS and very-long-chain inulin [51]. The present study did not reveal such specific increase in any of the members of the *Bifidobacterium* genus-like group (Figure S5.1), which could imply that FOS degradation is a more general characteristic of *Bifidobacterium* members compared to the ability to degrade other (fructo-) oligosaccharides like inulin. This is in agreement with *in vitro* studies that reported that all 55 *Bifidobacterium* strains tested could degrade FOS, whereas only eight could degrade inulin [55].

FOS consumption has previously been reported to affect physicochemical properties of the GI tract [17]. FOS degradation by *Bifidobacterium* takes place by the characteristic fructose-6-phosphate shunt (or bifidus pathway) that produces acetate and lactate as the major end-metabolites [56]. Previous measurements in the cohort used here indeed established increased fecal lactate concentrations upon FOS consumption [17]. Although lactate is poorly absorbed by the epithelial cells in the colon [57], it is normally either absent or found at very low levels (<5mM) in feces from healthy subjects [58-60]. This is due to the fact that lactate is considered as a fermentation intermediate that serves as a substrate for further metabolism to SCFAs by various intestinal bacteria [58, 61]. For example, lactate can be used as substrate for the production of propionate and acetate by members of the genera *Veillonella* [62] and *Propionibacterium* [63], or for the production of butyrate by some of the *Clostridium* cluster XIVa members, such as *Anaerostipes* spp or *Eubacterium hallii*. Notably, the latter cross-feeding interaction could also be verified in *in vitro* co-cultures of *Bifidobacterium* strains and fecal slurries using FOS (or inulin) as substrate [55]. This notion raises the question whether SCFA fluxes within the microbiota were altered or not upon FOS consumption. In this cohort the actual lactate levels after FOS consumption were significantly higher compared to the control diet, but remained far below the lactate concentrations typically seen in patients with functional bowel disorders [60, 64, 65], i.e. on average lower than 2mM [17]. This proposed cross-feeding is in agreement with the observed co-occurrence network modulations elicited by FOS, involving several butyrate producers and implying that butyrate production is enhanced upon FOS consumption (Figure 5.5). Remarkably, the butyryl-CoA transferase gene abundance is reduced in the FOS diet induced microbiota community as compared to the control diet (Figure 5.6B). However, although quantification of this gene has previously been reported to provide gene-specific approach to estimate the butyrate-

production capacity of bacteria in the colon [32], it targets only one of the several pathways for the production of this metabolite. The alternative butyrate production pathway is independent of the butyryl-CoA:acetate CoA transferase and involves phosphotransbutyrylase and butyrate kinase, and for example is found in *Clostridium acetobutylicum* [66]. Moreover, a recent study detected the lysine pathway in the human gut [67], which employs a different terminal enzyme, i.e. butyryl-CoA:acetoacetate CoA transferase, providing another butyrate producing pathway. Consequently, quantification of the butyryl-CoA:acetate CoA transferase encoding gene may underestimate the overall butyrate producing capacity within the community.

Most of the butyrate production capacity in the colon (more than 90%) is commonly assigned to the bacterial groups related to *Eubacterium rectale*, *Roseburia intestinalis* and *Faecalibacterium prausnitzii* of the Firmicutes phylum, none of which have been reported to have the ability to utilize lactate as a substrate [32, 68]. These non-lactate utilizing butyrate producers displayed a decrease upon FOS consumption, with the strongest reduction observed for the abundance of bacteria belonging to *Roseburia intestinalis* ($p=0.025$; Table S5.2). Conversely, the bacterial groups known to be able to convert lactate into butyrate, such as bacteria related to *Anaerostipes caccae* [58], *Eubacterium hallii* [58], and *Megasphaera elsdenii* [61], displayed an increase upon FOS consumption, with the strongest increase observed for the abundance of those related to *Anaerostipes caccae* ($p=0.040$; Table S5.2). Moreover, when analyzing the cumulative relative abundance of lactate-utilizing butyrate producers a significant increase of the group of bacteria was observed ($p=0.048$). Taken together these results support the modulation of butyrate metabolism by FOS consumption, in particular the abundance of community members that can convert lactate to butyrate appears to increase, possibly at the expense of members that produce butyrate from other substrates. Previous measurements on the samples of this cohort did not include butyrate or other any short-chain fatty acids (SCFA) content but did include lactate, which was increased upon FOS consumption [17].

The restructuring of the co-abundance networks upon FOS consumption is not restricted to the microbial groups that are known to produce butyrate (Figure 5.5). The Bacteroidetes members of the community display a higher degree of connectivity to other *Clostridium* cluster IV and XIVa members during the FOS intervention. The connectivity changes of the Bacteroidetes members upon FOS consumption could be the result of their general decrease in abundance. One of the factors that may suppress the Bacteroidetes members is the proposed reduced pH [17], and/or the increased removal of these bacteria by the elevated level of mucin production [17]. The *Clostridium* cluster IV members appear to have decreased connectivity within the co-abundance network during the FOS consumption period compared to the control period, whereas the abundance of the *Clostridium* cluster IV members was not affected by the dietary treatment. Notably, the *Clostridium* cluster IV connection network in the placebo samples resembles a network module in a healthy twin cohort that we studied previously [37] (**Chapter 3**). This module was hypothesized to be composed of primary fiber degraders, of which the fermentation products could be utilized by the so-called secondary fermenters that include several butyrate producers. When we interpret the higher connectivity as an indication of enhanced integration in the cooperative metabolism of the bacteria included in these ecosystem connections, the

hypothesized shift towards lactate driven butyrate production could explain the loss of ecosystem connectivity of the *Clostridium* cluster IV members. The metabolic capacity of these bacterial groups is commonly associated with primary fiber degradation, which may be relatively less demanded within the overall ecosystem when an alternative substrate for butyrate production (lactate) is becoming available at higher levels, e.g. upon FOS consumption.

Overall, our results indicate that FOS induces changes within the microbiota community by increasing the members *Bifidobacterium* genus universally, and decreasing the members belonging to the Bacteroidetes phylum. In addition, FOS consumption elicited a restructuring of co-abundance based connectivity within the ecosystem, which might reflect changes of the metabolic interplay of bacterial groups within the ecosystem's cooperative metabolic network, leading to an enhanced activity of the lactate to butyrate conversion pathway.

Supplementary information

Supplementary material and methods

Real-time quantitative PCR: Butyryl-CoA:acetate CoA transferase gene

Assay performed as described previously with minor modifications. Primers: BCoATscrF (5'-GCI GAI CAT TTC ACI TGG AAY WSI TGG CAY ATG-3') and BCoATscrR (5'-CCT GCC TTT GCA ATR TCI ACR AAN GC-3'). Amplification program included an initial denaturation step at 95°C for 3 min; followed by 40 cycles of denaturation at 95°C for 30s, primer annealing at 51°C for 30s, and primer extension at 72°C for 30s with data acquisition at 72°C; 1 final extension step at 72°C for 5 min. The butyryl-CoA CoA transferase template from *Faecalibacterium prausnitzii* strain M21/2 was used to generate standard curves.

Real-time quantitative PCR: *Bacteroides-Prevotella-Porphyromonas* ***group***

Assay performed as described previously. Primers: BPP-F (5'-GGT GTC GGC TTA AGT GCC AT-3') and BPP-R (5'-CGG A(C/T)G TAA GGG CCG TGC-3'). Amplification program included an initial denaturation step at 95°C for 5 min; followed by 30 cycles of denaturation at 95°C for 15s, primer annealing at 68°C for 20s and primer extension at 72°C for 45s with data acquisition at 72°C; 1 final extension step at 72°C for 5 min. The 16S rRNA gene from *Bacteroides fragilis* strain ATCC 25285 was used to generate standard curves.

Supplemental figures and tables

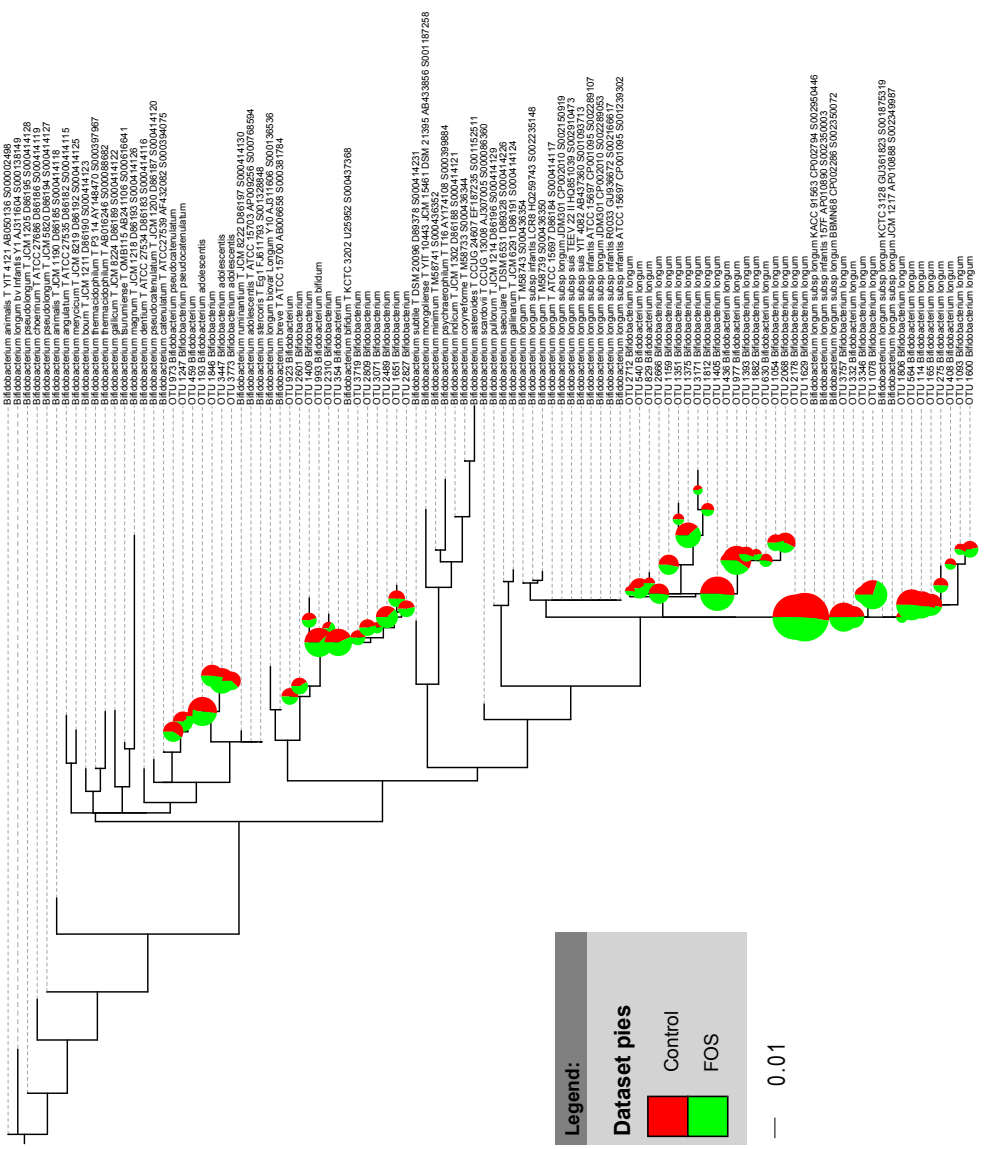


Figure S5.1 Maximum likelihood phylogenetic trees containing bifidobacterial reference sequences from the RDP database and a representative sequence from each OTU detected by the Bifidobacterium specific pyrosequencing. Pie-charts indicate the proportion detected in the control supplementation period (red) and the FOS supplementaioth period (green). Note that the relation between overall Pie-chart size and total abundance is non-linear.

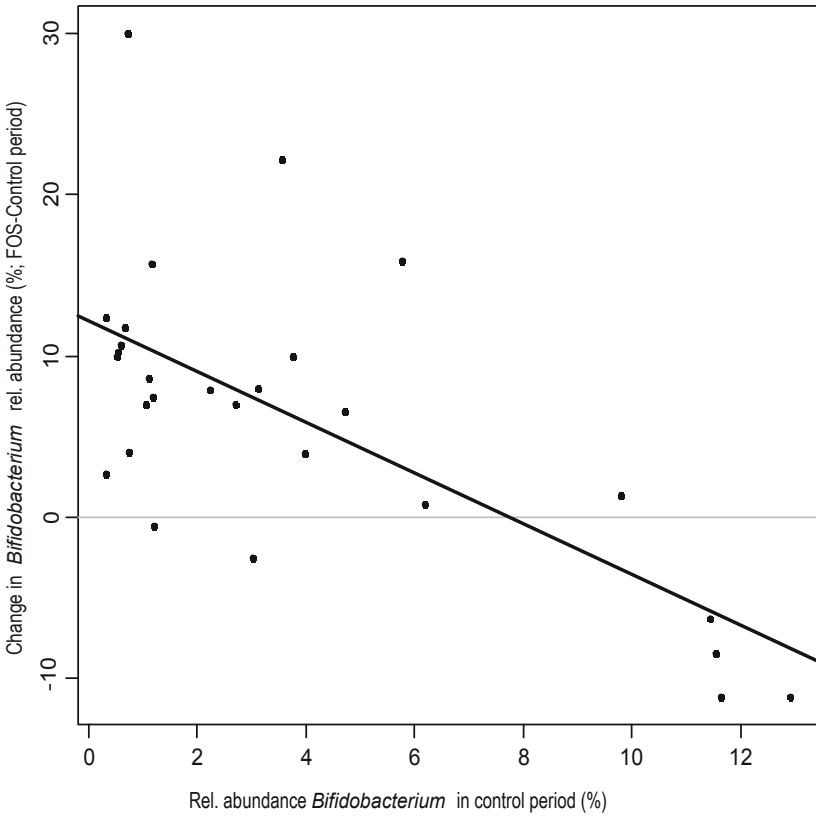


Figure S5.2 Correlation of the control period relative abundance of Bifidobacterium (%) in the control period samples and the difference in relative abundance of Bifidobacterium (%) between samples from the FOS and control period (ρ = -0.57, p = 7.33e-5).

Table S5.1 Mitochondrial DNA sequencing of six subjects.

Subject ID	Diet period	HV1 mtDNA variations (starting from 16090)	HV2 mtDNA variations (starting from 75)
2	I	16094C; 16130A; 16224T; 16264C; 16275A; 16312C; 16344G	77 -; 152C; 263G; 316C
	II	16094C; 16130A; 16224T; 16264C; 16275A; 16312C; 16344G	77 -; 152C; 263G; 316C
3	I	16105T; 16127C; 16295T; 16305C	106A; 152C; 194 ; 263G; 310CTC*
	II	16105T; 16127C; 16295T; 16305C	106 ; 152C; 194A; 263G; 310CTC*
5	I	None	196C; 263G; 310CTC
	II	None	196C; 263G; 310CTC
8	I	16097G; 16362C	239C; 263G; 310CTC
	II	16097G; 16362C	239C; 263G; 310CTC
11	I	16298C	263G; 310CTC
	II	16298C	263G; 310CTC
22	I	16264C; 16299C	263G; 310CTC
	II	16264C; 16299C	263G; 310CTC

*) Low sequencing quality

Table S5.2 Differences in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of the FOS and placebo consumption periods. Significant differences between control and FOS samples ($p < 0.05$; $q < 0.05$) are indicated in bold. Genus-like groups with significant differences between control and FOS samples, but with low relative abundance (< 0.05%) are considered as background variation and marked by a dagger (†). Genus-like groups (level 2) which harbor isolates with known butyrate producers are marked by an asterisk (*).

		Control group relative abundances		FOS group relative abundances		p-	q-
Phylum	**Genus-like group**	**Mean %**	**(± Std.err)**	**Mean %**	**(± Std.err)**	**value**	**value**
Actinobacteria	*Actinomycetaceae*	0.004	0.000	0.005	0.000	0.138	0.139
	Atopobium	0.017	0.001	0.020	0.001	0.053	0.114
	Bifidobacterium	3.810	0.757	10.028	1.360	**0.002**	**0.011**
	Collinsella	0.449	0.062	0.481	0.056	0.186	0.171
	Corynebacterium	0.013	0.001	0.016	0.001	0.050	0.114
	Eggerthella lenta et rel.	0.073	0.008	0.080	0.008	0.236	0.197
	Micrococcaceae	0.009	0.001	0.011	0.001	0.202	0.183
	Propionibacterium	0.026	0.002	0.039	0.012	0.695	0.373
Asteroleplasma	*Asteroleplasma* et rel.	0.010	0.001	0.011	0.001	0.109	0.130
Bacilli	*Aerococcus*	0.011	0.001	0.013	0.002	0.552	0.348
	Aneurinibacillus	0.016	0.001	0.018	0.001	0.157	0.151
	Bacillus	0.026	0.002	0.035	0.006	0.210	0.185
	Enterococcus	0.053	0.006	0.057	0.007	0.711	0.379
	Gemella	0.013	0.001	0.015	0.001	0.126	0.134
	Granulicatella	0.009	0.001	0.010	0.002	0.831	0.426
	Lactobacillus gasseri et rel.	0.459	0.240	0.362	0.141	0.036	0.096
	Lactobacillus plantarum et rel.	0.172	0.040	0.398	0.202	0.399	0.297
	Lactobacillus salivarius et rel.	0.037	0.005	0.271	0.215	0.274	0.220
	Lactococcus	0.023	0.002	0.027	0.002	0.144	0.143
	Staphylococcus	0.019	0.002	0.022	0.003	0.438	0.318
	Streptococcus bovis et rel.	0.737	0.128	0.613	0.063	0.646	0.365
	Streptococcus intermedius et rel.	0.125	0.020	0.197	0.072	0.374	0.282
	Streptococcus mitis et rel.	0.351	0.066	0.457	0.152	0.552	0.348
	Weissella et rel.	0.018	0.002	0.021	0.002	0.479	0.329
	Wissella et rel.	0.003	0.000	0.003	0.000	0.138	0.139
Bacteroidetes	***Allistipes et rel.***	1.748	(±0.429)	0.919	(±0.164)	**<0.001**	**0.002**
	Bacteroides fragilis et rel.	0.710	(±0.171)	0.342	(±0.062)	**0.001**	**0.010**
	Bacteroides intestinalis et rel.	0.357	(±0.066)	0.207	(±0.052)	**0.007**	**0.029**
	Bacteroides ovatus et rel.	0.868	(±0.168)	0.497	(±0.100)	**0.001**	**0.010**
	Bacteroides plebeius et rel.	1.239	(±0.241)	0.661	(±0.151)	**<0.001**	**0.003**
	Bacteroides splachnicus et rel.	1.102	(±0.239)	0.591	(±0.101)	**<0.001**	**0.002**
	Bacteroides stercoris et rel.	0.689	(±0.152)	0.356	(±0.075)	**0.002**	**0.011**

Table S5.2 (continued) Differences in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of the FOS and placebo consumption periods. Significant differences between control and FOS samples ($p < 0.05$; $q < 0.05$) are indicated in bold. Genus-like groups with significant differences between control and FOS samples, but with low relative abundance (< 0.05%) are considered as background variation and marked by a dagger (†). Genus-like groups (level 2) which harbor isolates with known butyrate producers are marked by an asterisk (*).

		Control group relative abundances		FOS group relative abundances		p-	q-
Phylum	**Genus-like group**	**Mean %**	**(± Std.err)**	**Mean %**	**(± Std.err)**	**value**	**value**
Bacteroidetes	*Bacteroides uniformis* et rel.	0.722	(±0.140)	0.393	(±0.083)	**0.016**	0.051
	Bacteroides vulgatus et rel.	1.260	(±0.262)	0.573	(±0.139)	**<0.001**	**<0.001**
	Parabacteroides distasonis et rel.	1.453	(±0.226)	0.772	(±0.129)	**<0.001**	**0.004**
	Prevotella melaninogenica et rel.	4.628	(±1.020)	4.324	(±1.076)	0.537	0.348
	Prevotella oralis et rel.	0.455	(±0.081)	0.313	(±0.060)	**0.007**	**0.029**
	Prevotella ruminicola et rel.	0.350	(±0.057)	0.235	(±0.049)	**0.007**	**0.029**
	Prevotella tannerae et rel.	0.905	(±0.130)	0.632	(±0.136)	**0.009**	**0.034**
	Tannerella et rel.	0.714	(±0.154)	0.436	(±0.073)	**<0.001**	**0.002**
	Uncultured Bacteroidetes	0.118	(±0.050)	0.053	(±0.019)	**0.002**	**0.013**
Clostridium cluster I	*Clostridia* *	6.511	2.062	7.842	2.136	0.582	0.353
Clostridium cluster III	*Clostridium stercorarium* et rel.	0.393	0.073	0.500	0.111	0.508	0.337
	Clostridium thermocellum et rel.	0.010	0.001	0.011	0.001	0.095	0.130
Clostridium cluster IV	*Anaerotruncus colihominis* et rel. *	0.273	0.041	0.281	0.062	0.662	0.365
	Clostridium cellulosi et rel.	0.512	0.072	0.632	0.124	0.678	0.367
	Clostridium leptum et rel. *	0.272	0.051	0.331	0.081	0.814	0.423
	Clostridium orbiscindens et rel.	0.825	0.095	0.691	0.079	0.099	0.130
	Eubacterium siraeum et rel.	0.354	0.097	0.252	0.058	0.849	0.426
	Faecalibacterium prausnitzii et rel. *	13.005	0.845	11.975	1.371	0.646	0.365
	Oscillospira guillermondii et rel.	1.173	0.229	0.851	0.147	0.115	0.130
	Papillibacter cinnamivorans et rel. *	0.527	0.043	0.405	0.043	**0.023**	0.068
	Ruminococcus bromii et rel.	0.414	0.096	0.588	0.245	0.552	0.348
	Ruminococcus callidus et rel.	1.018	0.161	1.205	0.441	0.063	0.121
	Sporobacter termitidis et rel.	0.493	0.105	0.541	0.123	1.000	0.472
	Subdoligranulum variable et rel. *	2.008	0.222	2.220	0.307	1.000	0.472
Clostridium cluster XI	*Dialister*	0.196	0.096	0.088	0.026	0.060	0.119
	Megamonas hypermegale et rel.	0.032	0.012	0.027	0.008	0.218	0.187
	Megasphaera elsdenii et rel. *	0.095	0.047	0.311	0.180	0.350	0.270
	Mitsuokella multiacida et rel.	0.341	0.182	0.607	0.298	0.425	0.312

Table S5.2 (continued) Differences in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of the FOS and placebo consumption periods. Significant differences between control and FOS samples ($p < 0.05$; $q < 0.05$) are indicated in bold. Genus-like groups with significant differences between control and FOS samples, but with low relative abundance (< 0.05%) are considered as background variation and marked by a dagger (†). Genus-like groups (level 2) which harbor isolates with known butyrate producers are marked by an asterisk (*).

		Control group relative abundances		FOS group relative abundances		p-	q-
Phylum	**Genus-like group**	**Mean %**	**(± Std.err)**	**Mean %**	**(± Std.err)**	**value**	**value**
Clostridium cluster XI	*Peptococcus niger* et rel.	0.067	0.006	0.060	0.004	0.316	0.250
	Phascolarctobacterium faecium et rel.	0.081	0.014	0.097	0.021	0.264	0.215
	Uncultured *Selenomonadaceae*	0.001	0.001	0.001	0.001	0.138	0.139
	Veillonella	0.043	0.024	0.019	0.003	0.374	0.282
	Anaerovorax odorimutans et rel. *	0.096	0.011	0.089	0.007	0.728	0.384
	Clostridium difficile et rel. *	14.699	1.946	13.274	2.354	0.567	0.350
	Clostridium felsineum et rel.	0.486	0.178	0.601	0.238	0.937	0.460
	Peptostreptococcus anaerobius et rel.	0.026	0.003	0.031	0.007	0.567	0.350
Clostridium cluster XIII	*Peptostreptococcus micros et rel.*	0.037	0.003	0.044	0.004	0.099	0.130
Clostridium cluster XIVa	*Anaerostipes caccae* et rel. *	1.407	0.124	2.079	0.270	0.040	0.104
	Bryantella formatexigens et rel. *	1.329	0.107	1.527	0.230	0.902	0.446
	Butyrivibrio crossotus et rel. *	1.190	0.201	1.169	0.215	0.646	0.365
	Clostridium colinum et rel.	0.381	0.045	0.358	0.049	0.646	0.365
	Clostridium nexile et rel.	1.328	0.123	1.474	0.239	0.991	0.472
	Clostridium sphenoides et rel.	1.340	0.137	1.488	0.194	0.451	0.321
	Clostridium symbiosum et rel. *	1.989	0.108	2.017	0.303	0.164	0.156
	Coprococcus eutactus et rel. *	1.282	0.149	0.982	0.095	0.081	0.130
	Dorea formicigenerans et rel.	2.242	0.210	2.199	0.242	0.662	0.365
	Eubacterium hallii et rel. *	0.936	0.087	1.028	0.153	0.973	0.472
	Eubacterium rectale et rel. *	2.407	0.210	2.338	0.255	0.849	0.426
	Eubacterium ventriosum et rel. *	1.009	0.115	0.992	0.181	0.508	0.337
	Lachnobacterium bovis et rel.	1.297	0.223	1.094	0.185	0.339	0.265
	Lachnospira pectinoschiza et rel. *	2.644	0.273	2.382	0.223	0.646	0.365
	Outgrouping clostridium cluster XIVa	1.190	0.111	1.202	0.102	0.991	0.472
	Roseburia intestinalis et rel. *	1.900	0.195	1.298	0.185	0.025	0.069
	Ruminococcus gnavus et rel.	0.516	0.068	0.497	0.056	0.646	0.365
	Ruminococcus lactaris et rel.	0.247	0.028	0.210	0.025	0.074	0.130
	Ruminococcus obeum et rel.	3.848	0.295	3.756	0.414	0.662	0.365

Table S5.2 (continued) Differences in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of the FOS and placebo consumption periods. Significant differences between control and FOS samples ($p < 0.05$; $q < 0.05$) are indicated in bold. Genus-like groups with significant differences between control and FOS samples, but with low relative abundance (< 0.05%) are considered as background variation and marked by a dagger (†). Genus-like groups (level 2) which harbor isolates with known butyrate producers are marked by an asterisk (*).

		Control group relative abundances		FOS group relative abundances			
Phylum	**Genus-like group**	**Mean %**	**(± Std.err)**	**Mean %**	**(± Std.err)**	**p-value**	**q-value**
Clostridium cluster XV	*Anaerofustis*	0.014	0.001	0.016	0.001	0.095	0.130
	Eubacterium limosum et rel. *	0.035	0.003	0.045	0.006	0.053	0.114
Clostridium cluster XVI	*Bulleidia moorei* et rel.	0.037	0.003	0.040	0.003	0.210	0.185
	Eubacterium biforme et rel. *	0.082	0.017	0.079	0.014	0.902	0.446
	Eubacterium cylindroides et rel. *	0.028	0.002	0.034	0.003	0.109	0.130
Clostridium cluster XVII	*Catenibacterium mitsuokai* et rel. †	0.013	0.002	0.021	0.006	**0.010**	**0.035**
	Lactobacillus catenaformis et rel.	0.002	0.000	0.002	0.000	0.050	0.114
Clostridium cluster XVIII	*Clostridium ramosum* et rel.	0.042	0.003	0.044	0.004	0.552	0.348
	Coprobacillus catenaformis et rel.	0.087	0.009	0.078	0.016	**0.017**	0.052
Cyanobacteria	Uncultured *Chroococcales*	0.009	0.001	0.010	0.001	0.465	0.323
Fusobacteria	*Fusobacteria*	0.028	0.002	0.034	0.003	0.086	0.130
Spirochaetes	*Brachyspira*	0.014	0.001	0.016	0.001	0.095	0.130
Uncultured *Clostridiales*	Uncultured *Clostridiales* I	0.913	0.277	0.820	0.198	0.646	0.365
	Uncultured *Clostridiales* II	0.438	0.067	0.476	0.095	0.849	0.426
Uncultured *Mollicutes*	Uncultured *Mollicutes*	0.190	0.019	0.160	0.017	0.218	0.187
Verrucomicrobia	*Akkermansia*	0.063	0.023	0.062	0.019	0.245	0.202

Table S5.2 (continued) Differences in fecal microbiota composition of 28 healthy subjects determined with the Human Intestinal Tract (HIT)Chip. Samples were taken at the end of the FOS and placebo consumption periods. Significant differences between control and FOS samples ($p < 0.05$; $q < 0.05$) are indicated in bold. Genus-like groups with significant differences between control and FOS samples, but with low relative abundance (< 0.05%) are considered as background variation and marked by a dagger (†). Genus-like groups (level 2) which harbor isolates with known butyrate producers are marked by an asterisk (*).

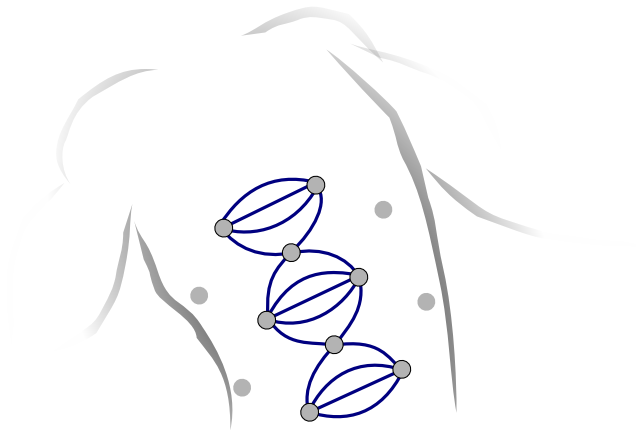
		Control group relative abundances		FOS group relative abundances		p-	q-
Phylum	**Genus-like group**	**Mean %**	**(± Std.err)**	**Mean %**	**(± Std.err)**	**value**	**value**
	Aeromonas	0.009	0.001	0.011	0.001	0.109	0.130
	Alcaligenes faecalis et rel.	0.012	0.002	0.013	0.002	0.120	0.130
	Anaerobiospirillum	0.008	0.001	0.009	0.001	0.120	0.130
	Aquabacterium †	0.009	0.003	0.012	0.003	<0.001	0.004
	Bilophila et rel.	0.017	0.001	0.020	0.002	0.120	0.130
	Burkholderia	0.013	0.003	0.014	0.003	0.779	0.408
	Campylobacter	0.031	0.003	0.038	0.003	0.109	0.130
	Desulfovibrio et rel.	0.027	0.002	0.033	0.003	0.104	0.130
	Enterobacter aerogenes et rel. †	0.024	0.002	0.032	0.003	0.010	0.035
	Escherichia coli et rel.	0.047	0.004	0.059	0.006	0.115	0.130
	Haemophilus	0.006	0.001	0.010	0.002	0.115	0.130
	Helicobacter	0.021	0.002	0.025	0.002	0.095	0.130
Proteobacteria	*Klebisiella pneumoniae* et rel.	0.023	0.002	0.032	0.005	0.050	0.114
	Leminorella	0.005	0.000	0.006	0.001	0.150	0.147
	Methylobacterium	0.003	0.000	0.004	0.000	0.115	0.130
	Moraxellaceae	0.013	0.001	0.015	0.001	0.109	0.130
	Novosphingobium	0.007	0.001	0.008	0.001	0.582	0.353
	Oceanospirillum	0.042	0.005	0.043	0.003	0.493	0.335
	Oxalobacter formigenes et rel.	0.063	0.009	0.062	0.016	0.451	0.321
	Proteus et rel.	0.020	0.002	0.025	0.002	0.099	0.130
	Pseudomonas	0.008	0.001	0.012	0.003	0.104	0.130
	Serratia	0.001	0.000	0.002	0.001	0.060	0.119
	Sutterella wadsworthia et rel.	0.464	0.064	0.425	0.061	0.678	0.367
	Vibrio	0.024	0.003	0.033	0.005	0.227	0.192
	Xanthomonadaceae	0.083	0.019	0.059	0.014	0.186	0.171
	Yersinia et rel.	0.019	0.002	0.023	0.002	0.086	0.130

References

1. **Clemente JC, Ursell LK, Parfrey LW, et al.** The impact of the gut microbiota on human health: an integrative view. Cell 2012;148(6):1258-70.
2. **Roberfroid M, Gibson GR, Hoyles L, et al.** Prebiotic effects: metabolic and health benefits. Br J Nutr 2010;104 Suppl 2:S1-63.
3. **Gibson GR, Scott KP, Rastall RA, et al.** Dietary prebiotics: current status and new definition. Food Science & Technology Bulletin: Functional Foods 2010;7(1):1-19.
4. **Gibson GR, Probert HM, Loo JV, et al.** Dietary modulation of the human colonic microbiota: updating the concept of prebiotics. Nutr Res Rev 2004;17(2):259-75.
5. **Alles MS, Hautvast JG, Nagengast FM, et al.** Fate of fructo-oligosaccharides in the human intestine. Br J Nutr 1996;76(2):211-21.
6. **Salvini F, Riva E, Salvatici E, et al.** A specific prebiotic mixture added to starting infant formula has long-lasting bifidogenic effects. J Nutr 2011;141(7):1335-9.
7. **Benjamin JL, Hedin CR, Koutsoumpas A, et al.** Randomised, double-blind, placebo-controlled trial of fructo-oligosaccharides in active Crohn's disease. Gut 2011;60(7):923-9.
8. **Yen CH, Kuo YW, Tseng YH, et al.** Beneficial effects of fructo-oligosaccharides supplementation on fecal bifidobacteria and index of peroxidation status in constipated nursing-home residents--a placebo-controlled, diet-controlled trial. Nutrition 2011;27(3):323-8.
9. **Wierdsma NJ, van Bodegraven AA, Uitdehaag BM, et al.** Fructo-oligosaccharides and fibre in enteral nutrition has a beneficial influence on microbiota and gastrointestinal quality of life. Scand J Gastroenterol 2009;44(7):804-12.
10. **Rajilic-Stojanovic M and de Vos WM.** The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 2014;38(5):996-1047.
11. **Bovee-Oudenhoven IM, ten Bruggencate SJ, Lettink-Wissink ML, et al.** Dietary fructo-oligosaccharides and lactulose inhibit intestinal colonisation but stimulate translocation of salmonella in rats. Gut 2003;52(11):1572-8.
12. **Haarman M and Knol J.** Quantitative real-time PCR assays to identify and quantify fecal Bifidobacterium species in infants receiving a prebiotic infant formula. Appl Environ Microbiol 2005;71(5):2318-24.
13. **Lindsay JO, Whelan K, Stagg AJ, et al.** Clinical, microbiological, and immunological effects of fructo-oligosaccharide in patients with Crohn's disease. Gut 2006;55(3):348-55.
14. **Dewulf EM, Cani PD, Claus SP, et al.** Insight into the prebiotic concept: lessons from an exploratory, double blind intervention study with inulin-type fructans in obese women. Gut 2013;62(8):1112-21.
15. **Cecchini DA, Laville E, Laguerre S, et al.** Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. PLoS One 2013;8(9):e72766.
16. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
17. **Ten Bruggencate SJ, Bovee-Oudenhoven IM, Lettink-Wissink ML, et al.** Dietary fructooligosaccharides affect intestinal barrier function in healthy men. J Nutr 2006;136(1):70-4.
18. **Salonen A, Nikkila J, Jalanka-Tuovinen J, et al.** Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods 2010;81(2):127-34.
19. **Zgonjanin D, Veselinovic I, Kubat M, et al.** Sequence polymorphism of the mitochondrial DNA control region in the population of Vojvodina Province, Serbia. Leg Med (Tokyo) 2010;12(2):104-7.
20. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009.
21. **Storey JD and Stanford University USA.** A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64:479-498.
22. **Butts CT, Handcock MS, and Hunter DR.** network: Classes for Relational Data. in R package version 1.7. 2011: Irvine, CA.
23. **Bastian M, Heymann S, and Jacomy M.** Gephi: An Open Source Software for Exploring and Manipulating Networks. in International AAAI Conference on Weblogs and Social Media. 2009: San Jose, California.
24. **Satokari RM, Vaughan EE, Akkermans ADL, et al.** Bifidobacterial Diversity in Human Feces Detected by Genus-Specific PCR and Denaturing Gradient Gel Electrophoresis. Applied and Environmental Microbiology 2001;67(2):504-513.
25. **Zeeuwen PLJM, Boekhorst J, van den Bogaard EH, et al.** Microbiome dynamics of human epidermis following skin barrier disruption. Genome Biology 2012;13(11):R101-R101.

26. **Caporaso JG, Kuczynski J, Stombaugh J, et al.** QIIME allows analysis of high-throughput community sequencing data. Nature methods 2010;7(5):335-336.
27. **Haas BJ, Gevers D, Earl AM, et al.** Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 2011;21(3):494-504.
28. **Caporaso G.** QIIME newsletter of 17 December 2010 QIIME, Quantitative Insights into Microbial Ecology. 2010 [cited 2014 24-Nov]; Available from: http://qiime.wordpress.com/2010/12/17/new-default-parameters-for-uclust-otu-pickers/.
29. **van den Bogert B, Erkus O, Boekhorst J, et al.** Diversity of human small intestinal Streptococcus and Veillonella populations. FEMS Microbiology Ecology 2013;85(2):376-388.
30. **SciPy-developers.** SciPy. 2014 [cited 2014 24-Nov]; Available from: www.scipy.org/.
31. **ter Braak C and Šmilauer P.** Canoco Reference Manual and User's Guide: Software for Ordination (version 5.0). 2012: Microcomputer power, Itaca, www. canoco. com.
32. **Louis P and Flint HJ.** Development of a semiquantitative degenerate real-time pcr-based assay for estimation of numbers of butyryl-coenzyme A (CoA) CoA transferase genes in complex bacterial samples. Appl Environ Microbiol 2007;73(6):2009-12.
33. **Rinttila T, Kassinen A, Malinen E, et al.** Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. J Appl Microbiol 2004;97(6):1166-77.
34. **Zoetendal EG, Akkermans ADL, Akkermans-van Vliet WM, et al.** The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. Microbial Ecology in Health and Disease 2001;13(3):129-134.
35. **Stewart JA, Chadwick VS, and Murray A.** Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. J Med Microbiol 2005;54(12):1239-1242.
36. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-484.
37. **Tims S, Derom C, Jonkers DM, et al.** Microbiota conservation and BMI signatures in adult monozygotic twins. ISME J 2013;7(4):707-17.
38. **Eckburg PB, Bik EM, Bernstein CN, et al.** Diversity of the human intestinal microbial flora. Science 2005;308(5728):1635-8.
39. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-4.
40. **Qin J, Li R, Raes J, et al.** A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59-65.
41. **Jalanka-Tuovinen J, Salonen A, Nikkila J, et al.** Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One 2011;6(7):e23035.
42. **Rajilic-Stojanovic M, Heilig H, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
43. **Rajilic-Stojanovic M, Heilig HG, Tims S, et al.** Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 2012.
44. **Zoetendal EG, Cheng B, Koike S, et al.** Molecular microbial ecology of the gastrointestinal tract: from phylogeny to function. Curr Issues Intest Microbiol 2004;5(2):31-47.
45. **Kolida S and Gibson GR.** Prebiotic capacity of inulin-type fructans. J Nutr 2007;137(11 Suppl):2503S-2506S.
46. **Claesson MJ, O'Sullivan O, Wang Q, et al.** Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One 2009;4(8):e6669.
47. **van den Bogert B, de Vos WM, Zoetendal EG, et al.** Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. Appl Environ Microbiol 2011;77(6):2071-80.
48. **Arumugam M, Raes J, Pelletier E, et al.** Enterotypes of the human gut microbiome. Nature 2011;473(7346):174-80.
49. **Whelan K, Judd PA, Preedy VR, et al.** Fructooligosaccharides and fiber partially prevent the alterations in fecal microbiota and short-chain fatty acid concentrations caused by standard enteral formula in healthy humans. J Nutr 2005;135(8):1896-902.
50. **Duncan SH, Louis P, Thomson JM, et al.** The role of pH in determining the species composition of the human colonic microbiota. Environmental Microbiology 2009;11(8):2112-2122.
51. **Davis LM, Martinez I, Walter J, et al.** Barcoded pyrosequencing reveals that consumption of galactooligosaccharides results in a highly specific bifidogenic response in humans. PLoS One 2011;6(9):e25200.

52. **Ramirez-Farias C, Slezak K, Fuller Z, et al.** Effect of inulin on the human gut microbiota: stimulation of Bifidobacterium adolescentis and Faecalibacterium prausnitzii. Br J Nutr 2009;101(4):541-50.
53. **Costabile A, Kolida S, Klinder A, et al.** A double-blind, placebo-controlled, cross-over study to establish the bifidogenic effect of a very-long-chain inulin extracted from globe artichoke (Cynara scolymus) in healthy human subjects. Br J Nutr 2010;104(7):1007-17.
54. **Kleessen B, Schwarz S, Boehm A, et al.** Jerusalem artichoke and chicory inulin in bakery products affect faecal microbiota of healthy volunteers. Br J Nutr 2007;98(3):540-9.
55. **Rossi M, Corradini C, Amaretti A, et al.** Fermentation of fructooligosaccharides and inulin by bifidobacteria: a comparative study of pure and fecal cultures. Appl Environ Microbiol 2005;71(10):6150-8.
56. **de Vries W and Stouthamer AH.** Pathway of glucose fermentation in relation to the taxonomy of bifidobacteria. J Bacteriol 1967;93(2):574-6.
57. **Umesaki Y, Yajima T, Yokokura T, et al.** Effect of organic acid absorption on bicarbonate transport in rat colon. Pflügers Archiv 1979;379(1):43-47.
58. **Duncan SH, Louis P, and Flint HJ.** Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Appl. Environ. Microbiol. 2004;70(10):5810-5817.
59. **Macfarlane S and Macfarlane GT.** Regulation of short-chain fatty acid production. Proc Nutr Soc 2003;62(1):67-72.
60. **Bustos D, Pons S, Pernas JC, et al.** Fecal lactate and short bowel syndrome. Dig Dis Sci 1994;39(11):2315-9.
61. **Hashizume K, Tsukahara T, Yamada K, et al.** Megasphaera elsdenii JCM1772T normalizes hyperlactate production in the large intestine of fructooligosaccharide-fed rats by stimulating butyrate production. J Nutr 2003;133(10):3187-90.
62. **Zoetendal EG, Raes J, van den Bogert B, et al.** The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. ISME J 2012;6(7):1415-26.
63. **Crow VL.** Utilization of Lactate Isomers by Propionibacterium freudenreichii subsp. shermanii: Regulatory Role for Intracellular Pyruvate. Appl Environ Microbiol 1986;52(2):352-8.
64. **Vernia P, Caprilli R, Latella G, et al.** Fecal lactate and ulcerative colitis. Gastroenterology 1988;95(6):1564-8.
65. **Vernia P, Gnaedinger A, Hauck W, et al.** Organic anions and the diarrhea of inflammatory bowel disease. Digestive Diseases and Sciences 1988;33(11):1353-1358.
66. **Walter KA, Nair RV, Cary JW, et al.** Sequence and arrangement of two genes of the butyrate-synthesis pathway of Clostridium acetobutylicum ATCC 824. Gene 1993;134(1):107-11.
67. **Vital M, Howe AC, and Tiedje JM.** Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta) genomic Data. mBio 2014;5(2):e00889-14.
68. **Duncan SH, Louis P, and Flint HJ.** Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Appl Environ Microbiol 2004;70(10):5810-7.



Chapter 6

Relating 16S rRNA gene profiling and shotgun metagenome data to predict functional capacities of currently uncultured bacteria from the human gut

Sebastian Tims, Erwin G. Zoetendal, Willem M. de Vos, Michiel Kleerebezem

Manuscript in preparation

Abstract

Improved understanding of the functioning of the gut microbiota in relation to human health is essential for successful new strategies to modulate the microbiota with the ambition to improve human and animal health. Current knowledge is limited in the functional reconstruction of the gut microbial ecosystem due to the fact that the majority of the gut microbiota consists of uncultured microbes. Standard metagenome analysis strategies, which compare the metagenomic content to reference genomes of organisms that have been cultured in the laboratory, therefore, cannot annotate a large portion of the data obtained. Furthermore, uncultured bacteria are frequently found among the microbial groups that have been identified as biomarkers for disease or health states of their host.

Here we show how the integration of two cultivation independent community-profiling approaches, i.e. 16S rRNA gene profiling and untargeted shotgun metagenomics data (after co-abundance gene clustering), can link functional properties to uncultured microbes from which only a 16S rRNA gene sequence is known. This combinatory assessment could provide insight into the functionality of the not yet cultured microbes as well as guide laboratory-media development for future isolation and cultivation attempts. For 13 exemplary uncultured organisms that on basis of their occurrence pattern among studied subjects that are of potential relevance in the context of host-health or –disease, we extracted functional clues which can potentially facilitate the design of advanced isolation strategies or provide genome-based functional information, such as the capacity to synthesize B-vitamins or amino acids, the presence of specific butyrate production pathways, or the manifestation of a homoacetogenic lifestyle.

Introduction

In the last decades the gut microbiome has increasingly been recognized as a functional counterpart of its host genome [1-3]. A great number of associations between the gut microbiota and disease have been described but, besides for specific infectious gut pathogens such as *Listeria monocytogenes* [4], *Vibrio cholerae* [5], or *Salmonella typhimurium* [6, 7], only limited causality has been demonstrated [8]. For a long time, isolation and subsequent cultivation and characterization of microbial strains in the laboratory was the only option to decipher functional properties of gut bacteria, and presently over 1,000 species of human intestine derived bacteria, archaea and fungi have been described [9]. The past decade has brought several new and next generation sequencing technologies that enabled the inventory of the microbial and metagenomic composition in complex ecosystems, such as the intestinal tract. The results of these approaches strengthened the view that the microbiota is a vast functional reservoir with the potential to orchestrate human physiology through various molecular and/or metabolic interactions, which include shaping the immune system, impacting metabolic functions, and protection against pathogens (for a review see [10]).

Improved understanding of the functioning of the gut microbiota in relation to human health and well-being can pave the way for innovative strategies to utilize or manage the microbiota with the intention to modulate and improve human and animal health by targeting homeostasis-maintenance or treating (intestinal) diseases. A healthy human adult gut is colonized by more than 1,000 bacterial species, that predominantly belong to the Actinobacteria, Bacteroidetes and Firmicutes phyla, and of which the majority has not yet been cultured in the laboratory [9]. The majority of studies addressing the gut community structure have focused on identifying its members based on the 16S rRNA gene sequence, which relies on sequence databases of known, though not necessarily cultured, bacteria. This approach does not measure the genetic potential of a given community, which can be assessed by shotgun metagenomic approaches that aim to determine the genetic potential of the microbiota by shotgun sequence analysis of total DNA extracted from the ecosystem. However, metagenomic analysis usually cannot provide high-confidence phylogenetic assignment for all sequenced genes that are detected [11]. This limitation can partly be attributed to the lack of genetic information from uncultured members of the community, since most metagenome analysis strategies depend on the comparison to reference genomes of organisms that have been cultured in the laboratory and these only account for a fraction of all microorganisms in the gut. Hence, this prevents accurate phylogeny-based analysis and subsequent predictions of the microbial ecosystem [12]. The metagenomic approaches are further complicated by the considerable genetic heterogeneity that is frequently observed between genomes of different isolates of the same bacterial species [13, 14]. As a consequence, the use of current reference genomes for metagenome mapping provides only a poor coverage of the actual biological diversity, and greatly limits the functional reconstruction of the gut microbial ecosystem. Not only do uncultured bacteria account for the majority of the gut microbiota, they also are frequently found among the microbial groups that have been reported as biomarkers for various physiological aspects of their host, such as BMI or metabolic health status [15]. Moreover, a recent large scale HITChip meta-analysis [16] of the fecal microbiota composition

of more than 1,000 healthy western subjects identified five bacterial groups that were proposed to act as ecosystem tipping-point species. In the vast majority of the samples, these bacterial groups were either present in relatively high or in very low (nearly absent) abundance levels: *Dialister* spp., relatives of *Bacteroides fragilis*, a *Prevotella* subgroup, and two groups of uncultured *Clostridiales* (UCI and UCII) [16]. The observed bimodal distribution pattern of these microbial groups could be associated to several host characteristics, such as age and health-related aspects, including BMI [16]. Interestingly, even for *Dialister* spp., relatives of *Bacteroides fragilis*, and the *Prevotella* subgroup, the 16S rRNA gene groups with the most prominent bimodal distribution were associated to uncultured isolates [16]. Hence, characterization of multiple representatives of the currently uncultured bacterial genera and species that inhabit the gut is needed to assign comprehensive functional and physiological information to these biomarker microbes.

Here we show how 16S rRNA gene profiling and shotgun-metagenomic data generated for intestinal samples studied within the MetaHIT consortium [17] can be combined to deduce the genetic potential harboured by uncultured bacteria that have so far only be identified based on their 16S rRNA gene sequence. To this end, 16S rRNA gene profiling data was generated for a large number of fecal samples for which the MetaHIT consortium had already generated metagenomic data and had recognized co-abundant gene groups that were labelled as metagenomic-species (MGS) [17]. An alternative approach for metagenomic species profiling using universal phylogenetic marker genes rather than 16S rRNA, has also been reported [18]. From the genetic potential of 13 MGS, with high correlations to 16S rRNA genes from uncultured bacteria that displayed intriguing occurrence patterns among the studied subjects, clues were extracted that could facilitate the design of laboratory media for their cultivation with respect to amino acid, folate (vitamin B9) and cobalamin (vitamin B12) requirements. Additionally, six out of the 13 MGS were found to encode one or more enzymes involved in butyrate production, and two of these MGS appear to harbour the uncommon combination of a butyryl-CoA transferase and a butyrate kinase. Moreover, three out of the 13 MGS appear to be homoacetogens, capable of reductive acetogenesis, which is a relevant way to produce acetate in the gut and may have ecosystem wide consequence for other fermentation processes. Overall, the combinatory approach presented here could guide the development of isolation strategies, target and set-up screening tools for specific uncultured organism on basis of their interesting genetic potential, and provide genome-based functional information that can be linked to health or disease states

Materials and methods

Metagenomic data set

The metagenome data set used here is an extension of the 124 subjects previously sequenced by Qin and co-workers, which originally resulted in a 3.3 million gene catalogue [11], and was generated from a total of 177 Danish and 141 Spanish subjects of which 78 were sampled twice, as described by Nielsen and co-workers in 2014 [17]. The combination of the new Illumnia GA reads with the ones from the 3.3 million gene catalogue resulted in an updated non-redundant 3.9 million gene catalogue [17].

HITChip analysis

The samples used for the construction of the 3.9M gene catalogue were also used for 16S rRNA gene amplification, RNA (reverse-)transcription and Human Intestinal Tract Chip (HITChip) microarray hybridization according to previously described methods [19]. When available, aliquots of the same DNA samples that were used for shotgun metagenomic sequencing were used (n=117), while for the remaining samples DNA was extracted from different aliquots of the same fecal samples (n=279) using the same procedure. The HITChip platform, which is based on a phylogenetic microarray with 3,699 unique probes that target over 1,000 intestinal phylotypes [19], was used for 16S rRNA gene-based microbiota profiling. Specific probe hybridization signals were assigned to different phylogenetic levels: order-like, genus-like (sequence similarity >90%), and species-like (sequence similarity >98%) 16S rRNA gene sequence groups [19].

Co-abundance gene clustering and paralleled co-abundance taxonomic mapping (3.9 M gene catalogue set)

Co-abundant gene groups (CAGs) were previously defined on the 396 samples from the 3.9M gene catalogue set, using a canopy-based clustering approach of the mapping results of all samples using the 3.9M gene catalogue [17]. CAG abundance profiles (median gene depth signal per sample) were correlated to HITChip relative abundance profiles (at all phylogenetic levels and at probe signal level) by calculating Spearman's correlation coefficient (ϱ). The bidirectional and parallel co-abundance taxonomic mapping of HITChip and gene group data resulted in predicted taxonomic assignments of the detected CAGs at the different phylogenetic levels. For consistency with the canopy-based clustering approach, only correlations with $\varrho > 0.6$ were used to assign HITChip taxonomy to the CAGs. Furthermore, CAGs were only assigned to HITChip phylogenetic groups when this assignment was "uniform", meaning that at any given phylogenetic level only a single taxonomic group is allowed with a $\varrho > 0.6$. The CAGs with the highest correlations ($\varrho > 0.8$) to uncultured isolates were selected for further investigation of their functional gene composition.

Functional annotation

The 3.9M gene catalogue was functionally annotated by BLASTP against the UniProt database (http://www.uniprot.org/) and the proteins from the eggNOG (v3) database [20], as described previously [17]. Gene sets and their functional annotation have been deposited at the European Nucleotide Archive (ENA) for the CAGs with >700 genes that passed the six high-quality draft assembly criteria from the Human Microbiome Project (HMP)[21]. Via the UniProt Knowledgebase (UniProtKB), UniprotIDs corresponding to enzymes from butyrate production pathways, as indicated by Vital and co-workers [22], were retrieved to identify potential butyrate producing organisms among CAGs of interest. Similarly, UniprotIDs corresponding to enzymes from the various amino acid biosynthetic pathways were retrieved. Furthermore, CAGs of interest were analysed with the online Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [23, 24], employing the KEGG Automatic Annotation Server (KAAS; [http://www.genome.jp/kegg/kaas/, 25]). KEGG Orthology (KO) assignment was obtained using the

recommended settings for metagenomes, which is the single-directional best hit (SBH) method combined with a representative gene data set for Prokaryotes. Annotations by KAAS yielded KO assignments per gene and KEGG pathway maps, in which these assignments were represented.

Results and discussion

Co-abundance clustering of metagenomic genes

Most metagenome annotation methods rely on the alignment with currently available reference genomes, which are strongly predominated by cultured isolates and represent only a small portion of the diversity within a gut microbiota metagenome. Thereby, our capacity to extract coherent biological information and genomic entities from metagenomic data is quite limited. Ideally, one would strive for *de novo* assembly of all biological entities present in a metagenome dataset. However, sequence ambiguities and limitations in sequencing depth render *de novo* assembly impossible [26]. Consequently, current assembly methods produce large sets of small and apparently independent contigs that do not allow biological entity reconstruction [27]. To overcome the difficulties in metagenomic data analysis a method was developed to segregate biological entities within large amounts of metagenomic data, utilizing co-abundance profiling of genes in multiple samples [17]. A similar method had already been applied before to cluster 120,723 metagenomic genes that were associated to bacterial richness [15], but computational constraints limited the mapping of the entire gene catalogue in a distance matrix that encompasses millions of genes. This computational constraint was overcome by extracting groups of genes based on their abundance correlation relative to a randomly picked set of seed genes [17].

The co-abundance method defines biological entities as co-abundant gene groups (CAGs), allowing a first evaluation of their genetic potential. Out of the 7,381 CAGs identified within the gut metagenome, 741 encompassed more than 700 genes, which was taken as a minimal number of genes to assign the term “metagenomic species” (MGS) to this CAG. A high level of consistency of nucleotide base composition and congruent taxonomic classification of the genes captured within a MGS supported the proposition that these MGS represent biological entities [17]. It should be noted that the MGS gene sets are not equivalent to complete genomes, but most likely are a reflection of the core genome of a particular species that is subject to clonal variations in different representatives of such species. The smaller CAGs may constitute specific sets of gene that are relatively frequently present within a species and as a consequence may display dependency associations with specific MGS [17]. Hence, these could be part of the so-called accessory genome of a species. Correspondingly, several of the smaller CAGs (< 700 genes) represent phages or antibiotic resistance cassettes that are among the typical clonal diversity within species. These accessory sets of genes of a species are part of its pan-genome [28] and not consistently present in each representative of such species, implying that these accessory sets of genes are unlikely to display consistent co-abundance with their corresponding core genome or MGS in different intestinal samples. Notably, the abundance of several smaller CAGs that displayed a strong dependency association to specific MGS clusters were explored for their

predictive value for persistence capacity of microbes in the gut environment in longitudinal samples [17].

The genes encompassed within the 741 MGS were taxonomically classified by sequence similarity to reference genomes using BLASTN, allowing the mapping of the annotated genes for 115 MGS at species level [17]. Importantly, for 56 of these 115 MGS the taxonomy could be confirmed by paralleled co-abundance taxonomic mapping using the HITChip analyses of the same samples (Table S6.1). The fact that the remainder of the MGS could not be assigned may be due to the limitations of the HITChip platform (such as absence of the species-specific probes or probe cross-hybridization), technical issues (such as low MGS abundance, resulting in unfavourable signal to noise ratios), or computational limitations (such as miss-assembly of the MGS). Remarkably, the HITChip co-abundance mapping could assign another 113 MGS to uncultured isolates that remained phylogenetically unassigned on basis of the BLASTN approach (accession numbers in Table S6.1). Hence, in this way a total of 169 MGS could be associated with cultured and uncultured taxa detected by the phylogenetic HITChip microarray.

Combining co-abundance clustering with 16S rRNA gene profiling

The above results indicate that combining co-abundance gene-set clustering methods and MGS identification with 16S rRNA gene profiling results of the HITChip can connect MGS and the associated smaller CAG gene-sets to uncultured bacterial species. We dedicated our primary attention to those MGS that have the highest correlating co-abundance with the so far uncultured genus-like groups determined by the HITChip (ϱ > 0.8; Table 6.1). This led to the identification of 15 MGS with high correlation to uncultured taxa, encompassing seven genus-like groups. Remarkably, eight of these taxa belong to three genus-like groups that have previously been identified as being prevalent in >90% of the subjects analysed by Qin and co-workers [11], i.e., bacteria related to *Clostridium cellulosi*, *Oscillospira guillermondii*, and *Sporobacter termitidis*. Both *O. guillermondii*, and *S. termitidis* have shown interesting associations to host BMI and are often co-occurring in fecal samples (Box 1). Moreover, five of the highly correlating MGS belonged to two genus-like groups that were defined as bimodal distributed groups [16], i.e., the uncultured *Clostridiales* I and II. This subgroup of 13 MGS was selected for further gene-function investigation, aiming to illustrate the potential of combining gene co-abundance clustering and 16S rRNA gene profiling by the exemplary analyses presented below.

Box 1. Human *Oscillospira guillermondii* & *Sporobacter termitidis* isolates: important members of the microbiota?

- More abundant in the leaner sibling of monozygotic twins discordant for BMI (**Chapter 3**)
- Part of the "core gut microbiota" ([11] & **Chapter 3**)
- High co-occurrence (ϱ > 0.7) among MetaHIT subjects (HITChip data subjects from: [11, 15, 17])
- Bacteria related to *O.guillermondii* and *S.termitidis* showed higher (2 and 2.5-fold, respectively) relative abundance in subjects with high bacterial gene richness (p = 3.24E-12 and p = 9.60E-15, respectively; [15])

Table 6.1 Metagenomic species (MGS) and co-abundant gene clusters (CAGs) based on high correlation to bacterial taxa from uncultured genus-like groups (ρ > 0.8). The number of genes, average gene length, and total genetic information size are provided for each of the MGS. Grey shading indicates the MGS that were selected for further analysis (see main text for rationale).

Genus-like group represented by relatives of the indicated species [17]	MGS	Correlation (ρ)	Nr of genes	Average gene length (bp)	Total genetic information size (bp)
Clostridium cellulosi *	MGS.177	0.823	1,197	1,060.0	1,268,877
	MGS.180	0.817	1,298	1,064.0	1,381,104
	CAG.355	0.803	475	1,168.2	554,892
Clostridium colinum	MGS.102	0.849	1,419	1,046.4	1,484,874
	CAG.2283	0.804	22	953.9	20,985
Oscillospira guillermondii *	CAG.1235	0.842	95	816.2	77,538
	CAG.1284	0.805	30	1,114.9	33,447
	MGS.176	0.814	1,003	997.7	1,000,722
	CAG.3291	0.850	77	799.1	61,527
Sporobacter termitidis *	MGS.110	0.869	1,448	1,058.1	1,532,133
	MGS.124	0.841	1,144	1,077.9	1,233,066
	CAG.1667	0.819	495	1,030.3	510,015
	MGS.204	0.822	1,713	779.5	1,335,369
	MGS.240	0.869	1,348	1,046.3	1,410,381
	CAG.3371	0.8181	308	819.4	252,387
	CAG.3522	0.831	6	1,175.0	7,050
	MGS.49	0.828	1,682	980.4	1,649,100
	CAG.934	0.819	69	1,080.4	74,547
Tannerella	MGS.20	0.848	1,810	1,199.7	2,171,382
Uncultured *Clostridia*les I ‡	CAG.1839	0.800	49	1,136.3	55,680
	MGS.245	0.819	797	1,062.5	846,798
Uncultured *Clostridia*les II ‡	MGS.111	0.806	2,132	952.9	2,031,483
	MGS.138	0.805	1,155	1,134.0	1,309,806
	MGS.159	0.859	1,286	1,015.4	1,305,825
	CAG.1722	0.818	88	521.5	45,888
	CAG.4622	0.805	14	502.7	7,038
	MGS.74	0.901	1,671	1,161.7	1,941,240

* prevalence of >1% in 90% of subjects

‡ bimodal

Example 1: essential nutrients for microbes

The simplest explanation for the lack of successful culturing and isolation of microorganisms to date is likely the lack of adequate *in vitro* reproduction of the environmental conditions that allow them to grow. A multitude of variables in culture medium composition and physicochemical conditions during *in vitro* growth (e.g. pH, oxygenation, temperature, osmolarity, and micro- and macro-nutrient availability) can be essential for any given organism. The possibilities in classical culturing approaches are virtually endless and systematic evaluation of all possibilities

can only be achieved with highly significant resources and dedicated approaches, such as the culturomics approach [29]. Moreover, mutualistic metabolic dependencies between microbes or between microbes and host are not readily mimicked in classical *in vitro* culturing approaches. Nevertheless, recent advances in high throughput, miniaturized culturing strategies may enable improved mimicking of the natural habitat, and could include co-culture strategies for two or more bacterial species or strains from the same habitat, and could thereby allow the cultivation of previously uncultured bacteria (for a review see [30]).

The elucidation of genomic information of the uncultured bacterial species through co-abundance mapping of MGS and smaller CAGs can accelerate the design of culture conditions that may facilitate the growth of particular species or strains, or that may enhance the production of a particular metabolite required for the co-culturing or co-occurring inhabitants of the same ecological niche. Despite the intrinsic incompleteness of MGS and CAG information, it may still provide essential information for the rational design of culturing approaches for particular intestinal bacteria that have not been cultured to date.

Among the universal requirements for growth is the availability of all amino acids, either from environmental sources or by endogenous production. The functional annotations provided by KAAS allowed the estimation of the capacity of the 13 selected MGS to synthetize amino acids, revealing that most of these MGS are predicted to be able to synthesize 10 or more amino acids endogenously (Table 6.2), although MGS:49 (*Sporobacter termitidis* relative), MGS:138 (uncultured *Clostridiales* II), and MGS:159 (uncultured *Clostridiales* II) were predicted to encode biosynthetic pathways for 9 amino acids only, and MGS:245 (uncultured *Clostridiales* I) appeared to encode only 2 amino acid synthesis pathways. Notably, none of the 13 MGS appeared to encode the enzymes involved in the last steps of the biosynthesis of phenylalanine or tyrosine (Table 6.2). These findings imply that the species represented by the MGS may have multiple auxotrophies for amino acids and thus are predicted to require relatively rich environmental conditions for their growth. However, it cannot be excluded that the missing genes are not encompassed within the MGS due to gene sequence variability among the different representatives of the corresponding species, eliminating their inclusion in the corresponding MGS. Therefore, the relative autotrophy among the deposited MGS species (in ENA) was estimated (Figure 6.1) and no single MGS was predicted to be autotrophic for all amino acids. One would expect that at least some MGS would encompass the gene repertoire for the synthesis of all amino acids. However, the gastrointestinal tract contains the highest levels of proteases of any organ in the human body [31], encompassing both host and microbial proteases [32, 33]. Although ileum effluent shows higher protease activity compared to the feces, previous studies that specifically inhibited bacterial proteases suggest that a substantial part of the proteolytic activity in feces is of bacterial origin whereas in the ileum proteolytic activity originates primarily from the host itself [32, 33]. However, a recent study in diarrhoea predominant irritable bowel syndrome patients show that the most abundant fecal proteases are of pancreatic origin [34]. Not surprisingly, a recent study reported that certain microbiota members are more associated to the fecal proteolytic activity than others [35]. Taken together, the colon microbiota may favor proteolytic activity and/or harvesting proteolytic break-down products resulting from proteolysis,

Table 6.2 Genes encoding for final enzymatic conversions involved in amino acid production pathways, and their representation in the selected 13 metagenomic species (MGS). – for further information see Table 6.1.

Amino acid	Clostridium cellulosi *		Oscillospira guillermondii *	Sporobacter termitidis *					Uncultured Clostridiales I ‡	Uncultured Clostridiales II ‡			
	MGS:177	MGS:180	MGS:176	MGS:49	MGS:110	MGS:124	MGS:204	MGS:240	MGS:245	MGS:74	MGS:111	MGS:138	MGS:159
Alanine	-	√	-	-	-	-	-	-	-	-	-	-	-
Arginine	√	√	-	-	√	√	√	-	-	√	√	-	-
Asparagine	-	-	-	-	√	√	√	√	-	√	√	-	√
Aspartic acid	-	-	-	√	√	√	√	√	-	√	√	√	√
Cysteine	√	√	√	√	√	√	√	√	√	√	√	-	√
Glutamic acid	√	√	√	√	√	√	√	√	√	√	√	√	√
Glutamine	√	√	√	√	√	√	√	√	-	√	√	√	√
Glycine	√	√	√	√	√	√	√	√	-	√	√	√	√
Histidine	-	√	-	-	-	-	-	-	-	√	√	-	-
Isoleucine	√	√	-	-	√	√	√	-	-	√	√	√	-
Leucine	√	√	-	-	√	√	√	-	-	√	√	√	-
Lysine	-	√	√	√	√	√	√	-	-	√	√	-	-
Methionine	-	√	√	√	√	√	√	√	-	√	√	-	-
Phenylalanine	-	-	-	-	-	-	-	-	-	-	-	-	-
Proline	√	√	√	-	-	-	-	√	-	√	√	-	√
Serine	√	√	√	√	-	√	√	√	-	√	√	√	√
Threonine	√	√	√	√	√	√	√	√	-	√	√	√	√
Tryptophan	-	-	√	-	-	-	√	-	-	-	√	-	-
Tyrosine	-	-	-	-	-	-	-	-	-	-	-	-	-
Valine	√	√	-	-	√	√	√	-	-	√	√	√	-

initiated by themselves, other members in their vicinity, or by host enzymes, instead of investing in anabolic pathways to accommodate their amino acid requirement. Moreover, obligate metabolic cross-feeding has been proposed to be an ecological advantageous trait that stabilizes community structures [36], which could imply that microbes autotrophic for all amino acids are less likely to be encountered in this complex microbial system. Therefore, even though MGS may not represent complete genomes, and could miss specific genes or pathways, the analyses presented here suggests that specific amino acid-dependencies for growth can be assigned to each of the 13 MGS.

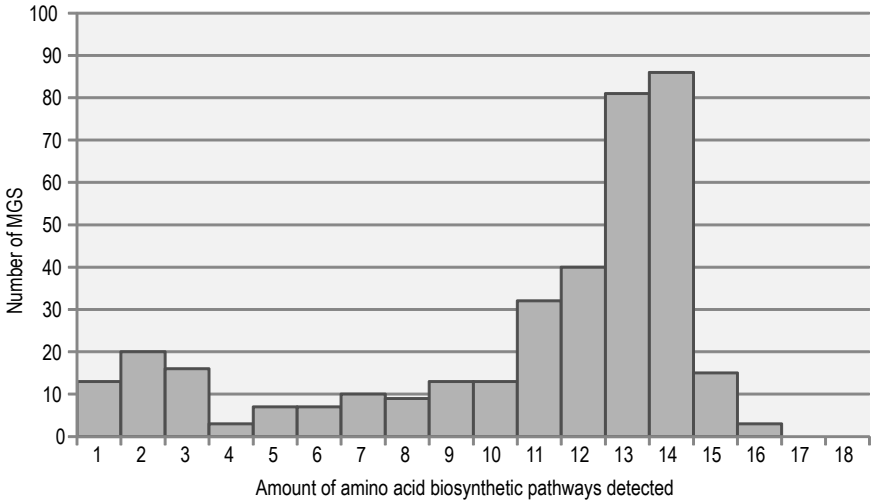
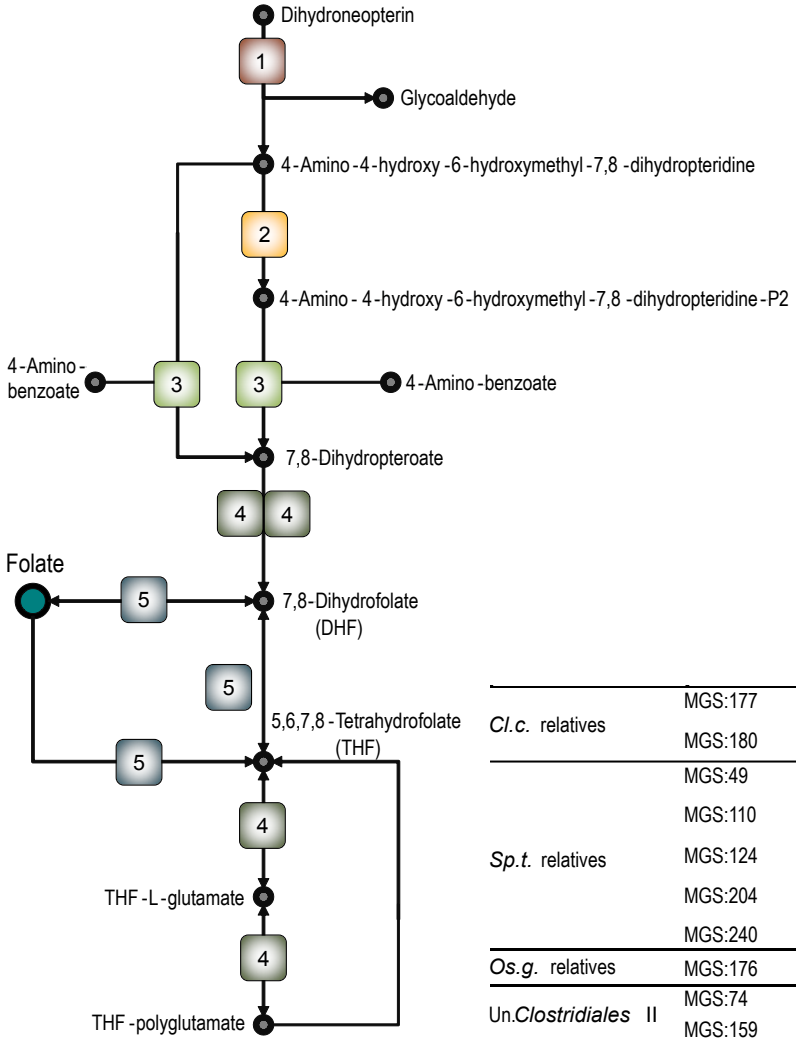


Figure 6.1 Histogram of the amount of amino acid biosynthesis pathways detected in the 373 deposited MGS assemblies that were deposited at the European Nucleotide Archive. These MGS passed the six high-quality draft assembly criteria from the Human Microbiome Project (HMP) [21]. The UniProt Knowledgebase (UniProtKB) was used to assess the amino acid biosynthetic pathway presence in the MGS.

Example 2: essential B-vitamins for host and microbe

The gut microbiota provides essential nutrients such as various B-vitamins to the human host [37, 38]. Human cells cannot synthesize various B-vitamins and rely on dietary sources, and/or their production and release by the gut microbiota. The selected MGS were evaluated for their predicted capacity to synthesize two of these B-vitamins, i.e., folate (vitamin B9) and cobalamin (vitamin B12). Interestingly, a previous study [39] analysed the presence of genes belonging to B-vitamin production pathways in 256 publicly available bacterial genomes, which were selected on basis of their prevalence and relative high abundance in many of the fecal sample metagenomic data that was employed to construct the 3.3M gene catalogue published by Qin and co-workers [11]. The results indicated that an apparently complete biosynthesis pathway for folate was present in 43% of the selected genomes, whereas cobalamin synthesis pathways were predicted to be encoded by 42% of these genomes [39]. The biosynthetic pathways for these two B-vitamins appeared to be present in most, if not all, *Fusobacteria*, and were concluded to be rare in Actinobacteria. In contrast, the distribution of these pathways among other phyla was more diverse, although most Bacteroidetes and Proteobacteria representatives encode the entire folate synthesis pathways, less than 14% of the Firmicutes encode for the complete folate synthesis pathway (although approximately 38% of the selected Firmicutes genomes encode for all enzymes except for those leading to 4-amino-benzoate formation). The cobalamin synthesis pathway distribution appeared to be even more scattered and was estimated to be present in half of the Bacteroidetes, in one out of four Proteobacteria genomes, and in 40% of the Firmicutes genomes [39]. Bacteria that do not synthesize these B-vitamins themselves are thought to express import systems and/or have effective retention systems (e.g., formation of polyglutamate tetrahydrofolate [40]) to enable the import and/or accumulation of folate and/or cobalamin (or one of its precursors) from the environment [41, 42].

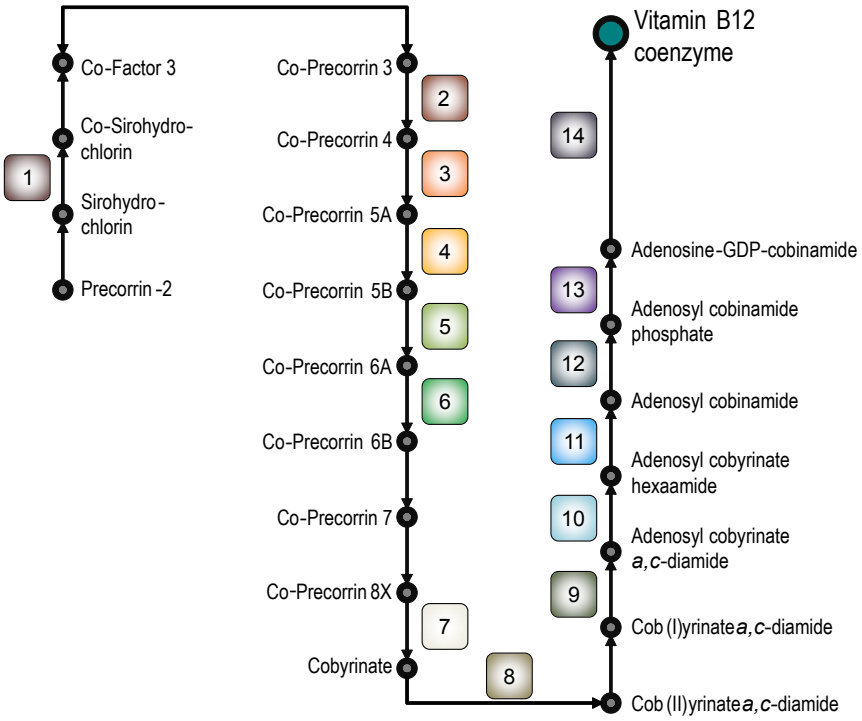
Figure 6.2 Subset of the folate biosynthesis pathway (KEGG map00790) on which the predicted enzymes are indicated from the selected metagenomic species (MGS). The table indicates which genes were predicted in each MGS. Predicted enzymes: 1 = EC 4.1.2.25; 2 = EC 2.7.6.3; 3 = EC 2.5.1.15; 4 = EC 6.3.2.12/17; 5 = EC 1.5.1.3. Abbreviations: Cl.c.; Clostridium cellulosi; Sp.t.: Sporobacter termitidis; Os.g.: Oscillospira guillermondii; Un.: Uncultured.



Cl.c. relatives	MGS:177	1, 3, 4
	MGS:180	1, 2, 3, 4
Sp.t. relatives	MGS:49	5
	MGS:110	5
	MGS:124	4
	MGS:204	4
	MGS:240	4, 5
Os.g. relatives	MGS:176	4, 5
Un.*Clostridiales* II	MGS:74	4
	MGS:159	5

Among the 13 selected MGS no complete pathways could be detected for either B-vitamin (Figure 6.2 and Figure 6.3), which could mean that this biosynthesis capacity is absent from these species or that it is not a part of their core genome. This analysis tends to imply that representative strains of the 13 selected MGS would require folate and cobalamin supplementation for growth. Considering that all 13 MGS are predicted to be Firmicutes species this is below the reported averages (see above; [39]) which may be related to their status of uncultured organism. However, import functions for these B-vitamins were also not detected in the selected MGS, which appears to be in apparent contradiction with the absence of the biosynthetic pathway, but may be due to the lack of an accurate prediction of substrate specificities of transport functions encoded by

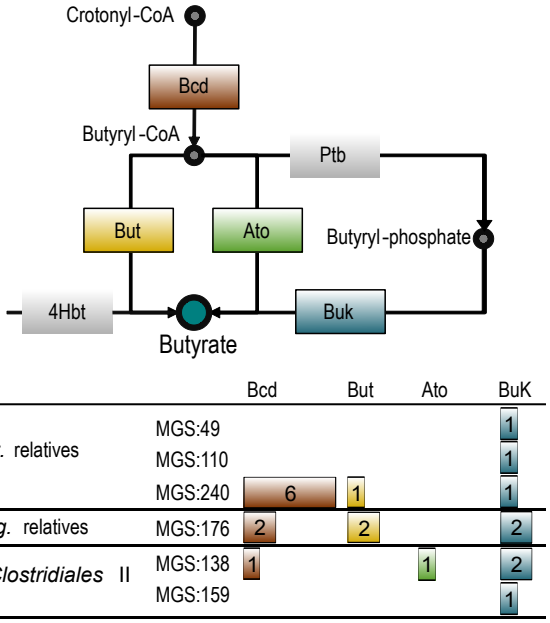
Figure 6.3 Cobalamin (vitamin B12) synthesis (anaerobic subset of KEGG map00860) on which the predicted enzymes are indicated from the selected metagenomic species (MGS). Cobalamin biosynthesis is one of the multi-enzyme B-vitamin production pathways, starting with adenosylcobalamin synthesis from precorrin-2. The table indicates which genes were predicted in each MGS. Predicted enzymes: 1 = EC 4.99.1.3; 2 = EC 2.1.1.131; 3 = EC 2.1.1.271; 4 = EC 3.7.1.12; 5 = EC 2.1.1.195; 6 = EC 1.3.1.106; 7 = EC 5.4.99.60; 8 = EC 6.3.5.11; 9 = EC 2.5.1.17; 10 = EC 6.3.5.10; 11 = EC 6.3.1.10; 12 = EC 2.7.1.156; 13 = EC 2.7.7.62; 14 = EC 2.7.8.26. Abbreviations see Fig 6.2.



Cl.c. relatives	MGS:180	9
Sp.t. relatives	MGS:124	2 3 4 5 7 8 10 11 12 13
	MGS:204	1 3 5 7 8 10
Os.g. relatives	MGS:176	2
Un.*Clostridiales* II	MGS:74	6 10 11 14

bacteria [41-43]. Conversely, most of the genes encoding enzymes involved in folate co-factor recycling were present in MGS:180 (*C. cellulosi*), and roughly half of the enzymes required for cobalamin biosynthesis appeared to be encoded within MGS:124 (*S. termitidis*). These observations suggest that representatives of these MGS could encode a complete cobalamin and/or folate synthesis pathway, particularly when taking into account the intrinsic incompleteness of MGS.

Figure 6.4 Schematic representation of butyrate production pathway indicating the involved enzymes and their presence in the metagenomic species (MGS) that were selected for analysis. Enzymes in grey boxes were not detected in any of the selected MGS. The table indicates which and how many different genes were predicted in each MGS. Abbreviations: Bcd) butyryl-CoA dehydrogenase (including electron transfer protein α, β subunits); 4Hbt) butyryl-CoA:4-hydroxybutyrate CoA transferase; But) butyryl-CoA:acetate CoA transferase; Ato) butyryl-CoA:acetoacetate CoA transferase (α, β subunits); Ptb) phosphate butyryltransferase; Buk) butyrate kinase. Other abbreviations as in Fig 6.2.

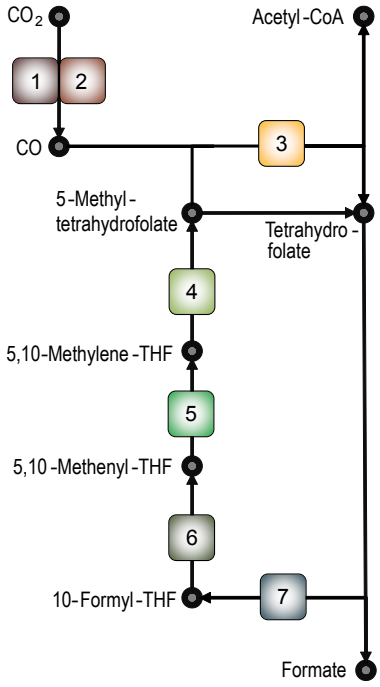


6

Example 3: SCFA production

The short chain fatty acids (SCFA) acetate, butyrate and propionate are the three main end products of anaerobic fermentation that occur in the human colon [44]. These SCFAs are readily absorbed by the host epithelial cells and have been proposed to play important and diverse roles in gut health (for a review see [45]). Especially butyrate has received a lot of attention since it is the preferential energy source of colonocytes, was shown to have anti-inflammatory properties by suppressing NK-$\varkappa$B activation [46], and has been implicated in modulation of genome wide gene expression in host cells through its inhibition of histone deacetylase [47, 48]. Moreover, in relation to its suggested key role in epithelial integrity and gut homeostasis, it has been shown that low levels of luminal butyrate or low abundance of butyrate producing microbial groups in the microbiota are frequently associated to different disease states, including inflammatory bowel diseases [49], diabetes [50, 51], and colon cancer [52-54]. Diet constituents, in particular dietary fibres or specific dietary polysaccharides such as resistant starch and xylans, have been shown to impact on luminal butyrate levels by affecting the growth and/or activity of butyrate producing bacteria in the gut either directly by polysaccharide substrate digestion, or indirectly through so-called cross-feeding mechanisms via the stimulation of other bacteria that produce lactate and/or acetate, which may subsequently serve as a substrate for secondary fermentation, and lead to enhanced butyrate production by co-occurring bacterial groups [55, 56].

Figure 6.5 Wood-Ljungdahl pathway (subset of KEGG map00720) on which the predicted enzymes are indicated from the selected metagenomic species (MGS). The table indicates which genes were predicted in each MGS. Predicted enzymes: 1 = EC 1.2.99.2; 2 = EC 1.2.7.4; 3 = EC 2.3.1.169; 4 = EC 1.5.1.20; 5 = EC 1.5.1.5; 6 = EC 3.5.4.9; 7 = EC 6.3.4.3. Other abbreviations as in Fig 6.2.



Cl.c. relatives	MGS:177	5 6 7
	MGS:180	5 6 7
Sp.t. relatives	MGS:49	1 2 3 4 5 6 7
	MGS:110	1 2 3 5 6 7
	MGS:124	1 2 3 4 5 6
	MGS:204	3 4 7
	MGS:240	1 7
Os.g. relatives	MGS:176	1 5 6
Un.*Clostridiales* I	MGS:245	5 6 7
Un.*Clostridiales* II	MGS:74	1 2 3 4 5 6 7
	MGS:111	4 5 6 7
	MGS:138	5 6
	MGS:159	5 6

Most studies addressing the butyrate production by the microbiota have evaluated the metabolism of the more abundant and cultured faecal isolates of the *Lachnospiraceae* and *Ruminococcaceae* groups and focussed on butyrate production via the acetyl-CoA pathway. However, investigating the butyrate production potential of the microbiota is far from trivial, as the "butyrate producers" are a heterogeneous, polyphyletic group of bacterial species, encompassing a diverse group of bacteria rather than few, well defined bacterial taxa. Moreover, from literature it is known that at least four pathways exist for butyrate production, i.e. the acetyl-CoA pathway and three amino acid dependent pathways: the 4-aminobutyrate, the glutarate, and the lysine pathways [22]. These pathways share one enzymatic step that involves crotonyl-CoA to butyryl-CoA conversion and is executed by the butyryl-CoA dehydrogenase electron-transferring flavoprotein complex (Bcd-Etfαβ). For each of the four main pathways the butyryl-CoA is transformed to butyrate either directly via butyryl-CoA transferase, or following phosphorylation and conversion via butyrate kinase (BuK). Recent fecal metagenomic analysis revealed that the most predominant butyrate production pathway in the gut microbiota of healthy individuals is the acetyl-CoA pathway [22], which is in agreement with observations made on the impact of dietary polysaccharides on butyrate production [55]. However, also the genes encoding the enzymes involved in amino acid-dependent pathways are detected in the gut metagenome, suggesting that the gut microbiota has more flexibility with respect to butyrate

production than initially assumed [22]. One of these, the butyrogenic lysine pathway, has recently been discovered in an intestinal isolate of the new intestinal genus *Intestinimonas* that is also capable of converting the Amadori product fructoselysine [57]. The butyrate production potential of the 13 selected MGS was investigated by determining whether they encompassed genes encoding the common butyryl-CoA dehydrogenase as well as genes involved in one or more of the four butyrate pathways (butyryl-CoA transferases or butyrate kinase). Six out of the 13 MGS encoded at least one enzyme involved in a butyrate production pathway, whereas three MGS encoded both the butyryl-CoA dehydrogenase and at least one of the terminal enzymes involved in the different pathways (Figure 6.4). Interestingly, MGS:176 (*O. guillermondii*) and MGS:138 (Uncultured *Clostridiales* II) encode both a butyryl-CoA transferase and a butyrate kinase, which has been reported before but is rather uncommon [22]. These indications should be interpreted with caution, as the enzyme function and substrate specificity prediction on basis of homology analysis can be hampered by the presence of conserved protein-sequences in enzymes with quite distinct metabolic functions [58]. Moreover, the existence of alternative and distinct butyryl-CoA transferase enzyme families has been proposed [59], but were not yet taken into consideration here. Consequently, the current prediction may underestimate the true amount of uncultured MGS that encompass butyrate production encoding genes and *in vitro* growth combined with biochemical tests would be required to reliably assess the butyrate production capacity of these organisms. Nevertheless, this analysis predicts that at least three of the uncultured intestinal MGS/taxa we analyse here, could produce butyrate, which reflects metabolic information of these bacteria that was unavailable *a priori*.

Interestingly, the relatively rare Wood-Ljundahl pathway for acetate production appeared to be present in at least three of the selected MGS, which enables bacteria to capture CO_2 and H_2, and leads to the production of acetyl-CoA (Figure 6.5)[60]. This pathway has been specifically described for homoacetogenic bacteria, such as *Blautia hydrogenotrophica*. Although acetate production in itself is not a remarkable capacity in fermentative bacteria, and is present in many gut-associated species, the homoacetogenic bacteria are not dominantly present among gut microbiota and are outnumbered by the fermentative (non-acetogenic) actetate producers [61]. It was estimated that approximately one third of the acetate pool is produced by homoacetogenic bacteria [60]. Nevertheless, the capacity of the homoacetogenic bacteria to remove H_2 from the gut-lumen may be of importance in the overall functioning of the gut ecosystem. While H_2 production is a common way for strict anaerobes to dispose of the reducing equivalents that are formed during fermentation [62], H_2 accumulation can have ecosystem-wide effects since primary fermenters are forced to hoard their reduced compounds, such as butyrate and ethanol, while re-oxidation of pyridine nucleotides is inhibited [63].

Conclusion

Various roles in orchestrating human physiology have been ascribed to the gut microbiota and the research involved in deciphering this role of the microbiota has expanded drastically over the past years. However, our current knowledge is not sufficient to fully grasp the vast function potential residing within the microbiota, which constrains our options to modulate the host-

microbiota interactions for the better of human health. Uncultured bacteria account for the largest knowledge gap, not only because they account for the majority of the gut eco-system but also because various important characteristics of the microbiota involve a dominant role of uncultured bacteria, such as the common core [11], bacterial richness [15], and bimodal taxa [16]. Currently, reference genomes cannot cover actual biological diversity of the gut microbiota to a satisfactory degree, which hampers attempts to functionally reconstruct the gut ecosystem. Therefore, characterization of the uncultured bacterial gut inhabitants will be crucial to fully comprehend gut functions as a composite of mammalian and microbial functions. As a first step in the characterization of the uncultured microbes, we have shown several examples where 16S rRNA gene profiling and shotgun-metagenomic analysis were combined to extract knowledge of the genetic potential harboured by the uncultured bacteria. Knowledge about the genetic potential of uncultured bacteria can further our insight in the driving forces within this complex ecological system, and can provide information that can potentially be exploited to obtain cultured representatives of so far uncultured representatives of the gut microbiota. Eventually, such an expanded gut-microbiota knowledgebase can support the rational design of interventions aiming to improve or sustain human health via modulation of the intestinal microbiota composition and/or activity.

The exemplary analyses that combine MGS and 16S rRNA gene abundance profiling presented here address several interesting aspects of the metabolic capacities of the microbiota that are considered of relevance for the functioning of the gut ecosystem. The results of the analyses of amino acid and vitamin production capacity illustrate how the integration of MGS and 16S rRNA composition data could tentatively predict essential nutrients that are required for the cultivation of gut bacteria that have not been cultured to date. Although these predictions are only tentative and may underestimate the metabolic capacity of individual isolates corresponding to a certain MGS, they do exclude the requirement for nutrients for which the MGS encodes the entire biosynthetic pathway. Moreover, these analyses underpin that MGS and 16S rRNA gene correlation analysis can predict specific metabolic capacities within the central carbon metabolism that MGS of interest are able to execute (e.g., butyrate production and acetogenic metabolism). With such knowledge growth media and conditions, for instance by introducing H_2 as electron donor for the homoacetogenic bacteria, as well as functional activity screening tools can be more specifically designed and employed to guide isolation and cultivation scenarios for representatives of these uncultured bacterial species.

The basis for the analyses described here is the correlation between MGS abundance patterns in individuals and the corresponding 16S rRNA gene abundance, which provides associations between the MGS and the species it represents and thereby can define the core genome of uncultured microorganisms from the intestinal niche. Analogously, this approach can be employed to investigate uncultured microorganisms in any other niche, provided that the appropriate combination of information (deep shotgun sequencing metagenomes and 16S rRNA gene composition profiles) is available in sufficient numbers of samples to allow co-abundance and correlation analyses. Improvement of the statistical algorithms for MGS detection may further improve their reliability, and functional predictions can be more accurately

assessed by analysing metabolic pathways rather than inferring their presence from taxonomy, as for instance can be done with PICRUSt [12]. However, the intrinsic incompleteness of the MGS gene-sets due to genomic diversity of the species they represent hampers the complete assessment of pathway presence and absence in members of a species. Moreover, incompleteness of a pathway may also not necessarily mean that the corresponding microorganism is unable to produce particular metabolic end-products, because especially in complex microbial consortia, pathway intermediates may be available in the environment because they are produced by other members of the community.

An important goal in the area of gut microbiota and cooperative metabolism of the ecosystem is the ability to predict and visualize the metabolic performance of the microbial community in metabolic network analyses that include nutrient flux across different members of the community. This ambition will require the integration of systems biology combined with metabolic modelling [64] and could potentially enable the prediction of the metabolic handling of dietary ingredients by the microbial community, allowing the rational design of dietary strategies intending to modulate the ecosystem's metabolic performance. Achieving such goals depends on improved characterization of the uncultured proportion of the microbiota that can be initiated using the culture independent approaches illustrated here, which may support the subsequent and crucial step of *in vitro* cultivation and characterization of representative isolates of the gut microbiota.

Acknowledgements

This research was partly funded by the European Community's Seventh Framework Programme (FP7/2007–2013) MetaHIT, Grant agreement HEALTH-F4-2007-201052, coordinated by S Dusko Ehrlich (Institut National de la Recherche Agronomique,France). We thank the MetaHIT Consortium members for discussions, DNA extracts of the fecal samples, and the (processed) sequence data. The MetaHIT Consortium: H Bjørn Nielsen, Mathieu Almeida, Agnieszka S Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R Plichta, Laurent Gautier, Anders G Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B. Quintanilha dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S 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, David W Ussery, Takuji Yamada, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, S Dusko Ehrlich, Alexandre Jamet, Alexandre Mérieux, Antonella Cultrone, Antonio Torrejon, Benoit Quinquis, Christian Brechot, Christine Delorme, Christine M'Rini, Emmanuelle Maguin, Encarna Varela, Eric Guedon, Falony Gwen, Florence Haimet, François Artiguenave, Gaetana Vandemeulebrouck, Gérard Denariaz, Ghalia Khaci, Hervé Blottière, Jan Knol, Jean Weissenbach, Johan E T van Hylckama Vlieg, Jørgensen Torben, Julian Parkhill, Keith Turner, Maarten van de Guchte, Maria Antolin, Maria Rescigno, Muriel Derrien, Nathalie Galleron, Nicolas Sanchez, Niels Grarup, Patrick Veiga, Raish Oozeer, Rozenn Dervyn, Séverine Layec, Thomas Bruls & Yohanan Winogradski.

Supplementary information

Supplementary tables

Table S6.1 MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:1	*Subdoligranulum* sp. 4_3_54A2FAA	95.4	2782	100	2915	0.800750581	
MGS:2	*Parabacteroides distasonis*	97.8	2550	100	2608	0.789958547	
MGS:3	*Bacteroides uniformis*	88.2	2160	99.4	2450	0.878694791	
MGS:4	*Escherichia coli*	96.1	3385	99.1	3523		
MGS:5	*Flavonifractor plautii*	97.8	2914	100	2981	0.853129993	
MGS:6	*Bacteroides vulgatus*	79.4	1803	85.7	2271	0.763021226	
MGS:7	-	-	0	-	2447	0.854735154	AF499907
MGS:8	-	-	1	-	2248	0.705235724	AB064762
MGS:9	*Blautia wexlerae* (*Ruminococcus* sp. 5_1_39BFAA)	85.4	1701	99.8	1992	0.857621093	
MGS:10	*Bilophila wadsworthia*	92.6	3606	93.2	3896		
MGS:12	*Eubacterium hallii*	92.9	1952	100	2101	0.725750825	
MGS:13	*Roseburia intestinalis*	91.2	2293	100	2513	0.701942406	
MGS:14	*Odoribacter splanchnicus*	96.6	2443	100	2530	0.76749819	
MGS:15	*Roseburia inulinivorans*	85.3	1810	100	2123	0.800780468	
MGS:16	*Roseburia hominis*	92.7	2164	100	2334	0.800832354	
MGS:18	-	-	2	-	1949	0.754364903	AJ408979
MGS:19	*Coprococcus comes*	93.1	1789	100	1922	0.710622466	
MGS:20	-	-	1	-	2119	0.847189783	AY916176
MGS:21	*Bacteroides caccae*	96.6	2047	100	2119	0.823914125	
MGS:22	*Bacteroides ovatus*	52.4	1377	66.6	2630	0.696670046	
MGS:24	-	-	0	-	1959	0.911908345	AF132253
MGS:25	*Eubacterium hadrum*	96.7	1629	100	1684	0.838773536	
MGS:27	*Clostridium leptum*	98.7	2254	100	2283	0.778172788	

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:28	*Dorea formicigenerans*	98.3	1774	100	1805	0.603034984	
MGS:30	*Lachnospira*ceae bacterium 3_1_57FAA_CT1	91	5309	100	5836		
MGS:31	-	-	0	-	2884	0.625966951	AF132268
MGS:32	*Bacteroides faecis*	94.1	1985	99.9	2110	0.696484633	
MGS:34	-	-	11	-	1799	0.865178704	AB099720
MGS:35	*Haemophilus parainfluenzae*	37.2	1431	93	3843		
MGS:36	*Eubacterium rectale*	94.5	1464	100	1550	0.871710506	
MGS:37	-	-	0	-	1919	0.756576771	AJ408969
MGS:39	*Ruminococcus obeum*	92.7	1614	100	1741	0.73590172	
MGS:40	*Bacteroides thetaiotaomicron*	94.1	1798	100	1910		
MGS:41	-	-	0	-	1763	0.859549422	AJ608246
MGS:42	*Dorea longicatena*	85.9	1708	100	1988	0.757108974	
MGS:43	-	-	0	-	1928	0.797246631	AB064711
MGS:44	-	-	0	-	2284	0.641552774	AF132242
MGS:45	-	-	0	-	2019	0.647801468	AY684388
MGS:46	-	-	1	-	1817	0.704097802	AF132258
MGS:47	*Bacteroides fragilis*	96.8	2831	100	2926	0.780466164	
MGS:48	*Parabacteroides merdae*	88.3	2007	97.4	2273	0.774121177	
MGS:49	-	-	0	-	1942	0.828067192	AY916335
MGS:50	-	-	0	-	1900	0.661875247	L34623
MGS:51	*Tannerella* sp. 6_1_58FAA_CT1	92.9	2218	100	2387		
MGS:52	-	-	1	-	2045	0.806138613	AF132274
MGS:54	-	-	1	-	2684	0.646489719	AB099796
MGS:55	-	-	2	-	1599	0.733158041	AY920004
MGS:56	-	-	1	-	1952	0.857241378	AY920027
MGS:57	-	-	16	-	1835	0.611610553	X85097
MGS:58	-	-	0	-	2386	0.864916876	AY916157
MGS:59	*Clostridium bolteae*	88.2	2319	99.7	2628		
MGS:60	-	-	1	-	2511	0.665787303	AY916245
MGS:61	*Ruminococcus torques*	94	1651	100	1756	0.84373442	
MGS:62	-	-	1	-	1868	0.675568843	AJ409006
MGS:64	*Clostridium innocuum*	99.8	3390	100	3397		
MGS:65	-	-	1	-	1879	0.880814958	AF153851
MGS:67	*Alistipes putredinis*	93.6	1381	99.8	1476	0.874399678	

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:68	*Alistipes finegoldii*	60.2	1066	61.7	1770		
MGS:69	*Bifidobacterium longum*	95.4	1275	96.5	1337		
MGS:70	-	-	2	-	1578	0.632620344	AJ408989
MGS:72	*Eubacterium eligens*	63.7	1171	99.9	1837	0.738074987	
MGS:74	-	-	0	-	2085	0.900765499	AY920203
MGS:75	-	-	0	-	2059	0.907713564	AY920012
MGS:76	-	-	1	-	1716	0.631253075	AY916189
MGS:78	-	-	0	-	1516	0.725588634	AY916166
MGS:79	*Streptococcus salivarius*	74.1	1026	87.8	1385	0.770695476	
MGS:80	*Eubacterium siraeum*	85.2	1515	99.9	1779	0.711974788	
MGS:82	-	-	2	-	1297	0.788641875	AY916290
MGS:84	-	-	1	-	2327	0.730802573	AB080886
MGS:86	-	-	1	-	1858	0.628211201	AJ608250
MGS:87	-	-	0	-	1877	0.644177063	AY920092
MGS:88	*Alistipes indistinctus*	98.3	2215	100	2254		
MGS:89	*Eubacterium ventriosum*	98.2	1510	100	1537	0.849673098	
MGS:90	-	-	0	-	1172	0.73170705	AB064755
MGS:91	-	-	1	-	1100	0.601466751	AY920013
MGS:92	-	-	0	-	2772	0.601297203	AY916203
MGS:93	-	-	1	-	1809	0.683960071	AJ608220
MGS:95	-	-	0	-	2120	0.654530033	AY916383
MGS:96	-	-	0	-	2877	0.785437169	AY916380
MGS:98	-	-	1	-	2183	0.682909589	AY920210
MGS:101	-	-	0	-	2685	0.769451944	AB099779
MGS:102	-	-	0	-	1745	0.848679929	AY684406
MGS:103	-	-	0	-	1526	0.814126544	AY916281
MGS:104	*Ruminococcus lactaris*	95.2	1720	100	1807	0.832677958	
MGS:106	-	-	1	-	1006	0.84993019	AJ408966
MGS:108	-	-	1	-	1445	0.862600278	L76600
MGS:109	*Bacteroides eggerthii*	94.2	1896	99.8	2013		
MGS:110	-	-	0	-	1843	0.868686789	AB099729
MGS:111	-	-	0	-	2632	0.806259532	AB080881
MGS:112	*Veillonella parvula*	87.4	1872	98.4	2141	0.707328454	
MGS:115	-	-	0	-	2058	0.776524147	AJ608247
MGS:116	*Paraprevotella clara*	84.6	1854	93.2	2191	0.651714168	
MGS:119	*Bifidobacterium adolescentis*	95.5	1065	99.8	1115		

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:120	*Bacteroides stercoris*	91	1450	99.9	1594		
MGS:121	-	-	1	-	1396	0.742432527	AF132255
MGS:124	-	-	0	-	1442	0.840611076	AJ608220
MGS:125	*Faecalibacterium prausnitzii*	93.4	960	100	1028	0.615437206	
MGS:126	*Ruminococcus gnavus*	86.5	1957	99.9	2263	0.602614754	
MGS:127	-	-	0	-	1921	0.771561067	AF132276
MGS:128	*Faecalibacterium prausnitzii*	87.5	852	100	974	0.799197062	
MGS:129	-	-	0	-	1447	0.703131914	AY916357
MGS:131	*Coprococcus eutactus*	22.2	449	100	2024	0.839887665	
MGS:132	*Clostridium clostridioforme*	86.2	2548	93.6	2955	0.613844233	
MGS:134	*Clostridium symbiosum*	99.8	4166	100	4173		
MGS:135	*Sutterella wadsworthensis*	93.7	1988	100	2121	0.601941048	
MGS:138	-	-	0	-	1627	0.804518065	AY916346
MGS:139	*Parasutterella excrementihominis*	15.2	270	100	1777	0.781581752	
MGS:140	-	-	0	-	2391	0.758260725	AJ270485
MGS:147	*Blautia hydrogenotrophica*	97.9	2839	100	2899		
MGS:148	*Streptococcus parasanguinis*	84.3	1448	99	1718	0.677453688	
MGS:149	*Clostridium* sp. M62/1	97.5	2439	100	2502		
MGS:151	-	-	1	-	1706	0.842928562	AB099737
MGS:154	*Akkermansia muciniphila*	88.4	1499	100	1696		
MGS:156	-	-	0	-	1742	0.626700077	AY916360
MGS:157	-	-	1	-	1764	0.659368192	AF153864
MGS:158	*Bacteroides cellulosilyticus*	83.7	2715	92.1	3242	0.733702906	
MGS:159	-	-	0	-	1568	0.859235715	AY920207
MGS:160	*Bacteroides clarus*	97.2	1390	100	1430		
MGS:162	*Bacteroides coprocola*	89.9	1864	99.9	2073		
MGS:163	-	-	0	-	1690	0.864696412	AY916263
MGS:164	*Prevotella copri*	77.4	1276	99.9	1648	0.772046182	
MGS:165	-	-	0	-	1556	0.743687078	AB064903
MGS:166	*Collinsella aerofaciens*	39.8	511	99.4	1283		
MGS:169	-	-	0	-	2053	0.727916088	AJ408957
MGS:170	-	-	0	-	2055	0.634510951	AB099728
MGS:171	-	-	1	-	2260	0.667662236	AB064712
MGS:172	-	-	0	-	1415	0.604219366	AF052411
MGS:173	-	-	0	-	2346	0.658881928	AY916376

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:176	-	-	1	-	1209	0.814344539	AY916203
MGS:177	-	-	0	-	1506	0.823468102	AB099734
MGS:178	-	-	0	-	2065	0.756774277	AY920195
MGS:180	-	-	1	-	1580	0.816957589	AB099730
MGS:183	*Clostridium ramosum*	99.4	2943	100	2961		
MGS:186	*Methanobrevibacter smithii*	94.4	1438	100	1523		
MGS:188	-	-	0	-	1364	0.741863125	AJ608220
MGS:193	-	-	1	-	1735	0.604614395	AY916263
MGS:195	*Faecalibacterium prausnitzii*	67	650	100	970	0.600218324	
MGS:196	-	-	1	-	1661	0.696494022	AY916143
MGS:198	-	-	1	-	1116	0.622574494	AJ408991
MGS:199	-	-	0	-	1358	0.667295035	AB099721
MGS:201	-	-	0	-	2523	0.717076784	AY916333
MGS:202	-	-	0	-	1504	0.603306693	AJ408987
MGS:203	*Bacteroides finegoldii*	94.3	1377	100	1460		
MGS:204	-	-	0	-	1883	0.822788632	AB099728
MGS:206	-	-	0	-	782	0.683997125	AF052411
MGS:207	-	-	0	-	1942	0.637055835	AB099771
MGS:210	-	-	2	-	1922	0.657651625	AY916147
MGS:211	*Bacteroides plebeius*	61	1012	100	1660	0.633731965	
MGS:215	*Clostridium glycyrrhizinilyticum*	89.7	2179	100	2429		
MGS:216	*Coprobacillus* sp. 29_1	91.6	3189	100	3481		
MGS:217	-	-	0	-	1473	0.792762524	AY916292
MGS:218	*Dialister invisus*	87.6	1245	100	1422	0.678120844	
MGS:220	-	-	0	-	989	0.641884158	AF132255
MGS:222	*Bacteroides dorei*	70.9	568	99	801	0.88736384	
MGS:223	-	-	0	-	2194	0.6910974	AB099793
MGS:224	*Clostridium hathewayi*	97.7	3251	100	3327		
MGS:225	-	-	0	-	1429	0.631745505	AY919923
MGS:226	-	-	0	-	1687	0.75655982	AY916160
MGS:227	-	-	0	-	2152	0.674938428	AB064749
MGS:228	-	-	0	-	2919	0.744722695	AY920204
MGS:232	*Eubacterium biforme*	15.2	210	100	1386	0.726386301	
MGS:233	*Parasutterella excrementihominis*	89.9	1045	100	1162	0.722278933	

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:234	*Bifidobacterium bifidum*	98.4	1431	99.9	1455		
MGS:235	-	-	2	-	1641	0.779056605	AY920075
MGS:236	*Streptococcus thermophilus*	93	959	98.2	1031		
MGS:237	-	-	0	-	1596	0.673160028	AY916255
MGS:240	-	-	0	-	1715	0.869494096	AY916387
MGS:241	-	-	0	-	821	0.651748757	AJ408991
MGS:244	-	-	0	-	2004	0.609030396	AF153860
MGS:245	-	-	0	-	1074	0.819153578	AY920189
MGS:246	*Parabacteroides johnsonii*	89.3	820	96.8	918	0.664988761	
MGS:247	*Veillonella atypica*	90.2	1347	99.9	1494	0.610745939	
MGS:249	-	-	0	-	1173	0.68853907	AY916203
MGS:254	-	-	0	-	1752	0.678765749	AB064810
MGS:256	-	-	0	-	999	0.676705275	AY916227
MGS:258	*Ruminococcaceae* bacterium D16	83	1446	100	1743		
MGS:259	*Butyrivibrio crossotus*	97.1	1649	100	1699	0.667831334	
MGS:263	*Bifidobacterium pseudocatenulatum*	75.9	870	87.6	1147		
MGS:264	*Clostridium* sp. L2-50	87.5	1111	100	1270	0.754232497	
MGS:266	-	-	1	-	1440	0.65587255	AF132234
MGS:269	-	-	0	-	1153	0.686100342	AY916150
MGS:273	-	-	0	-	1040	0.661449841	AY916291
MGS:276	*Anaerostipes caccae*	99.8	2823	100	2830		
MGS:280	*Eggerthella lenta*	77.4	1180	78.7	1525		
MGS:287	*Phascolarctobacterium succinatutens*	55.4	917	100	1656	0.61482286	
MGS:290	*Catenibacterium mitsuokai*	48	776	100	1615	0.612950974	
MGS:295	*Clostridium aldenense*	6.6	173	98.3	2602		
MGS:301	-	-	0	-	1751	0.667040118	AY916148
MGS:302	-	-	0	-	1134	0.614424208	AY916348
MGS:309	-	-	0	-	2046	0.630613637	AY916382
MGS:310	*Clostridium spiroforme*	91.4	1630	100	1784		
MGS:315	*Bacteroides intestinalis*	90.2	1069	100	1185		
MGS:321	-	-	0	-	865	0.662151962	AY920083
MGS:322	*Bacteroides thetaiotaomicron*	67.5	495	74.8	733	0.658772034	
MGS:323	*Erysipelotrichaceae bacterium*	99	2995	100	3025		

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:325	*Acidaminococcus intestini*	94.5	1739	96.4	1841		
MGS:329	-	-	0	-	1121	0.681050991	AJ408971
MGS:331	-	-	0	-	813	0.648405543	AY916179
MGS:333	*Bacteroides coprophilus*	96.1	1957	100	2036		
MGS:335	*Desulfovibrio piger*	77.7	1128	99.9	1452		
MGS:337	*Bifidobacterium animalis*	98.6	1738	100	1763		
MGS:339	*Veillonella sp.* oral taxon 158	79.6	868	98.3	1091		
MGS:347	-	-	0	-	1260	0.690033686	AB064863
MGS:348	*Clostridium nexile*	92	1701	99.8	1848		
MGS:350	-	-	0	-	1000	0.62198715	AY919923
MGS:354	-	-	0	-	1059	0.601126792	AY916165
MGS:364	*Blautia hansenii*	92.4	1783	100	1929		
MGS:367	*Lactobacillus ruminis*	98.6	1651	100	1675		
MGS:373	[*Lachnospiraceae* bacterium]	94.3	1973	100	2093		
MGS:375	*Eubacterium dolichum*	93.9	1755	100	1870		
MGS:377	*Megamonas funiformis*	95	1680	100	1768		
MGS:387	*Mitsuokella multacida*	7.5	167	98.8	2238		
MGS:401	-	-	0	-	1595	0.809113016	AB064779
MGS:405	*Enterobacter hormaechei*	40.5	1051	94.8	2592		
MGS:407	-	-	0	-	864	0.80554459	AY916263
MGS:418	-	-	0	-	1326	0.75222403	AY982155
MGS:453	*Enterococcus faecium*	99.5	2053	100	2064		
MGS:480	-	-	0	-	703	0.631452041	AY916269
MGS:507	-	-	0	-	904	0.65242484	AB080873
MGS:511	*Clostridium clostridioforme*	93.3	706	100	757		
MGS:512	*Dialister succinatiphilus*	87.4	1449	100	1658		
MGS:544	*Mitsuokella multacida*	85.4	857	100	1003		
MGS:546	*Acidaminococcus fermentans*	93.7	1603	100	1711		
MGS:558	*Bacteroides fragilis*	94.4	1492	99.9	1580		
MGS:561	*Odoribacter laneus*	93.5	2541	100	2717		
MGS:564	*Bacteroides intestinalis*	85.3	632	100	741		
MGS:570	*Megasphaera elsdenii*	90.3	1300	100	1440		
MGS:621	*Peptostreptococcus anaerobius*	94.4	1432	99.6	1517		
MGS:629	*Prevotella* stercorea	86.1	876	100	1017		

Table S6.1 (continued) MGS with taxonomic identification by BLASTn and/or HITChip association.

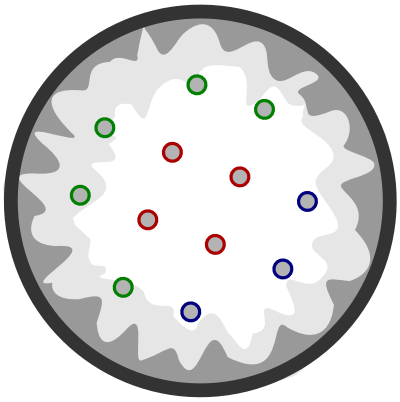
MGS ID	Species (primary taxonomy)	MGS genes that match the primary taxonomy			Number of genes in the MGS	HITChip confirmation (Spearman correlation coefficient)	Accession nr
		% of total genes	nr. of genes	% of annotated genes			
MGS:630	*Lactobacillus salivarius*	93.4	1955	99.8	2094		
MGS:634	*Klebsiella variicola*	96.3	1276	99.5	1325		
MGS:640	*Collinsella intestinalis*	90.9	966	99.9	1063		
MGS:649	*Fusobacterium nucleatum*	95.7	1346	98.9	1407		
MGS:665	*Coprococcus eutactus*	97.2	1072	100	1103		
MGS:679	*Fusobacterium gonidiaformans*	93.4	2205	99.4	2362		
MGS:692	*Parvimonas micra*	79.4	620	86.6	781		
MGS:702	*Bacteroides salanitronis*	18.6	320	100	1721		
MGS:719	*Lactobacillus amylovorus*	96.8	1557	99.4	1609		
MGS:1329	*Clostridium bartlettii*	94.3	2303	100	2442	0.658218395	
MGS:1353	*Collinsella stercoris*	62.5	815	99.1	1304		

References

1. **Savage DC.** Microbial ecology of the gastrointestinal tract. Ann. Rev. Microbiol. 1977;31:107-133.
2. **Yang X, Xie L, Li Y, et al.** More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body. PLoS One 2009;4(6):e6074.
3. **Wei C and Brent MR.** Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 2006;7:327.
4. **O'Hara AM and Shanahan F.** The gut flora as a forgotten organ. EMBO Rep. 2006;7(7):688-693.
5. **Lecuit M, Vandormael-Pournin S, Lefort J, et al.** A transgenic model for listeriosis: role of internalin in crossing the intestinal barrier. Science 2001;292(5522):1722-5.
6. **Bhowmick R, Ghosal A, Das B, et al.** Intestinal adherence of Vibrio cholerae involves a coordinated interaction between colonization factor GbpA and mucin. Infect Immun 2008;76(11):4968-77.
7. **McCormick BA, Colgan SP, Delp-Archer C, et al.** Salmonella typhimurium attachment to human intestinal epithelial monolayers: transcellular signalling to subepithelial neutrophils. J Cell Biol 1993;123(4):895-907.
8. **Winter SE, Thiennimitr P, Winter MG, et al.** Gut inflammation provides a respiratory electron acceptor for Salmonella. Nature 2010;467(7314):426-9.
9. **de Vos WM and de Vos EA.** Role of the intestinal microbiome in health and disease: from correlation to causation. Nutrition Reviews 2012;70(suppl 1):S45-S56.
10. **Rajilic-Stojanovic M and de Vos WM.** The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 2014;38(5):996-1047.
11. **Shreiner AB, Kao JY, and Young VB.** The gut microbiome in health and in disease. Current Opinion in Gastroenterology 2015;31(1):69-75.
12. **Qin J, Li R, Raes J, et al.** A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59-65.
13. **Langille MGI, Zaneveld J, Caporaso JG, et al.** Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotech 2013;31(9):814-821.
14. **Fitzsimons MS, Novotny M, Lo CC, et al.** Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. Genome Res 2013;23(5):878-88.
15. **Lukjancenko O, Wassenaar TM, and Ussery DW.** Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol 2010;60(4):708-20.
16. **Le Chatelier E, Nielsen T, Qin J, et al.** Richness of human gut microbiome correlates with metabolic markers. Nature 2013;500(7464):541-6.
17. **Lahti L, Salojärvi J, Salonen A, et al.** Tipping elements in the human intestinal ecosystem. Nat Commun 2014;5.
18. **Nielsen HB, Almeida M, Juncker AS, et al.** Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 2014;32(8):822-8.
19. **Sunagawa S, Mende DR, Zeller G, et al.** Metagenomic species profiling using universal phylogenetic marker genes. Nat Meth 2013;10(12):1196-1199.
20. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
21. **Powell S, Szklarczyk D, Trachana K, et al.** eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res 2012;40(Database issue):D284-9.
22. **Chain PS, Grafham DV, Fulton RS, et al.** Genomics. Genome project standards in a new era of sequencing. Science 2009;326(5950):236-7.
23. **Vital M, Howe AC, and Tiedje JM.** Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta) genomic Data. mBio 2014;5(2):e00889-14.
24. **Kanehisa M and Goto S.** KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27-30.
25. **Kanehisa M, Goto S, Sato Y, et al.** Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42(Database issue):D199-205.
26. **Moriya Y, Itoh M, Okuda S, et al.** KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 2007;35(Web Server issue):W182-5.
27. **Pop M.** Genome assembly reborn: recent computational challenges. Brief Bioinform 2009;10(4):354-66.
28. **Wooley JC, Godzik A, and Friedberg I.** A primer on metagenomics. PLoS Comput Biol 2010;6(2):e1000667.
29. **Medini D, Donati C, Tettelin H, et al.** The microbial pan-genome. Current Opinion in Genetics & Development 2005;15(6):589-594.
30. **Lagier JC, Armougom F, Million M, et al.** Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 2012;18(12):1185-93.
31. **Stewart EJ.** Growing unculturable bacteria. J Bacteriol 2012;194(16):4151-60.

32. **Antalis TM, Shea-Donohue T, Vogel SN, et al.** Mechanisms of Disease: protease functions in intestinal mucosal pathobiology. Nature clinical practice. Gastroenterology & hepatology 2007;4(7):393-402.
33. **Gibson SA, McFarlan C, Hay S, et al.** Significance of microflora in proteolysis in the colon. Applied and Environmental Microbiology 1989;55(3):679-683.
34. **Macfarlane GT, Allison C, Gibson SA, et al.** Contribution of the microflora to proteolysis in the human large intestine. J. Appl. Bacteriol. 1988;64(1):37-46.
35. **Tooth D, Garsed K, Singh G, et al.** Characterisation of faecal protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. Gut 2014;63(5):753-760.
36. **Carroll IM, Ringel-Kulka T, Ferrier L, et al.** Fecal Protease Activity Is Associated with Compositional Alterations in the Intestinal Microbiota. PLoS ONE 2013;8(10):e78017.
37. **Pande S, Merker H, Bohl K, et al.** Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. ISME J 2014;8(5):953-962.
38. **Hill MJ.** Intestinal flora and endogenous vitamin synthesis. Eur. J. Cancer. Prev. 1997;6(S43-5).
39. **Rossi M, Amaretti A, and Raimondi S.** Folate production by probiotic bacteria. Nutrients 2011;3(1):118-34.
40. **Magnúsdóttir S, Ravcheev D, de Crécy-Lagard V, et al.** Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. Frontiers in Genetics 2015;6:148.
41. **Sybesma W, Van Den Born E, Starrenburg M, et al.** Controlled modulation of folate polyglutamyl tail length by metabolic engineering of Lactococcus lactis. Appl Environ Microbiol 2003;69(12):7101-7.
42. **de Crecy-Lagard V, El Yacoubi B, de la Garza RD, et al.** Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. BMC Genomics 2007;8:245.
43. **Rodionov DA, Vitreschak AG, Mironov AA, et al.** Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. J Biol Chem 2003;278(42):41148-59.
44. **Degnan PH, Barry NA, Mok KC, et al.** Human gut microbes use multiple transporters to distinguish vitamin B(1) (2) analogs and compete in the gut. Cell Host Microbe 2014;15(1):47-57.
45. **Pryde SE, Duncan SH, Hold GL, et al.** The microbiology of butyrate formation in the human colon. FEMS Microbiol. Lett. 2002;217(2):133.
46. **den Besten G, van Eunen K, Groen AK, et al.** The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. Journal of Lipid Research 2013;54(9):2325-2340.
47. **Hamer HM, Jonkers D, Venema K, et al.** Review article: the role of butyrate on colonic function. Aliment. Pharmacol. Ther. 2008;27(2):104-119.
48. **Gibson PR, Rosella O, Wilson AJ, et al.** Colonic epithelial cell activation and the paradoxical effects of butyrate. Carcinogenesis 1999;20(4):539-44.
49. **Daly K and Shirazi-Beechey SP.** Microarray analysis of butyrate regulated genes in colonic epithelial cells. DNA Cell Biol 2006;25(1):49-62.
50. **Macia L, Thorburn AN, Binge LC, et al.** Microbial influences on epithelial integrity and immune function as a basis for inflammatory diseases. Immunological Reviews 2012;245(1):164-176.
51. **Cani PD, Amar J, Iglesias MA, et al.** Metabolic Endotoxemia Initiates Obesity and Insulin Resistance. Diabetes 2007;56(7):1761-1772.
52. **Qin J, Li Y, Cai Z, et al.** A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55-60.
53. **Archer SY, Meng S, Shei A, et al.** p21WAF1 is required for butyrate-mediated growth inhibition of human colon cancer cells. Proceedings of the National Academy of Sciences 1998;95(12):6791-6796.
54. **McIntyre A, Gibson PR, and Young GP.** Butyrate production from dietary fibre and protection against large bowel cancer in a rat model. Gut 1993;34(3):386-391.
55. **Leschelle X, Delpal S, Goubern M, et al.** Butyrate metabolism upstream and downstream acetyl-CoA synthesis and growth control of human colon carcinoma cells. European Journal of Biochemistry 2000;267(21):6435-6442.
56. **Scott KP, Duncan SH, and Flint HJ.** Dietary fibre and the gut microbiota. Nutrition Bulletin 2008;33(3):201-211.
57. **Louis P and Flint HJ.** Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS Microbiol Lett 2009;294(1):1-8.
58. **Bui TP, Ritari J, Boeren S, et al.** Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. Nat Commun 2015;6:10062.
59. **Vital M, Penton C, Wang Q, et al.** A gene-targeted approach to investigate the intestinal butyrate-producing bacterial community. Microbiome 2013;1(1):8.
60. **Eeckhaut V, Van Immerseel F, Croubels S, et al.** Butyrate production in phylogenetically diverse Firmicutes isolated from the chicken caecum. Microbial Biotechnology 2011;4(4):503-512.

61. **Miller TL and Wolin MJ.** Pathways of acetate, propionate, and butyrate formation by the human fecal microbial flora. Applied and Environmental Microbiology 1996;62(5):1589-1592.
62. **Louis P, Hold GL, and Flint HJ.** The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol 2014;12(10):661-72.
63. **Rey FE, Gonzalez MD, Cheng J, et al.** Metabolic niche of a prominent sulfate-reducing human gut bacterium. Proc Natl Acad Sci U S A 2013;110(33):13582-7.
64. **Wolin MJ and Miller TL.** Interactions of microbial populations in cellulose fermentation. Fed Proc 1983;42(1):109-13.
65. **Segata N, Boernigen D, Tickle TL, et al.** Computational meta'omics for microbial community studies. Mol Syst Biol 2013;9:666.



Chapter 7

Functional characterization of fibronectin-binding domains mined from metagenomic databases

**Sebastian Tims, Victor C.L. de Jager,
Erwin G. Zoetendal, Willem M. de Vos,
Sacha A.F.T. van Hijum, Michiel Kleerebezem**

Manuscript in preparation

Abstract

For a long time the gut microbiota has been recognized for functionally complementing the human host. However, for the majority of the genes in the gut metagenome no function can be predicted to date, and due to the inability to culture all gut microbes the understanding of the functional properties of the microbiota remains incomplete. A powerful approach to improve functional prediction of complete protein sequence is to perform domain searches to identify functional domains within a protein sequence. This study assesses the predictive value of mining assembled metagenomic sequences using a supervised Hidden Markov Model (HMM) mining-method to guide the selection process towards specific domains of functional interest. Physical adherence of microbes to host tissue, as for instance due to the binding of fibronectin by bacteria, establishes first contact in host-microbe interactions. Therefore, fibronectin-binding domains were used as a proof of principal for mining metagenomic gene catalogues and extracting candidate domain sequences that were subjected to functional characterization using an assay based on fibronectin coupled to magnetic beads. Six fibronectin-binding protein families were used as seed for the HMMER mining procedure on the MetaHIT 3.9M gene catalog and this resulted in the identification of 2 – 1,085 metagenomic domains per protein family. For two protein families, PF08533 and PF11966, four domain sequences were selected, cloned, and expressed in *Escherichia coli* BL21-AI. While employing the fibronectin-beads it was shown that all four sequences of the larger PF05833 indeed were positive for fibronectin-binding capacity, but the four sequences of the smaller PF11966 were either unsuccessfully expressed or negative in the binding characterization. This approach, coupled to functional characterization in the laboratory, can specify and expand the function-domain sequence-space and corresponding HMMs to more accurately assign function to genes encoding unknown (or poorly predicted) functions.

Introduction

Already for nearly four decades the gastrointestinal microbiota has been a "prime suspect" in several beneficial and detrimental modulations of host health, by providing important nutritional complements, but also providing a port of entree for pathogens [1]. The inability to culture all microbes from the intestinal tract limits the complete understanding of the functional properties of the microbiota. By the end of the 20th century it was estimated that less than 1% of the bacteria cells in various ecosystems could be cultured [2]. Despite the fact that the relative fraction of the intestinal microbiota that can be cultured is substantially higher (estimated around 10-30%; [3-5]), the largest fraction of the overall population remained uncultured in the laboratory, even after introduction of high throughput and other novel cultivation methods [6]. Hence, cultivation-independent approaches, mostly rooted on DNA based techniques, have been introduced to study the overall microbial community as they capture the composition and genetic repertoire of both "cultured and uncultured" microbes by 16S ribosomal RNA gene profiling and random shotgun metagenomic sequencing of total DNA extracts from a sample, respectively. The latter approach could theoretically provide the complete genetic repertoire of a community, and large scale projects like MetaHIT [7] and the Human Microbiome Project (HMP) [8] have generated large amounts of metagenome sequence data of the human gut microbiota. Following metagenomic sequencing and assembly, the next step is usually functional annotation of the assembled (gene) sequences using sequence identity based approaches and databases with genes of known or predicted function. However, these analyses employ alignment with reference genomes from cultivated microbes, which represent only a fraction of the gastrointestinal inhabitants. As a consequence a substantial fraction of the metagenomic sequences remains unaccounted for, and is assigned to encode proteins of (currently) unknown function [7]. Therefore, next to cataloguing all genes present in a certain ecosystem, the functional characterization of genes and encoded proteins is a requirement for better understanding of microbial ecosystems, and presents an important and challenging task to the scientific community.

To improve the function prediction for metagenomic sequences, more advanced methods such as PSI-BLAST [9] or HMMER (analysis tool using profile hidden Markov models) based domain searches [10] can be employed. Particularly domain searches, that recognize functional domains in partial protein sequences rather than the complete protein sequence provides a powerful approach to achieve function prediction. Based on the functional domains of known enzymes iterative procedures, such as JackHMMer [11], can identify novel homologs and align these to the existing hidden Markov model (HMM) and rebuilt this at each iteration until no more homologs can be found (or a maximum number of iterations is reached). Such procedures enable the recognition of sets of candidate proteins that contain domains with functional properties that are assumed to be similar to the query domain. However these iterative procedures often suffer from false-positive hits, capturing distant proteins that due to sequence divergence do not necessarily encode the predicted functional property. Here we present a supervised method (HMMER based) that builds cluster and domain specific HMMs that can be iterated to identify protein domain families of similar predicted function. This study assesses the predictive value of mining assembled metagenomic sequences using this supervised HMMER method and guiding

the selection process towards specific domains of functional interest, which may expand the function-domain sequence-space and the accuracy of the corresponding HMMs.

One of the most direct forms of host-microbe interactions is achieved by the microbe physically adhering to a variety of membrane-bound and secreted host cell components, such as extracellular matrix (ECM) or mucins [12, 13]. Among the ECM the glycoprotein fibronectin has been identified as the main binding target to host tissue for several bacteria [13-20]. Therefore, we chose the fibronectin-binding domain to assess the predictive qualities of domain targeted mining of metagenomes using the supervised HMMER method.

Materials and methods

Experimental set-up

One of the most direct forms of host-microbe interactions is achieved by the microbe physically adhering to host cells. Successful adherence to host tissue can be achieved by different mechanisms, and can prelude to commensal or pathogenic interactions [21, 22]. Fibronectin is a glycoprotein of approximately 440kD and is a major part of the ECM present in various tissues [23], including the gut where it is vital for tissue repair and survival after intestinal damage [24]. For several bacteria, including both pathogens and probiotic strains, fibronectin has been identified as a main ECM-associated binding target [13-20]. Therefore, fibronectin-binding by bacteria can be regarded as an important potential mechanism of adherence, involved in establishing first contact host-microbe interactions.

Bio-informatics

Amino acid (AA) sequences of characterized fibronectin-binding proteins or domains (see Table 7.2) were obtained from the PFAM database (PFAM26.0; October 2012; [http://PFAM.sanger.ac.uk/, 25]). In order to retrieve genes of interest with similarity to the selected PFAMs, HMM were used as seed for mining the metagenomic DNA sequences from the MetaHIT 3.9M gene catalog [26]. This mining procedure employed HMMER and the complete overall approach is depicted in Figure 7.1 (see SI figure 7.1 for more details on the bio-informatics part).

Quality filtering of the retrieved gut metagenome DNA sequences was performed according the recommendation of the HMMER documentation [27, 28]: the full retrieved sequence should have an E-value < 1 and a full sequence score five-fold higher than the bias (correction term that HMMER applies to the full sequence bit score [28]), while the domain sequence should have an E-value < 1 and a domain score three-fold higher than the domain bias (correction term that HMMER applies to the best-scoring domain sequence bit score [28]). From the retrieved domain sequences, from one of shortest and one of the longest domains, randomly selected domain sequences were further characterized for their fibronectin-binding potential. The taxonomic origin of these selected domain sequences was assessed independently by searching the non-redundant protein sequences (nr) database using BLASTP (protein-protein BLAST [9]). Similarity was assessed by multiple sequence alignment using Clustal Omega [29] (available from the EMBL-EBI bioinformatics analysis tools framework [30]).

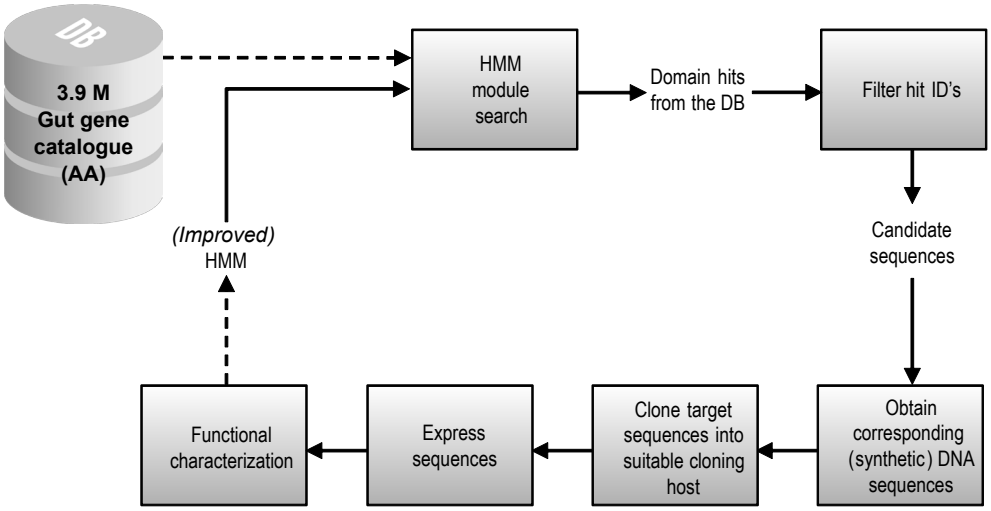


Figure 7.1 Schematic flow-chart of the mining to laboratory characterization procedure. More details on the white boxes, which comprise the developed bioinformatics tool, could be found in SI figure 7.1.

Cloning of the selected domains

After *in silico* introduction of restriction sites for further downstream processing (BamHI restriction site at 5' domain end and HindIII restriction site at the 3' end), the selected domain-encoding sequences (Table S7.1) were codon-optimized for *E.coli* (Table S7.2), synthetized and cloned into pMA or pMK-RQ plasmids by GeneArt (GeneArt®, Life Technologies). The plasmids containing the synthetic gene were transformed into *E.coli* strain MC1061 [31] for plasmid propagation (for a list of plasmids see Table 7.6). The synthetic domains were obtained from these plasmids by BamHI-HindIII digestion, purified from gel with the Zymoclean™ Gel DNA Recovery Kit (Zymo Research), and cloned in the BamHI-HindIII digested expression vector pET24d-AviHisC (Novagen, Darmstadt, Germany), yielding domain expression derivatives designated pET24d-domain-AviHisC (Table 7.6). Subsequently, these domain expression vectors were transformed to *E.coli* strain BL21-AI (Invitrogen) that contains the pBirA plasmid, encoding the BirA biotin ligase [32, 33].

Preparation fibronectin-beads

Fibronectin from human plasma (Sigma) was dissolved in PBS (pH 7.4) to a final concentration of 0.5 mg/mL and subsequently biotinylated with the EZ-link Sulfo-NHS-LC Biotinylation kit (Thermo Scientific), according to the manufacturer's protocol. In short, fibronectin was incubated with a 20-fold molar excess of Sulfo-NHS-LC-Biotin and incubated on ice for 2 hours. The biotinylation mixture was applied three times to a Zebra Spin Desalting Column (Thermo Sciences) to remove excess Sulfo-NHS-LC-Biotin. Biotin binding assay employing the 4'-hydroxyazobenzene-2-carboxylic acid (HABA) reagent was performed to measure the level of biotin incorporation, which was >20 moles of biotin per mole of fibronectin. Fibronectin-beads were produced by activating and coupling Dynabeads® M-280 Streptavidin beads (Invitrogen)

Table 7.1 Strains and plasmids used for cloning in this study.

Strain	Description	Source
E.coli MC1061	F– λ– Δ(ara-leu)7697 [araD139]B/r Δ(codB-lacI)3 galK16 galE15 e14– mcrA0 relA1 rpsL150(StrR) spoT1 mcrB1 hsdR2(r–m+)	[31]
E.coli BL21-AI	F– ompT gal dcm lon hsdSB(rB- mB-) araB::T7RNAP-tetA	Invitrogen
Plasmid	**Description and order of genes (5'-3') ****	**Source**
pBirA	Biotin ligase containing helper plasmid	[32, 33]
pMA plasmids	Propagation plasmids of the synthetic domain sequences	Invitrogen
pMA-Fib1MG1	Fib1MG1, AmpR, Col E1 origin	this study
pMA-Fib1MG2	Fib1MG2, AmpR, Col E1 origin	this study
pMA-Fib1MG3	Fib1MG3, AmpR, Col E1 origin	this study
pMA-Fib1MG4	Fib1MG4, AmpR, Col E1 origin	this study
pMA-Fib1Pos1	Fib1Pos1, AmpR, Col E1 origin	this study
pMA-Fib1Pos2	Fib1Pos1, AmpR, Col E1 origin	this study
pMA-Fib2MG1	Fib2MG1, AmpR, Col E1 origin	this study
pMK-RQ plasmids	Propagation plasmids of the synthetic domain sequences	Invitrogen
pMK-RQ-Fib2MG2	Fib2MG2, KanR, Col E1 origin	this study
pMK-RQ-Fib2MG3	Fib2MG3, KanR, Col E1 origin	this study
pMK-RQ-Fib2MG4	Fib2MG4, KanR, Col E1 origin	this study
pMK-RQ-Fib2Pos1	Fib2Pos1, KanR, Col E1 origin	this study
pMK-RQ-Fib2Pos2	Fib2Pos2, KanR, Col E1 origin	this study
pET24d-AviHisC	T7 RNA polymerase based expression vector, KanR	Novagen
pFib1MG1	Fib1MG1 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib1MG2	Fib1MG2 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib1MG3	Fib1MG3 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib1MG4	Fib1MG4 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib1Pos1	Fib1Pos1 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib1Pos2	Fib1Pos2 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib2MG1	Fib2MG1 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib2MG2	Fib2MG2 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib2MG3	Fib2MG3 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib2MG4	Fib2MG4 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib2Pos1	Fib2Pos1 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study
pFib2Pos2	Fib2Pos2 in pET24d-AviHisC, with Avi-tag and His-tag (C-terminal) *	this study

Fib1 = PF05833-like domain sequence; Fib2 = PF11966-like domain sequence

* for detailed graphical summary, see Figure 7.3

to the biotinylated fibronectin, according to the manufacturer's protocol, resulting in a $1x10^7$ beads/µl PBS solution. A Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) was used to assess the success of the coupling by measuring the protein concentration of the fibronectin solution before and after coupling to the M-280 beads. Bovine serum albumin (BSA)-beads were made by activating and blocking M-280 beads with BSA and these beads were utilized as controls for non-specific binding.

Expression and processing of selected domains

Overnight cultures of the domain expression transformants of BL21-AI (harboring the expression vectors listed in Table 7.6) were inoculated 1/100 into fresh 2xYT medium with 50 μg/ml kanamycin and 10 μg/ml chloramphenicol and were incubated at 37°C while gently shaking. Following three hour incubation, when the cultures had reached an OD600 value of 0.2-0.3, protein expression was induced by isopropyl-β-D-thiogalactopyranoside (IPTG) addition, to a final concentration of 1 mM, L-arabinose to a final concentration of 0.02%, and biotin (from a 10mM tricin buffer pH 8.3) to a final concentration of 50 μM. For each clone a non-induced, negative control culture was taken along to which no IPTG, L-arabinose and biotin were added. After three hour induction, at an OD600 value of 0.8-1.2, the cells were harvested by centrifugation at 4,800 x g, 4°C, for 30 min and pellets were either immediately lysed, or stored at -20°C until further processing. Cell pellets were resuspended in Bacterial Protein Extraction Reagents (B-PER; Thermo Scientific) and the cell suspension was lysed by passing it three times through a French pressure cell (SLM, Aminco, USA) at 110 MPa. Lysates were centrifuged at 4,700 rpm, 4°C, for 30 min and the cleared cell free extract (CFE) was either used directly or stored at 4°C for a maximum of 72 hours prior to their use in fibronectin-binding assays.

Fibronectin binding assay

To 50 μl of CFE, 1 μl of fibronectin-coupled magnetic beads was added and incubated for 10 min at 37°C while gently shaking. Afterwards the fibronectin-beads were separated from the CFE by applying a magnet for 2 min and removing the supernatant, which was discarded. Next, the beads were washed three times for 5 min with 1 volume of PBS (pH 7.4) while gently rotating, each time using a magnet for 2 min to separate the beads from the supernatant. Finally, the fibronectin beads were resuspended in 25 μl of PBS of which 10 μl was mixed with loading buffer (0.1 M sodium phosphate buffer, 10% 2-mercaptoethanol, 4% SDS, 20% glycerol, pH 6.8) and prepared for SDS-PAGE analysis by heating for 10 min at 98°C. SDS-PAGE was performed with 15% or 18% acrylamide gels in a MiniProean III system (BioRad, Hercules, CA, USA). Coomassie brilliant blue staining was applied to visualize the protein bands. As a non-specific binding control, the CFEs of the induced cultures were incubated with BSA-beads. The insoluble fractions of the CFEs were analyzed by SDS-PAGE gels by directly boiling the cell-debris pellets in SDS-PAGE loading buffer and loading on gels.

Results

PFAM selection and metagenomic mining

The PFAM26.0 database [25] contains six protein families with reported fibronectin-binding capacity (Table 7.2). These six PFAMs were used as seed for the HMMER mining procedure on the MetaHIT 3.9M gene catalog [26], which identified a remarkably variable number of metagenome sequences that matched above the quality thresholds depending on the PFAM family (ranging from 2 to 1085 domains; Table 7.2). The number of retrieved metagenomic

sequences did not seem to be related to the size of the original PFAM query sequence employed, e.g. the second largest PFAM (PF07174) identified only 3 metagenomic genes.

Table 7.2 Results of mining a gut microbiota metagenomic dataset with known fibronectin-binding PFAMs.

PFAM accession	Description	PFAM ID	PFAM domain length (AA)	Mining result (domains / gene entries*)
PF02986	Fibronectin binding repeat	Fn_bind	38	2 / 2
PF05833	Fibronectin-binding protein A N-terminus	FbpA	455	1085 / 1063
PF07174	Fibronectin-attachment protein	FAP	298	3 / 3
PF07299	Fibronectin-binding protein	FBP	208	18 / 16
PF08341	Fibronectin-binding signal sequence	Fb_signal	72	4 / 4
PF11966	Fibronectin-binding repeat	SSURE	81	10 / 8

* Metagenomic DNA sequences from the MetaHIT 3.9M gene catalog [26]; the majority of which are predicted to be complete genes.

Metagenomic sequence selection and *in silico* characterization

Out of the three PFAMs that yielded more than five metagenomic domain sequences (Table 7.2) the longest and shortest domains, i.e. PF05833 (Fib1) and PF11966 (Fib2), were chosen to validate their fibronectin-binding capacity. For both Fib1 and Fib2 two randomly chosen domain sequences from the original PFAM alignment, as well as four randomly chosen domain sequences from the candidate metagenomic sequences, were selected for cloning, expression, and characterization of the fibronectin-binding capacity of the protein domains (Table 7.3; full DNA sequences used for DNA synthesis and subsequent cloning, see Table S7.1). The similarity of the selected domain sequences was explored by aligning the amino acid sequences of the domains (Figure 7.2; full amino acid alignment, see Table S7.3), and BLASTP was employed to predict the most likely taxonomic origin of the Fib1 and Fib2 domain sequences, revealing that from the retrieved Fib1 associated domain sequences Fib1MG1 and Fib1MG3 are most similar to the sequences employed to determine the original PFAM (PF05833) (Table 7.3; Figure 7.2A). Conversely, the Fib1-like metagenomic sequences Fib1MG2 and Fib1MG4 were not related to the taxa that are reported to contain the original PFAM (PF05833) (Table 7.3; Figure 7.2B).

Figure 7.2 Neighbor-joining tree (without distance corrections) of the domain sequences selected for fibronectin-binding assay. A) domain sequences retrieved with PF05833. B) domain sequences retrieved with PF11966.

Similar analysis of the Fib2 domain sequences selected indicated that the most retrieved Fib2 associated sequences were most similar to proteins found in streptococci. Only Fib2MG4 (Table 7.3) could not be reliably assigned to any currently known taxonomic origin (Table 7.3; Figure 7.2B).

Table 7.3 Domain sequences selected for fibronectin-binding assay. Entries shaded in gray are control sequences (from PFAM with known fibronectin-binding activity), other entries are candidate sequences of metagenomic origin.

PFAM	Sequence origin (UniProt entry or MetaHIT 3.9M gene catalog entry)	Position (from .. to)	Length (AA)	Name	Top BLASTP taxonomy	(% identity; matches)
Fib1 (PF05833)	P95752_STRGN	4 .. 427	423	Fib1Pos1	*Streptococcus gordonii*	(100%; 423/423)
	Q8KWV4_CLODI	4 .. 460	456	Fib1Pos2	*Clostridium difficile*	(100%; 456/456)
	MC3.MG3.AS1.GP1.C11318.G1	4 .. 326	323	Fib1MG1	*Streptococcus parasanguinis*	(100%; 322/322)
	MC3.MG288.AS1.GP1.C14423.G2	4 .. 461	457	Fib1MG2	Peptostreptococcaceae bacterium VA2	(78%; 403/446)
	MC3.MG290.AS1.GP1.C5166.G2	4 .. 456	452	Fib1MG3	*Clostridium* sp. CAG:7	(93%; 439/452)
	MC3.MG182.AS1.GP1.C1515.G2	4 .. 447	443	Fib1MG4	Firmicutes bacterium CAG:176	(68%; 346/444)
Fib2 (PF11966)	Q97T70_STRPN	667 .. 749	82	Fib2Pos1	*Streptococcus pneumoniae*	(100%; 82/82)
	Q97T70_STRPN	213 .. 293	80	Fib2Pos2	*Streptococcus pneumoniae*	(100%; 80/80)
	MC3.MG353.AS1.GP1.C31364.G4	27 .. 106	79	Fib2MG1	*Streptococcus* sp. SR4	(100%; 79/79)
	MC3.MG353.AS1.GP1.C31364.G4	176 .. 235	60	Fib2MG2	*Streptococcus salivarius* *Streptococcus* sp. SR4	(100%; 60/60) (100%; 60/60)
	MC3.MG3.AS1.GP1.C17136.G1	203 .. 124	79	Fib2MG3	*Streptococcus vestibularis*	(99%; 79/79)
	MC3.MG361.AS1.GP1.C2304.G9	619 .. 665	46	Fib2MG4	*Coprococcus comes*	(44%; 30/45)

Cloning and functional characterization of selected metagenomic domain sequences

The synthetic domains, codon optimized for *E.coli*, were successfully cloned into a pET24d-AviHisC expression vector, where the domain sequence was C-terminally fused to an Avi- and His-tag tandem, under control of the T7 promoter which contains a lac-repressor binding site (the 7-promoter operator sequence lacO). Additionally, this expression vector contains a constitutively expressed lacI gene, encoding the lac-repressor that binds the lacO site in the T7 promoter and thereby represses transcription. Addition of lactose or similar molecule (e.g. IPTG, as used here) can be used to prevent the lac-repressor from binding to the T7 promoter, thereby barring the repression of the T7 promoter and subsequently exposing the cloned recombinant sequence to T7-polymerase driven expression. The utilized expression host, *E.coli* BL21-AI, has a chromosomal insertion of the T7 RNA polymerase gene under control of an araB promoter,

addition of L-arabinose induces the expression of the T7 polymerase from the genome. To achieve expression of the recombinant, domain sequence containing protein both L-arabinose and lactose (or IPTG) are required. For a complete overview of the non-induced and induced states of the expression clones see Figure 7.3.

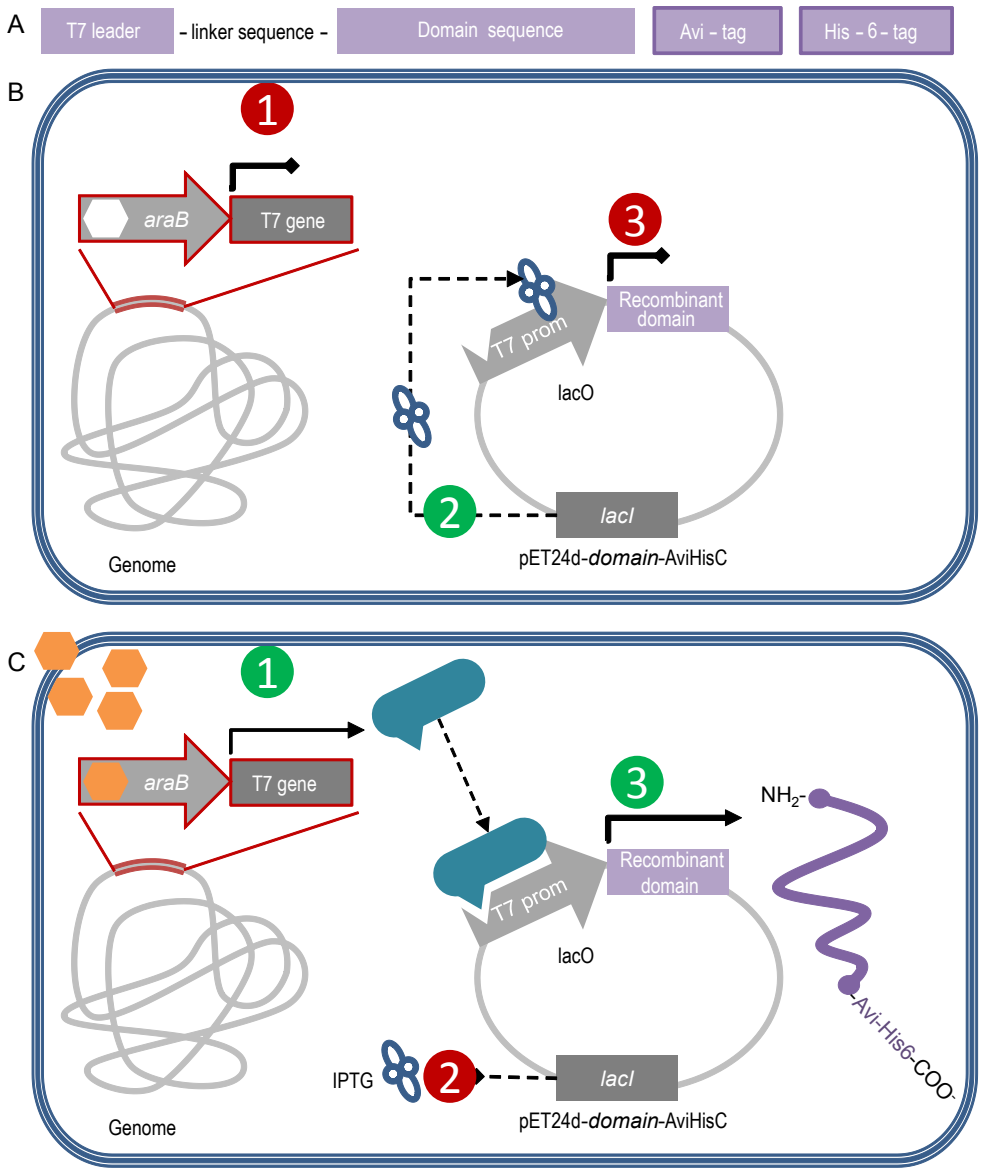


Figure 7.3 Schematic representation of the domain expression system employed. A) General organization of the recombinant domain-sequence to be expressed. B) The BL21-AI host-cell without induction with the cocktail of IPTG, L-arabinose, and exogenously provided biotin. C) The BL21-AI host-cell after induction with IPTG and L-arabinose, in a medium containing biotin.

After a three hour induction (with 1 mM IPTG, 0.02% L-arabinose, and 10 mM biotin), successful expression of the Fib1-type domains was achieved for both positive-control sequences and all four of the metagenomic derived domains as judged by visual inspection of coomassie brilliant blue stained SDS-PAGE gels for the presence of the expected protein bands in the CFEs following IPTG induction, as compared to non-induced conditions (Figure 7.4; SI figure 7.2), indicating that all Fib1-type domains were successfully expressed in *E.coli*. The CFEs were incubated with the fibronectin-beads for 10 min at 37 °C under neutral conditions (PBS, pH 7.4), without the addition of any co-factors. Following standardized washing procedures (PBS washing steps) the expected protein band could be detected in all of the CFEs, revealing that all of the Fib1-type domains can effectively bind to the bead-bound fibronectin (SI figure 7.2). The fibronectin-binding could be shown to be specific by the addition of free fibronectin, indicating that free unbound fibronectin could compete for binding of the Fib1-type domain products (Figure 7.4; SI figure 7.2). Furthermore, none of the Fib1-type domains bound to the control BSA-beads, further supporting the specificity of these domains for fibronectin-binding (Figure 7.4; SI figure 7.2). Taken together these observations indicate that all Fib1-type domains could be expressed in *E. coli*, were shown to specifically and reversibly bind to fibronectin, and therefore confirm and expand this family of fibronectin domains. These results illustrate the potential of the combination of *in silico* and experimental domain-function mining in assembled shotgun metagenomic data to establish and expand domain function predictions.

The IPTG-induced expression of the clones containing the Fib2-type domain expression constructs yielded successful expression for Fib2Pos2, Fib2MG1 and Fib2MG4 as judged by visual inspection of coomassie brilliant blue stained SDS-PAGE gels for the presence of the expected protein bands in the CFEs following IPTG induction compared to non-induced conditions (Figure 7.4; SI figure 7.3). In contrast, induction of Fib2Pos1, Fib2MG2 and Fib2MG3 expression did not yield an additional protein band of the anticipated molecular weight in comparison with the non-induced negative control strain (SI figure 7.3). For the unsuccessfully induced constructs also the insoluble cell-fractions of the induced and non-induced cultures were analyzed by SDS-PAGE, but in these fractions no additional protein bands were detected upon induction either (data not shown). All CFEs of the induced cultures containing the Fib2-expression constructs (irrespective of the detection of a visible protein product migrating at the apparent molecular weight that corresponds to the anticipated Fib2 domain-product) were used in the fibronectin-bead binding assay. The CFEs containing the induced Fib2-type domains were analyzed using the same fibronectin-binding protocol, but no protein could be detected in any of the CFEs (expected molecular weight (MW) of approximately 12 kDa; Figure 7.4; SI figure 7.3). Therefore, for none of the Fib2-like domains it could be shown that they could specifically bind to the bead-associated fibronectin, including those Fib2-like domain-proteins for which a substantial protein band was detected by SDS-PAGE in the CFE (Figure 7.4).

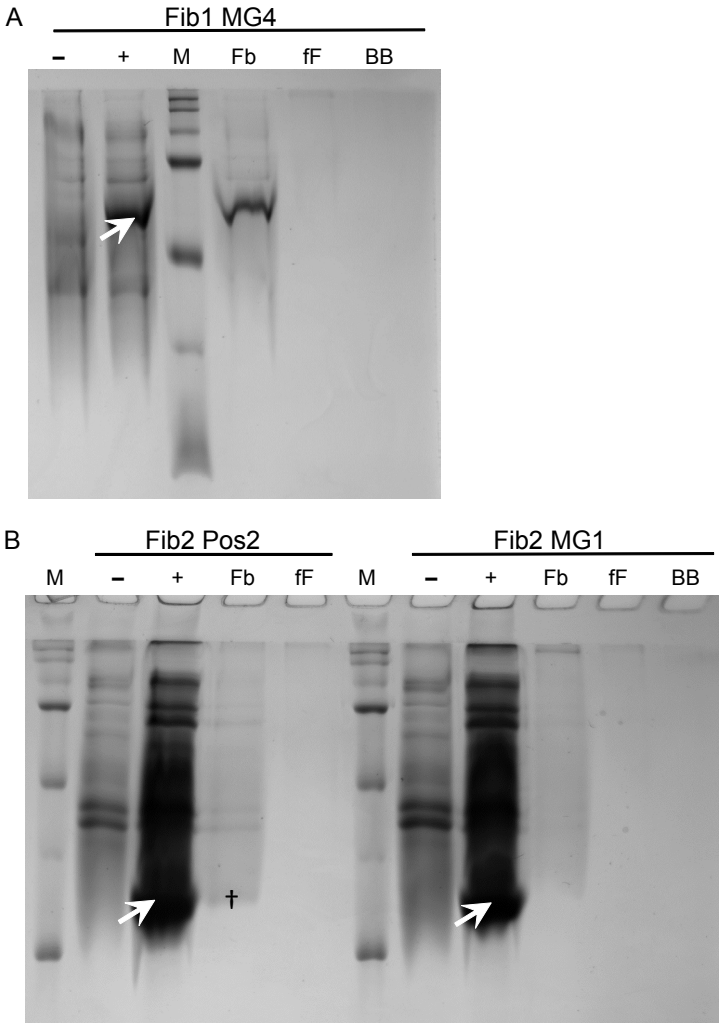


Figure 7.4 A) Example of expression of domain sequences retrieved with PF05833 (A) and PF11966 (B) and the characterization of their fibronectin-binding capability as visualized by SDS-PAGE gels. For PF05833 (A) the domain sequence is one of the randomly chosen metagenomic sequences (Fib1MG4). For PF11966 (B) the domain sequences originate from a random domain sequences from the original PFAM alignment (Fib2Pos2) and a randomly chosen metagenomic sequences (Fib2MG1). White arrows indicate the anticipated apparent molecular weight of the domain-containing recombinant protein constructs (54 kDa for Fib1MG4, and 12 kDa for the Fib2 domains). Lane labels: "–", cell free extract (CFE) of non-induced cultures; "+", CFE of induced cultures; "Fb", fibronectin-bead bound fraction after 10 min incubation of the beads with induced CFE at 37 °C; "fF", fibronectin-bead bound fraction after 10 min incubation at 37 °C with free fibronectin of the Fb-fraction beads; "BB" BSA-bead bound fraction after 10 min incubation of the beads with of induced CFE at 37 °C; "M", Precision Plus Protein Standards (Bio-Rad) displaying bands sizes of 250, 150, 100, 75, 50, 25, and 10 kDa (marker is loaded in the third and sixth lane in panel A, and in the third lane in panel B).

Discussion

In the gut fibronectin-binding is a mechanism to adhere and persist in this host environment employed by pathogenic strains and likely by probiotic strains as well [13-20]. Since microbial adherence to host cells is a direct form of host-microbe interactions that can either result in beneficial or detrimental consequences for the host cells [21, 22], it is relevant to obtain insight in

protein-domain sequences that enable bacteria to bind to fibronectin. We employed HMM-mining to search the available metagenomic gene catalogue of the human intestinal tract microbiota to identify candidate fibronectin domains within the intestinal microbiota metagenome [26]. The PFAM database specified six fibronectin-binding domain families, which displayed a remarkably variation in frequency of detection within the available metagenome database, and allowed the detection of novel candidate fibronectin-binding domain sequences from the metagenomic gene catalogue (Table 7.2). The number of retrieved metagenomic sequences did not seem to be related to the average length of the 'seed' sequences in the original query PFAM, suggesting that the results may reflect actual biodiversity of these PFAM domains within the intestinal gene catalogue. Notably, the second largest PFAM (PF07174) was deduced from the alignment of fibronectin-binding proteins from mycobacteria [34] and still appears to be restricted to this group of organisms. This likely explains the lack of homologous domains in the intestinal metagenome catalogue, since mycobacteria normally do not reside in the human gastrointestinal tract and have only been detected in the gut of severely immune-compromised individuals [35]. The candidate fibronectin-binding domain sequences detected within the metagenomic gene catalogue included sequences that were identical to the query sequences employed to construct the PFAM sequence matrix, but also included nearly unrecognizable sequences that have less than 50% sequence identity with any of the sequences that determine the PFAM. Particularly, this latter group of more distant domain sequences are of interest since these can be employed to (iteratively) expand the sequence matrix for fibronectin-binding domain recognition, provided that one can show that the identified candidate fibronectin-binding domains can specifically recognize and bind to fibronectin. Moreover, this finding illustrates the potential of the HMMER-based approach, since methods to identify such metagenomic genes based on sequence alignment alone (e.g., BLASTP [9]) did not identify these candidate fibronectin-binding proteins (see also below).

Functional analysis was pursued for randomly selected candidate sequences of two fibronectin-binding domains of different sizes, i.e. PF05833 (average length of 455 residues; Fib1) and PF11966 (average length 81 residues; Fib2). Although randomly selected from the mining result, the selected candidate domains were chosen in such a way that they spanned the sequence variability encompassed within the overall mining result, and included very distantly related protein sequences (Table 7.3). Although the protein sequences from which the Fib1 domains were derived did align these proteins to proteins that are annotated as fibronectin-binding proteins, the best hits in these analyses were classified as putative or hypothetical proteins except for the original sequence from which Fib1MG1 originated (Table S7.4). In contrast, the protein sequences from which the Fib2 domains were derived could not be assigned to any function using BLASTP sequence alignment, indicating that sequence alignment approaches fail to recognize the candidate Fib2 domain sequences (best hits in Table S7.4). The capacity of the selected candidate domains to bind to fibronectin was evaluated using a qualitative fibronectin-binding assay that does not provide information concerning binding affinity, but illustrates the "mining-to-binding" approach. This procedure allows the expansion of the sequence diversity matrix of legitimate binding domains and could eventually enable the determination of the sequence-boundaries of the specific domain family. However, with the limited number of candidate

domain-sequence examples studied here, uniform results per domain-type were obtained; (I) all of the larger Fib1-type domains displayed specific and reversible binding to fibronectin, (II) and none of the smaller-sized Fib2-type domains displayed fibronectin-binding, including the positive control domains from established fibronectin-binding proteins [36]. Consequently, our study did not yet reach the boundaries of the sequence flexibility of the Fib1-type domains, and iterative rounds of HMMER mining would be required to identify those sequence boundaries and thereby specify a definite functional Fib1-HMM.

Besides the domain sequence itself, the approach employed also depends on the ability of the cloning host to express the domain, its intrinsic capacity to achieve appropriate folding to display its function, and its capacity to display target binding under the conditions employed in the binding assay. Nonetheless, all Fib1-type domains were successfully expressed and in all cases displayed specific and reversible fibronectin-binding. Our approach in these constructs did not include N-terminal or C-terminal extensions. Especially the lack of N-terminal extensions for this domain family are not expected to interfere with its function, since most of the PF05833 (Fib1) containing proteins (> 99%) have the domain positioned at their N-terminus (Table S7.5; [http://pfam.xfam.org/]). Moreover, our results show that the intrinsic folding of the candidate sequences identified by PFAM-PF05833 is sufficient for functional display of the fibronectin-binding pocket, thereby, clearly contrasting the substantially shorter candidate Fib2-type sequences (see below). The function of the PF08533-domain containing proteins in fibronectin-binding and adhesion to the ECM of *in vitro* cell cultures has been determined for various bacteria, including *Streptococcus pyogenes*, FBP54 [37], *S. gordonii*, FbpA [38], *Clostridium difficile*, Fbp68 [39], and S. *pneumoniae*, PavA [40]. Notably, the N-terminal positioning of PF05833 implies that no signal sequence is present in these proteins, and indeed none of these proteins is predicted to be transported to the bacterial cell envelope on basis of their sequence (Table S7.5). However, most of the functionally confirmed fibronectin-binding proteins of this family were shown to be cell-surface localized, i.e. PavA [40], Fbp68 [39], the N-terminal part of FBP54 [41]. The mechanism by which these bacteria succeed to expose these fibronectin domains (PF05833) to their cell surface remains unknown.

Although fibronectin-binding capacity of genes harboring PF05833 have only been characterized for more pathogen-like species [37-40], this does not exclude commensal or even probiotic strains from exerting this binding strategy. Successful probiotic bacterial strains are thought to encompass the ability to (temporarily) bind the intestinal lining, in order to prevent the attachment of intestinal or food-borne pathogens ([42-44] by competing over the same binding sites [45]. Fibronectin located among the shed ECM has been proposed to be such a binding site where the probiotic bacteria, or other beneficial commensal bacteria, compete with the pathogens [17]. The suspected taxonomic origin of Fib1MG2 and Fib1MG3 are the *Peptostreptococcaceae* family and the *Clostridium* sensu stricto (real clostridia), respectively. With *Clostridium difficile* as the most notorious member, the *Peptostreptococcaceae* family members are often negatively associated with host health [6], suggesting Fib1MG2 comes from a more pathogenic source. Isolates belonging to *Clostridium* sensu stricto are often perceived as pathogens as well, yet real clostridia have been found in considerable amounts in both healthy adults [46] and healthy infants

[47], which does not allow further speculation on the nature of Fib1MG3's species of origin. Interestingly, Fib1MG4 is most closely related to the metagenomic gene group cluster CAG:176, which is suspected to represent (part of a) species [26] and appears to be a relative of *Oscillospira guillermondii* (**Chapter 6**), an uncultured bacteria which appears to be more pronounced in leaner siblings of monozygotic twins discordant for BMI (**Chapter 3**) and in individuals with high bacterial gene richness ([48]; **Chapter 6**). The latter suggests that Fib1MG4 originates from a beneficial commensal and, moreover, gives another clue about the large and morphologically unique *O. guillermondii* which has already been described over a century ago by Chatton and Perard (according to Bergey's Manual of Systematic Bacteriology [49]) but still has not been grown in pure culture.

The limitations of heterologous expression of selected domains is particularly obvious with the Fib2-type constructs, where half of the selected domains appeared not expressed or expressed below detection by SDS-PAGE. There may be a variety of explanations for this observation. Nevertheless, neither of the three successfully expressed Fib2-type domains could be shown to bind to fibronectin, including the positive control domain that is derived from the adherence and virulence factor B (PavB) from S. *pneumoniae* TIGR4 [36, 50], which to the best of our knowledge is the only experimentally verified fibronectin-binding protein with the Fib2-type domain. PavB binds fibronectin in the extracellular matrix (ECM) of the human host cells and contains four Fib2-type repeats [36, 50]. We expressed the Fib2-type domain as a single copy, whereas most (72%) are present in multi-copy tandem-repeats in fibronectin-binding proteins (Table S7.6; [http://pfam.xfam.org/]). The presence of multiple copies of the Fib2-type domain might imply that appropriate folding of the domain depends on its protein context, which could explain why the single copies of this domain that we expressed lacked fibronectin-binding function due to inappropriate folding. However, it has been reported that both multiple and single Fib2-like domains of PavB could be expressed in *E. coli* and could bind fibronectin [36, 50], but the Fib2-domains expressed in these studies included a 70 residue N-terminal extension relative to the minimal domain region. This extension may be critical for its fibronectin-binding capacity and/or the appropriate folding of the domain region [36, 50]. Taken together these results may suggest that the current definition of the PF11966 domain lacks additional, less conserved, N-terminal residues that are required for appropriate domain folding and/or function.

The approach presented here demonstrates that from shotgun metagenomic sequences novel functions can be pinpointed for assembled metagenomic genes. Although, this approach is still targeted towards a recognizable domain sequence within a protein, and will not enable direct detection of completely unknown functions, it allows to verify the current domain families and to expand their sequence biodiversity. Moreover, it may enable the detection and recognition of subfamilies within a domain-family. To this end, iteration of the sequence-mining and function-validation approach that initiates on basis of existing PFAM HMMs, can drive the expansion of functionally validated domain-HMMs. Full analysis of the complete genes in which the domains reside may further refine our understanding of the role of the different types of fibronectin-binding domains in the intestinal tract. The successful implementation of the approach presented here depends on the accuracy of the original PFAM domain definition,

which may explain the lack of success in verification of the fibronectin-binding capacity of Fib2-type domains when expressed out of their native protein context. The approach precisely targets predicted domain functions that are identified by a generic sequence-mining module, while the use of synthetic genes avoids the requirement for appropriate PCR templates, and at the same time allows for codon optimization to facilitate expression in the chosen expression host. The approach is principally driven by sequence-mining, which is fundamentally different from the function-driven random cloning strategies that have been applied before, for example for the identification of screening of novel bile salt hydrolase encoding genes [51] and carbohydrate hydrolases that enable hydrolysis of a specific prebiotic molecule [52]. These random cloning approaches are strongly constrained by the expression of the heterologous genes in the chosen expression host, and require high-throughput functional-screening capacity to identify positive clones. Thereby the function-target oriented approach has specific advantages over the random cloning and function-screening scenario, and the combination or sequential application of random-clone high-throughput screening, with function-driven domain mining could most adequately decipher the functional repertoire encoded within the microbiota of the intestinal tract.

Acknowledgements

We thank Vincent Ruigrok for his excellent technical advice. ST and EGZ received funding from the European Community's Seventh Framework Programme (FP7/2007–2013): MetaHIT, Grant agreement HEALTH-F4-2007-201052, coordinated by S Dusko Ehrlich (Institut National de la Recherche Agronomique, France). Furthermore we would like to thank Tim van Keeke for his efforts to set-up the bioinformatics pipeline and the entire MetaHIT Consortium for the discussions and data sets. MetaHIT Consortium: H Bjørn Nielsen, Mathieu Almeida, Agnieszka S Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R Plichta, Laurent Gautier, Anders G Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B. Quintanilha dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S 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, David W Ussery, Takuji Yamada, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, S Dusko Ehrlich, Alexandre Jamet, Alexandre Mérieux, Antonella Cultrone, Antonio Torrejon, Benoit Quinquis, Christian Brechot, Christine Delorme, Christine M'Rini, Emmanuelle Maguin, Encarna Varela, Eric Guedon, Falony Gwen, Florence Haimet, François Artiguenave, Gaetana Vandemeulebrouck, Gérard Denariaz, Ghalia Khaci, Hervé Blottière, Jan Knol, Jean Weissenbach, Johan E T van Hylckama Vlieg, Jørgensen Torben, Julian Parkhill, Keith Turner, Maarten van de Guchte, Maria Antolin, Maria Rescigno, Muriel Derrien, Nathalie Galleron, Nicolas Sanchez, Niels Grarup, Patrick Veiga, Raish Oozeer, Rozenn Dervyn, Séverine Layec, Thomas Bruls & Yohanan Winogradski.

Supplementary figures and tables

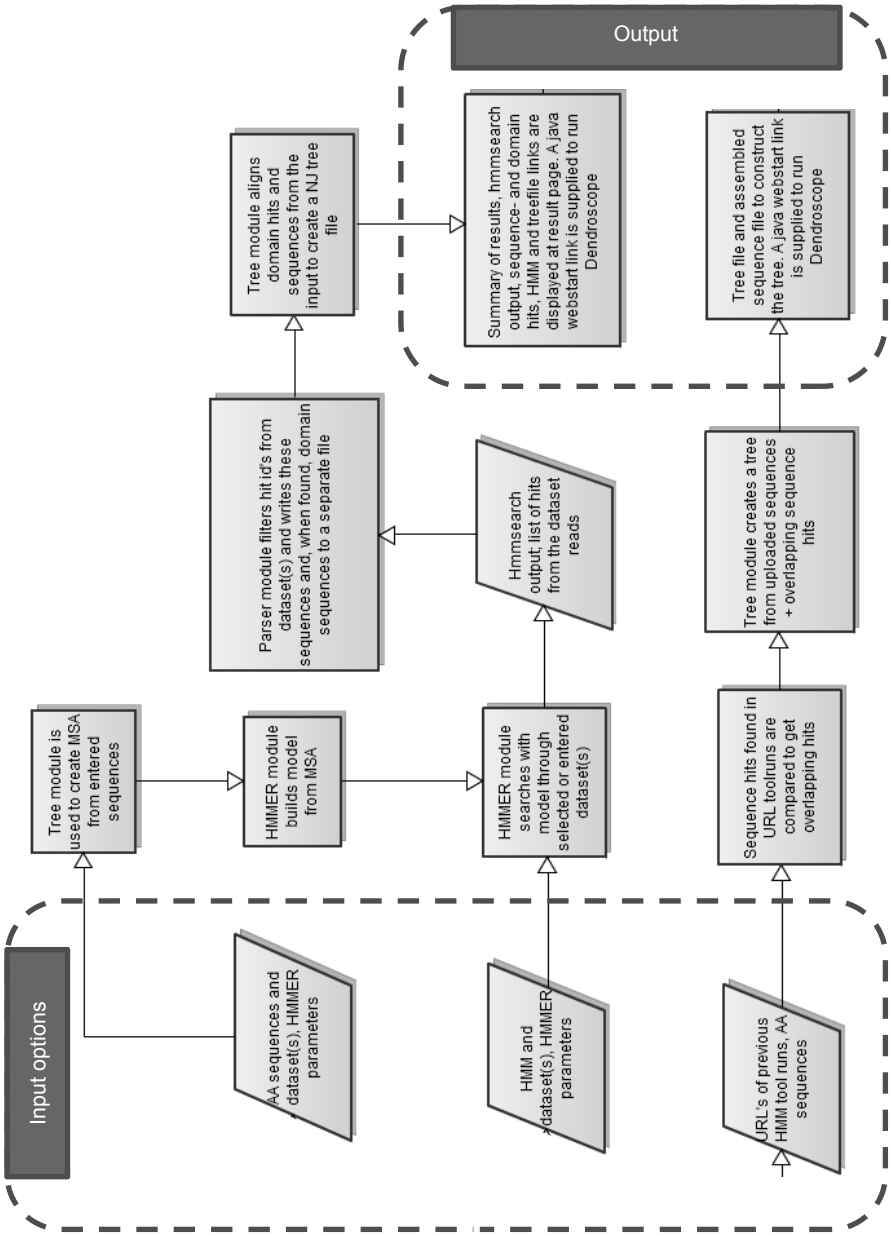


Figure S7.1 Newly developed approach that optimizes the query model based on cluster analysis of the results from the previous round. Interactive selection of a cluster and subsequent searches with cluster specific HMMs guides the user towards novel target domains distant from the initial functional domain.

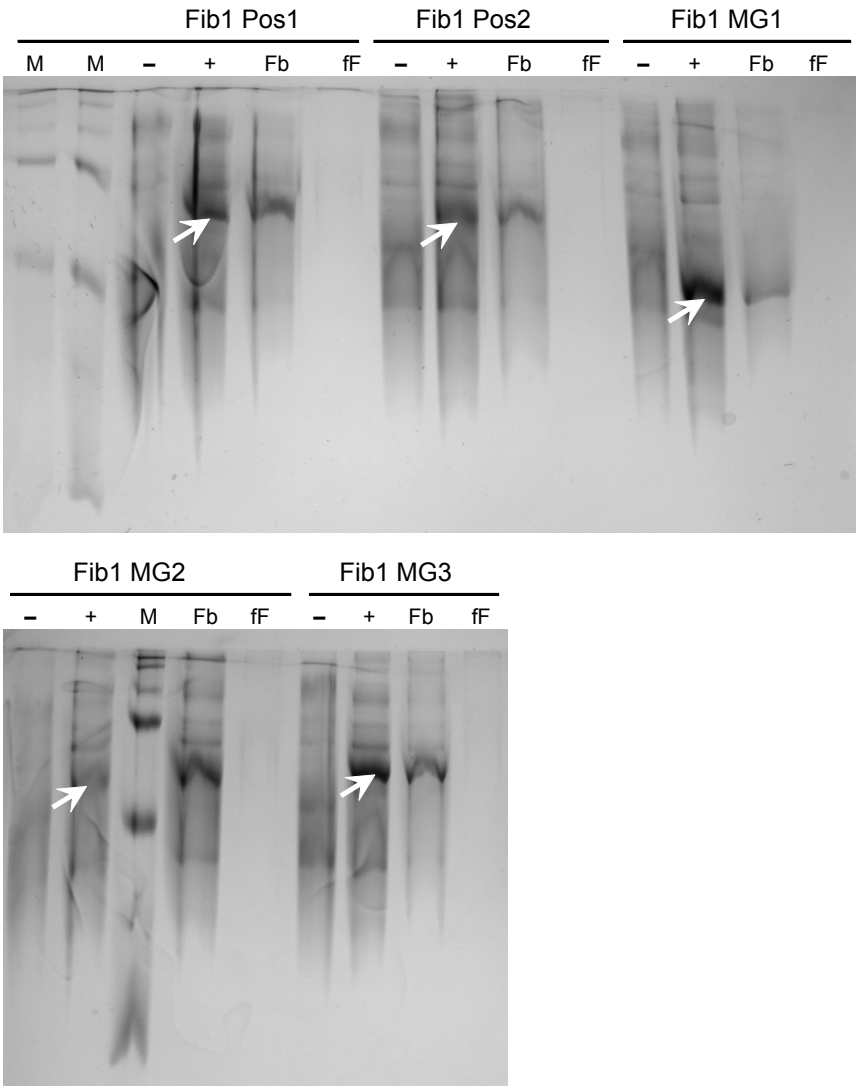


Figure S7.2 Expression of domain sequences retrieved with PF05833 and characterization of their fibronectin binding capability visualized on a 12% SDS-PAGE gel. The domain sequences originate from two random domain sequences from the original PFAM alignment (Fib1Pos1, Fib1Pos2) as well as four randomly chosen metagenomic sequences (Fib1MG1 – Fib1MG4). White arrows indicate the anticipated apparent molecular weight of the domain-containing recombinant protein constructs (40 kDa for the Fib1MG1, and 54 kDa for the other Fib1 domains. Lane labels: "–", cell free extract (CFE) of non-induced cultures; "+", CFE of induced cultures; "Fb", fibronectin-bead bound fraction after 10 min incubation of the beads with induced CFE at 37 °C; "fF", fibronectin-bead bound fraction after 10 min incubation at 37 °C with free fibronectin of the Fb-fraction beads; "BB" BSA-bead bound fraction after 10 min incubation of the beads with of induced CFE at 37 °C; "M", Precision Plus Protein Standards (Bio-Rad) displaying bands sizes of 250, 150, 100, 75, 50, 25, and 10 kDa.

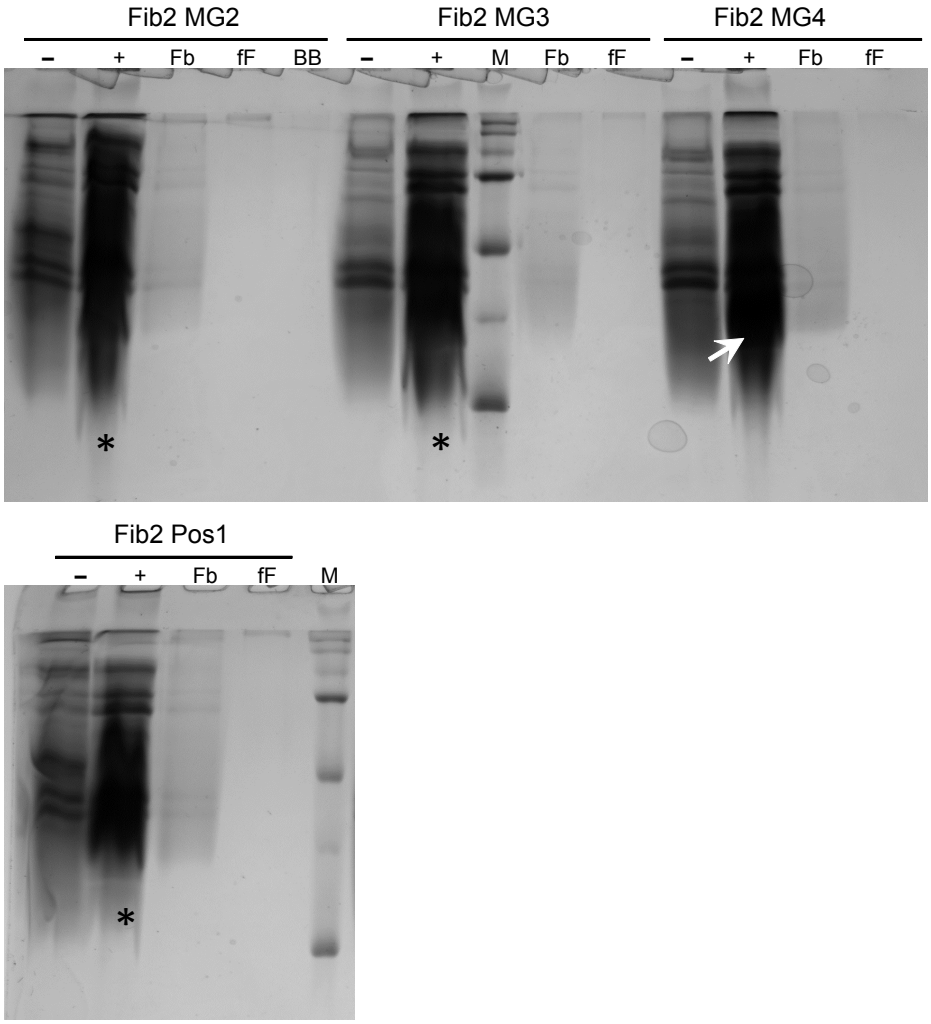


Figure S7.3 Expression of domain sequences retrieved with PF11966 and characterization of their fibronectin binding capability visualized on an 18% SDS-PAGE gel. The domain sequences originate from two random domain sequences from the original PFAM alignment (Fib2Pos1, Fib2Pos2) as well as four randomly chosen metagenomic sequences (Fib2MG1 – Fib2MG4). White arrows indicate the anticipated apparent molecular weight of the domain-containing recombinant protein constructs (11-12 kDa for the Fib2 domains). Asterisk marks positions were the domain protein band was expected but were not detected, indicating either very low level or failed protein expression. Lane labels: "–", cell free extract (CFE) of non-induced cultures; "+", CFE of induced cultures; "Fb", fibronectin-bead bound fraction after 10 min incubation of the beads with induced CFE at 37 °C; "fF", fibronectin-bead bound fraction after 10 min incubation at 37 °C with free fibronectin of the Fb-fraction beads; "BB" BSA-bead bound fraction after 10 min incubation of the beads with of induced CFE at 37 °C; "M", Precision Plus Protein Standards (Bio-Rad) displaying bands sizes of 250, 150, 100, 75, 50, 25, and 10 kDa.

Table S7.1 Selected sequences selected for fibronectin binding assay.

PFAM	Name	Type	DNA sequence
PF05833	Fib1Pos1	Control	TTTGGATCCCT ggctttttctgcatcatatgaccgaagaactgcgccatgaactggtgggcggccgcatt cagaaaattaaccagccgtttgaacaggaactggtgctgcagattcgcagcaaccgcaaa agcctgaaactgctgctgagcgcgcatagcgtgtttggccgcgtgcagctgaccgatacc acctttgaaaacccggcggtgccgaacacctttattatggtgatgcgcaaatatctgcag ggcgcggtgattgaagcgattcagcaggtggaaaacgatcgcattctggaaattagcgtg agcaacaaaaacgaaattggcgatagcgtggcggtgaccctggtgattgaaattatgggc aaacatagcaacattattctgctggataaagcgagcggcaaaattattgaagcgattaaa catgtgggctttagccagaacagctatcgcaccattctgccgggcagcacctatgtggcg ccgccgcagaccggcagcctgaacccgtttaccgtgggcgatgaaaaactgtttgaaatt ctgcagaccgaagaaattgaaccgaaacgcctgctgcagatttttcagggcctgggccgc gataccgcgaccgaactgagcggccgcctgaccaccgatcgcctgaaaacctttcgcgcg ttttttgcgagcccgacccagccgagcctgaccgaaaaaagctttagcgcgctgctgttt agcgatagcaaaacccagatgagcaccctgagcgaactgctggatacctttttataaagat aaagcggaacgctatcgcgtgaaccagcaggcgagcgaactgattcgccgcgtggaaaac gaactggaaaaaaaccgcaaaaaactgggcaaacaggaagatgaactgctggcgaccgaa aaagcggaagaatttcgccagaaaggcgaactgctgaccacctttctgcatcaggtgccg aacgatcaggatcaggtggaactggataactattataccggcgaaaaaattctgattacc ctggataaagcgctgaccccgaaccagaacgcgcagcgctatttttaaacgctatcagaaa ctgaaagaagcggtgaaacatctgaccagcctgattgaagaaacccgcaccaccattctg tatctggaaagcgtggaaaccgcgctggcgcaggcgagcctgaccgaaattgcggaaatt cgcgaagaactgattcagaccggctttattcgccgccgccagcgcgaaaaaattcagaaa cgcaaaaaa AAGCTTGTG
PF05833	Fib1Pos2	Control	TTTGGATCCCT ggcctggtgattcatagcattgtggatgaactgagcagcaaactgaccggcggcaaaatt gataaaattcatcagccggaagatgatgaagtgatttttaacattcgcaacaacaaagaa aactttcgcctggtgctgagcgcgagcgcgagcaacccgcgcgtgtatctgaccagcaac tatcagaaagaaaacccgctgaaagcgccgatgttttgcatgctgctgcgcaaatatatt cagggcggcaacattgtggaaattagccagattggctttgaacgcattattaaaattagc gtggaaagcctggatgaactgaaagaaaaaaccgtgaaaaacattatgattgaaatttatg ggccgccatagcaacattattattacccatggcgaagaaaacaaaattattgatagcatt aaacgcgtgccgtttagcattagccgcgtgcgccaggtgctgccgggccatgattatagc ctgccgccggaacagaacaaactgaacccgctggatgatattagcaaagatctgtttatt aaaaacctggaagaactggaaggcccgatttttaaaagcatttatagccgctttctgggc attagcccgattattgcgaaagaaatttgctatcgcgcgggcgtgaaccagaacgcgatt attaaagatattagcgatgaacagtttgatagcctgcataaagtgttttgcaacctgttt aacgatattaacagcaacaaatatagcccgtgcattattattgataaaaaagtggataaa gtggtggattttagctgcattaacctgaccctgtttagcgatctgagctatattaacaaa gatagcatgagcccgcattctggaagatttttatcgcaccaaagatattaaagatcgcatt aaccagcgcagcagcgatctgaaaaaaagcattagcgtgaaactggatcgcctgtataac aaactgaaaaaacaggaagaagaactgagcgaaagcgaaaacgcggatatttataaaatt aaaggcgaactgattaccagctatatttatatggtggaaaaaggcatggaaagcattgaa gtggcgaacttttatgatgaaaactgcaacgatgtgaccattgaactgaacaaaaacctg accccgagcgaaaacgcgcagaaatattttaaaaaatataacaaaatgaaacatgcgaaa gtggaaattagccatcagattagcctgaacaaagaagaaattgattatctggaaaacatt attctgagcattgaaaactgcgaaaacctggcggaactgcaggatattaaagaagaactg gcgaaagtgggctatattaaaacccagaaaaaaaacagcaaaaaagat AAGCTTGTG
PF05833	Fib1MG1	Metagenomic	TTTGGATCCCT ggctttttctgcatcatatggtgcaggaactgaaagcggaactgctgagcggccgcatt cagaaaattaaccagccgtttgaacaggaactggtgctgcagattcgcggcaaccgcaaa aaccagaaactgctgctgagcgcgcatagcgtgtttggccgcattcagcgcacccagacc aactttgaaaacccggcgtttccgaacacctttattatggtgatgcgcaaatatctgcag ggcgcggtgattgaaggcattgaacagatggaaaacgatcgcattctggaaattcgcgtg agcaacaaaaacgaaattggcgatgcgattagcgtgagcctgatgattgaaattatgggc aaacatagcaacattattctgctggatcgcaccagcaacaaaattattgaagcgattaaa catgtgggctttagccagaacagctatcgcaccattctgccgggcagcacctatattgcg ccgccgaaaaccgatgcggtgaacccgtttaccattggcgatgaagcgctgtttgcgctg ctgcataaagaagaactgagcccgaaaaacctgcagaaatgctttcagggcctgggccgc gataccgcgcaggaactggcgaaacgcctggaaaccgatgaaaaactgaaaacctttcgc gcgtttttgaagcgccgagcgatccgcatctgaccaccaaaagctttagcgcgattccg tttgcggatgcgaccagccagacctttgaaaccctgagcgatctgctggatgattattat cgcgataaagcggaacgcgatcgcgtgcagcagcaggcgagcgaactgattcgcaaagtg gaaaacgatctggaaaaaaaccgcaaaaaactggcgaaacaggaagcggaactggcggcg accgataacgcggaagaatttcgccagaaaggcgaactgctgaccacctttctgcatcag gtgccg AAGCTTGTG

Table S7.1 (continued) Selected sequences selected for fibronectin binding assay.

PFAM	Name	Type	DNA sequence
PF05833	Fib1MG2	Metagenomic	TTTGGATCCCT ggcctggtgattcatagcattgtggatgaactgcataaaaaaactgctgggcggcaaaatt gataaagtgtatcagccggaaaacgatgaagtggtgctgcatattcgcaacaacaaagaa aactttaaactggtgctgagctgcagcgcgagcaaccccgcgcgtgtatctggcgagcgat tataaaaaagaaaacccgattaacgcgccgatgttttgcatgctgtttcgcaaatatatt cagggcggcaacattgtgaacgtgagccaggtggattttgaacgcattattaaaattagc gtggaaagctttgatgaactgaaagaaaaaaccaccaaagatattattattgaaatttatg ggccgccatagcaacattattctgacccatagcagcaacaacaaaattattgatagcgcg aaacgcattccgccgagcgtgagccgcgtgcgccagattctgccgggccagacctatgtg ctgccgccgaaacaggataaactgaacccgattaccgatattagcctgaacagctttgtg gataccctgagcagctttaacggcccgatttttaaagcgatttatagcaaatttctgggc attagcccggtgattgcgaaagaaatttgctttcgcgcgaacattgatgaaaacctgctg gtgagcgaaattagcagcgatgatattagcaaaatttatcgcgaatttcataacctgttt aaatatattaaagataacatttataacccgtgcatggtgattgataccagcattgataaa gtgctggattttagctgcattaacctgagcctgtttagcaacctgagcattattaacgat gatagcattagcaaaattctggaaaactattatgcgaccaaagatattaaagatcgcatt catcagcgcagcagcgatctgcgcaaaagcattagcattaaactggatcgcctgtataac aaactgaacaaacaggaaaaagaactgattgaaagcgaaaacgcggatatttataaaatt aaaggcgaactgattaccagctatatttatatgattgaaaaaggcatggaaagcgtggaa gtggcgaacttttatgatccggaatataaaaacattaaaattagcctgaacaccaacttt accccgagcgaaaacgcgcagaaatattttaaaaaatataacaaactgaaaaccgcgaaa aaagaaattaccagccagatggaaattaccaaagaagaaattgattatctggaaaacatt atgctgagcattgaaaactgcgaaaacctggcggaactgatggatattcgcgaagaactg ggcaaagtgggctatctgcgcagcaaaaacaacagcaaaaaagaaaccaaa AAGCTTGTG
PF05833	Fib1MG3	Metagenomic	TTTGGATCCCT ggcattaccattgcgaacctggtgtgggaatttaaacataccctggaaggcggcaaaatt gcgaaaattgcgcagccggaaaaagatgaactgctgattaccattaaaaacaacaaagaa aactatcgcctgcagattagcgcgagcgcgagcctgccgctgatttatctgaccgcgaac aacaaaaccagcccgctgaccgcgccgaacttttgcatgctgctgcgcaaacatattggc agcgcgcgcattattagcgtgaaacagccgggcctggaacgcattctggaatttgaactg gaacatctggatgaactgggcgatctgtgccgcaaacgcctgattgtggaaattatgggc aaacatagcaacattatttttttgcaaagaagatggcaccattattgatagcattaaacat gtgagcgcgagcatgagcagcgtgcgcgaagtgctgccgggccgcgaatattttattccg cagaccattgcgaaagaaaacccgctggaagtgaccgaagatgtgtttaaaaactgcatt agcaccagcccgaccagcgtgcagaaagcgctgtatggccatctgaccggcattagcccg attattgcggaagaactgtgccatctggcgagcattgatagcgatcgcagcgcgaccgaa ctgaccgaaccggaactgattcatctgtatcataccttcgcctgatgatggaagatgtg aaagatggccatttagcccgagcgtgatttatgatggcgataccccgattgaatatgcg agcgtgccgctgagctgctatgatagcaaaggctattgccgcaaagcgtatgatgcatt agcgcgctgctggaaaactattatgcgagccgcgataccattacccgcattcgccagaaa agcagcgatctgcgccgcattgtgcagaccgcgctggaacgcagctgcaaaaaatatgat ctgcagctgaaacagctgaaagataccgaaaaacgcgaaaaatatcgcatttatggcgaa ctgctgaacacctatggctatgaactgaaaggcggcgaaaaaagctttaaatgcattaac tattatgataacaaagaaattaccattccgctggatccgcagctgaccgcgcgcgaaaac gcgcagaaacattttgataaatataacaaactgaaacgcacctatgaagcgctgagccag ctgaccaaagaaaccaaagcggaagtggatcatctggaaagcgtgagcagcgcgctggat attgcgctggaagaaaacgatctggtgcagattaaagaagaactgatggaatttggctat gtgaaaaaacgccgcgcgaacgaaaaacgcccgaaa AAGCTTGTG
PF05833	Fib1MG4	Metagenomic	TTTGGATCCCT gcgatttgcctgcagggcgtggtgggcgaactggcgccgcagctgaccggcagccgcatt gaaaaaattcagcagccggcgcgcgatcagattattctgctgctgcgcggcagccgccgc ctgtttctgaacgcgggcgcgaaccagccgcgcattcatctgaccgaacagctgcgcgat aacccgagccagccgccgatgttttgcatgctgctgcgcaaacatctgagcggcggcatt attgaaagcgtgcgccaggaaccgctggaacgcgtggtgaccctgaccgtgctggcgagc gatgaaatgggcgaacgcagccgctttaccctggtgtgggaaggcatgccgcgccgcgcg aacctgattctgtgcgatcgcgatggccgcattattgattgcctgcgccgcgtggatctg gaagcggaacaggatcgccaggtgctgccgggcctgttttatcgcctgccgacccgccag gataaacgcagcccgctgagcgtgaccgaagaagaatttgcggcgctgctgggccgcgcg gcgccggatgcgccgctggatgattggctgctggatacctttaccgcgattagcccgctg gtggcgcgcgaactgaccgtgcgcgcgtgcggcagcaccgatgcgccggcgagccagggc aacgcgctgtgggatgtgtttagccgctggcagaaagatgtgaacgaaaacacctttacc ccgaccctgattaaacgcaacggcagcctggcggattttacctatggcctggtgacccag tatggcacctatgcggaaaccgaagtgtatgatagctttagccatctgctggatgatttt tatgaaaaacgcgaacaggcggaacgcgtgaaacagaaaggccgcgatctgctgaaaacc gcgaccaccgcgcgcgatcgcgtgcgccgcaaactggcggcgcaggaaaaagaactggcg gcgtgcctggatcgcgatcatctgcgcatttgcggcgaactgattaccgcgaacctgtat cgcatggaacgcggccagagccgcctgaccgcgcagaactattatgatgaaaactgcgcg gatgtggatattccgctggatgtgcgcctgagcccgcaggaaaacgcggcgcgctatttt aaacagtatgcgaaagcgaaaaccgcggaaaaatatctgaccgcgcagctgcagcgcggc cgcgaagaactgcagtatctggaaagcgtgctgcaggaactggcgcaggcggaaagcgaa caggattttaacgatattcgcaccgaactgaccgatggcggctatctgcgcggccgcggc aaaaaacag AAGCTTGTG

Table S7.1 (continued) Selected sequences selected for fibronectin binding assay.

PFAM	Name	Type	DNA sequence
PF11966	Fib2Pos1	Control	TTTGGATCCCT aacggcgatgcgaaaaacccggcgctgagcccgctgggcgaaaacgtgaaaaccaaaggc cagtatttttatcaggtggcgctggatggcaacgtggcgggcaaagaaaaacaggcgctg attgatcagtttcgcgcgaacggcacccagacctatagcgcgaccgtgaacgtgtatggc aacaaagatggcaaaccggatctggataacattgtggcgaccaaaaaagtgaccattaaa attaac AAGCTTGTG
PF11966	Fib2Pos2	Control	TTTGGATCCCT aacggcaccgcgaaaaacccggcgctgccgccgctggaaggcctgaccaaaggcaaatat ttttatgaagtggatctgaacggcaacaccgtgggcaaacagggccaggcgctgattgat cagctgcgcgcgaacggcacccagacctataaagcgaccgtgaaagtgtatggcaacaaa gatggcaaagcggatctgaccaacctggtggcgaccaaaaacgtggatattaacattaac AAGCTTGTG
PF11966	Fib2MG1	Metagenomic	TTTGGATCCCT aacggcagcgcgaaaaacccggtgctgccgccggtggaaaaactgggcaaaggcctgtat ttttatgaagtggatctggcggatacccagggcaaaagcgataaagaactgctggatctg ctgaaacagaacggcacccagagctataaagcgaccattaaagtgtatggcgcgaaagat ggcaaaccggatctgaccaacctggtggcgaccaaagatctgaccgtgaacctgaac AAGCTTGTG
PF11966	Fib2MG2	Metagenomic	TTTGGATCCCT aacggcatggataaaaacccggcgctgctgccgctggaaggcctggcgaaaggccagtat ttttatgaagtggatctgaacggcaacaccgtgggcaaagatggccaggcgctgctggaa caggtgcgcgcgaacggcacccatacctatctggcgaccgtgaaagtgtatggcgcgaaa AAGCTTGTG
PF11966	Fib2MG3	Metagenomic	TTTGGATCCCT gatagcggcgcgtattttgtggaactgaaactgagcggcaacgatatgggcaaaaaagtg caggtgattgtgaacggcaaaaaatttaaccagagcaacgcgtatagcgtggcggaaatt agcaactatggcgtggaa AAGCTTGTG
PF11966	Fib2MG4	Metagenomic	TTTGGATCCCT aacggcaccgcgaacaacccggcgctgctgccggtggaaggcctggcgaaaggccagtat ttttatgaagtggatctgaacggcaacaccaccggcaaagaaggccaggcgctgctggat cagctgcgcgcgaacggcacccatacctatcaggcgaccgtgaaagtgtatggcagcaaa gatggcaaaccggatctgagcaacctggtggcgacccgccaggtgaccattcgcctg AAGCTTGTG

Table S7.2 For *E.coli* optimized selected fibronectin binding domain sequences.

Domain name	For *E.coli* optimized domain DNA sequence
Fib1MG1	GGTTTTTTTCTGCATCATATGGTGCAAGAACTGAAAGCAGAACTGCTGAG CGGTCGTATTCAGAAAATTAACCAGCCGTTTGAACAAGAACTGGTTCTGC AGATTCGTGGCAATCGTAAAAATCAGAAACTGCTGCTGAGTGCCCATAGC GTTTTTGGTCGCATTCAGCGTACCCAGACCAATTTTGAAAATCCGGCATT TCCGAACACCTTCATTATGGTGATGCGTAAATATCTGCAGGGTGCAGTTA TTGAAGGTATTGAGCAGATGGAAAATGATCGCATTCTGGAAATTCGTGTG AGCAACAAAAACGAAATTGGTGATGCAATTAGCGTGAGCCTGATGATTGA AATTATGGGCAAACATAGCAACATCATTCTGCTGGATCGTACCAGCAATA AAATCATCGAAGCCATTAAACATGTGGGCTTTAGCCAGAATAGCTATCGT ACCATTCTGCCTGGTAGCACCTATATTGCACCGCCTAAAACCGATGCAGT TAATCCGTTTACCATCGGTGATGAAGCACTGTTTGCACTGCTGCATAAAG AAGAACTGAGCCCGAAAAACCTGCAGAAATGTTTTCAGGGTCTGGGTCGT GATACAGCACAAGAGCTGGCAAAACGTCTGGAAACCGATGAAAAACTGAA AACCTTTCGTGCCTTTTTTGAAGCACCGAGCGATCCGCATCTGACCACCA AAAGTTTTAGCGCAATTCCGTTTGCAGATGCAACCAGCCAGACCTTTGAA ACCCTGAGCGATCTGCTGGACGATTATTATCGTGATAAAGCCGAACGTGA TCGTGTGCAGCAGCAGGCAAGCGAACTGATTCGTAAAGTTGAGAATGACC TGGAAAAAAAATCGCAAAAAACTGGCCAAACAAGAAGCCGAACTGGCAGCA ACCGATAATGCAGAAGAATTTCGTCAGAAAGGTGAGCTGCTGACAACCTT TCTGCACCAGGTTCCG
Fib1MG2	GGTCTGGTTATTCATAGCATTGTTGATGAGCTGCACAAAAAACTGCTGGG TGGCAAAATTGATAAAGTGTATCAGCCGGAAAATGATGAAGTGGTTCTGC ATATTCGCAACAACAAAGAGAACTTCAAACTGGTTCTGAGCTGTAGCGCA AGCAATCCGCGTGTTTATCTGGCAAGCGATTACAAAAAAGAAAACCCGAT TAACGCACCGATGTTTTGTATGCTGTTTCGCAAATACATTCAGGGTGGCA ATATCGTTAATGTTAGCCAGGTTGATTTTGAGCGCATCATTAAAATCAGC GTCGAGAGCTTTGATGAACTGAAAGAAAAAACCACCAAAGACATCATCAT CGAAATTATGGGTCGCCATAGCAACATTATTCTGACCCATAGCAGCAATA ACAAAATCATCGATAGCGCCAAACGTATTCCGCCTAGCGTTAGCCGTGTT CGTCAGATTCTGCCTGGTCAGACCTATGTTCTGCCTCCGAAACAGGATAA ACTGAATCCGATTACCGATATTAGCCTGAACAGCTTTGTTGATACCCTGA GCAGCTTTAATGGTCCGATCTTTAAAGCCATCTATAGCAAATTTCTGGGC ATTAGTCCGGTGATCGCAAAAGAAATTTGTTTTCGTGCCAACATCGATGA AAATCTGCTGGTTAGCGAAATCAGCAGTGATGATATCAGCAAAATCTATC GCGAATTCCATAACCTGTTCAAATACATCAAAGACAACATCTACAATCCG TGCATGGTTATTGATACCAGCATCGATAAAGTGCTGGATTTTAGCTGTAT TAACCTGAGCCTGTTTAGCAACCTGAGCATTATTAACGATGACAGCATTA GCAAAATCCTGGAAAACTATTATGCGACCAAAGATATCAAAGACCGTATT CATCAGCGTAGCAGCGATCTGCGTAAATCCATTAGCATTAAACTGGATCG CCTGTATAACAAACTGAACAAACAAGAGAAAGAGCTGATCGAAAGCGAAA ACGCCGATATCTATAAAATCAAAGGCGAACTGATCACCAGCTATATCTAC ATGATTGAGAAAGGTATGGAAAGCGTTGAAGTGGCCAACTTTTATGATCC GGAATACAAAAACATCAAAATCTCCCTGAACACCAATTTTACCCCGAGCG AAAATGCGCAGAAATACTTCAAAAAATACAATAAACTGAAAACCGCGAAA AAAGAAATCACCAGCCAGATGGAAATCACGAAAGAAGAGATTGACTATCT GGAAAACATCATGCTGAGCATCGAAAATTGCGAAAATCTGGCAGAACTGA TGGATATTCGTGAAGAACTGGGTAAAGTTGGTTATCTGCGTAGCAAAAAC AACAGCAAAAAAGAGACAAA
Fib1MG3	GGTATTACCATTGCAAATCTGGTGTGGGAATTCAAACATACCCTGGAAGG TGGTAAAATTGCAAAAATTGCCCAGCCGGAAAAAGATGAACTGCTGATTA CCATCAAAAACAACAAAGAGAATTACCGCCTGCAGATTAGCGCAAGCGCA AGCCTGCCGCTGATTTATCTGACCGCAAATAACAAAACCAGTCCGCTGAC CGCACCGAATTTTTGTATGCTGCTGCGTAAACATATTGGTAGCGCACGTA TTATTAGCGTTAAACAGCCTGGTCTGGAACGCATTCTGGAATTTGAACTG GAACATCTGGATGAACTGGGTGATCTGTGTCGTAAACGTCTGATTGTTGA AATCATGGGCAAACACAGCAACATCATCTTTTGTAAAGAAGATGGCACCA TCATCGACAGCATTAAACATGTTAGCGCCAGCATGAGCAGCGTTCGTGAA GTTCTGCCTGGTCGTGAATATTTCATTCCGCAGACCATTGCCAAAGAAAA TCCGCTGGAAGTTACCGAAGATGTGTTTAAAAACTGTATTAGCACCAGCC CGACCAGCGTTCAGAAAGCACTGTATGGTCATCTGACCGGTATTAGCCCG ATTATTGCCGAAGAACTGTGTCATCTGGCAAGCATTGATAGCGATCGTAG CGCAACCGAACTGACCGAACCGGAACTGATTCATCTGTATCATACCTTTC GTCTGATGATGGAAGATGTGAAAGATGGTCATTTTAGCCCGAGCGTTATT TATGATGGTGATACCCCGATTGAATATGCAAGCGTTCCGCTGAGCTGTTA TGATAGCAAAGGTTATTGTCGCAAAGCCTACGATAGCATTAGCGCACTGC TGGAAAACTATTATGCAAGCCGTGATACCATTACCCGTATTCGTCAGAAA AGCAGCGATCTGCGTCGTATTGTTCAGACCGCACTGGAACGTAGCTGCAA AAAATACGATCTGCAGCTGAAACAACTGAAAGATACCGAAAAACGCGAGA AATATCGCATTTATGGTGAGCTGCTGAATACCTATGGTTATGAACTGAAA GGTGGCGAAAAATCCTTCAAATGCATCAACTACTATGATAACAAAGAAAT CACCATTCCGCTGGACCCGCAGCTGACAGCACGTGAAAATGCACAGAAAC ACTTCGATAAATACAACAAACTGAAACGCACCTATGAAGCACTGAGCCAG CTGACCAAAGAAACCAAAGCAGAAGTTGATCATCTGGAAAGCGTTAGCAG CGCACTGGATATTGCACTGGAAGAAAATGATCTGGTGCAGATCAAAGAAG AACTGATGGAATTTGGCTACGTGAAAAAACGTCGTGCCAATGAAAAACGT CCGAAA

Table S7.2 (continued) For *E.coli* optimized selected fibronectin binding domain sequences.

Domain name	For *E.coli* optimized domain DNA sequence
Fib1MG4	GCAATTTGTCTGCAGGGTGTTGTTGGTGAACTGGCACCGCAGCTGACCGG TAGCCGTATTGAAAAAATTCAGCAGCCTGCACGTGATCAGATTATTCTGC TGCTGCGTGGTAGCCGTCGTCTGTTTCTGAATGCCGGTGCAAATCAGCCT CGTATTCATCTGACCGAACAGCTGCGTGATAATCCGAGCCAGCCTCCGAT GTTTTGTATGCTGCTGCGCAAACATCTGAGCGGTGGTATTATTGAAAGCG TTCGTCAAGAACCGCTGGAACGTGTTGTTACCCTGACCGTTCTGGCAAGT GATGAAATGGGTGAACGTAGCCGTTTTACCCTGGTTTGGGAAGGTATGCC TCGTCGTGCAAATCTGATTCTGTGTGATCGTGATGGTCGCATTATTGATT GTCTGCGTCGTGTTGATCTGGAAGCAGAACAGGATCGTCAGGTTCTGCCT GGTCTGTTTTATCGTCTGCCGACCCGTCAGGATAAACGTAGTCCGCTGAG CGTTACCGAAGAAGAATTTGCAGCACTGCTGGGTCGTGCAGCACCGGATG CACCGCTGGATGATTGGCTGCTGGATACCTTTACCGCAATTTCACCGCTG GTTGCACGTGAACTGACCGTGCGTGCATGTGGTAGTACCGATGCTCCGGC AAGCCAGGGTAATGCACTGTGGGATGTTTTTAGCCGTTGGCAGAAAGATG TGAACGAAAATACATTTACCCCGACCCTGATTAAACGTAATGGTAGCCTG GCAGATTTTACCTATGGTCTGGTTACCCAGTATGGCACCTATGCAGAAAC CGAAGTTTATGATAGCTTTAGCCATCTGCTGGATGACTTTTATGAAAAAC GTGAACAGGCCGAACGTGTGAAACAGAAAAGGTCGTGATCTGCTGAAAACC GCAACCACCGCACGCGATCGTGTTCGTCGTAAACTGGCAGCACAAGAAAA AGAGCTGGCAGCCTGTCTGGATCGTGATCATCTGCGTATTTGCGGTGAAC TGATTACCGCAAATCTGTATCGTATGGAACGTGGTCAGAGCCGTCTGACC GCACAGAACTATTATGATGAAAATTGTGCCGATGTGGATATTCCTCTGGA TGTTCGTCTGAGTCCGCAAGAAAATGCAGCACGTTATTTCAAACAGTACG CAAAAGCAAAAACCGCAGAGAAATATCTGACAGCCCAGCTGCAGCGTGGT CGTGAAGAACTGCAGTATCTGGAAAGCGTGCTGCAAGAACTGGCGCAGGC AGAAAGTGAACAGGATTTTAATGATATTCGTACCGAACTGACGGATGGTG GTTATCTGCGTGGTCGCGGTAAAAAACAG
Fib1Pos1	GGTTTTTTTCTGCATCACATGACCGAAGAACTGCGTCATGAACTGGTTGG TGGTCGTATTCAGAAAATTAACCAGCCGTTTGAACAAGAACTGGTTCTGC AGATTCGTAGCAATCGTAAAAGCCTGAAACTGCTGCTGAGCGCACATAGC GTTTTTGGTCGTGTTCAGCTGACCGATACCACCTTTGAAAATCCGGCAGT TCCGAATACCTTTATTATGGTGATGCGTAAATATCTGCAGGGTGCAGTTA TTGAAGCAATTCAGCAGGTTGAAAATGATCGCATTCTGGAAATTAGCGTG AGCAACAAAAATGAAATTGGTGATAGCGTTGCAGTGACCCTGGTTATTGA AATTATGGGTAAACACAGCAACATCATCCTGCTGGATAAAGCAAGCGGTA AAATCATTGAAGCCATTAAACATGTGGGCTTTAGCCAGAATAGCTATCGT ACCATTCTGCCTGGTAGCACCTATGTTGCACCGCCTCAGACCGGTAGCCT GAATCCGTTTACCGTTGGTGATGAAAAACTGTTTGAAATCCTGCAGACCG AAGAGATTGAACCGAAACGTCTGCTGCAGATTTTTCAGGGTCTGGGTCGT GATACCGCAACCGAACTGAGCGGTCGTCTGACCACCGATCGTCTGAAAAC CTTTCGTGCATTTTTTGCAAGCCCGACCCAGCCGAGCCTGACCGAAAAAA GTTTTAGCGCACTGCTGTTTAGCGATAGCAAAACCCAGATGAGCACCCTG AGCGAACTGCTGGACACCTTCTATAAAGATAAAGCAGAACGTTATCGCGT TAATCAGCAGGCAAGCGAACTGATTCGTCGTGTGGAAAATGAACTGGAAA AAAATCGCAAAAAACTGGGCAAACAAGAAGATGAGCTGCTGGCAACCGAA AAAGCAGAAGAATTTCGTCAGAAAGGTGAACTGCTGACCACATTTCTGCA CCAGGTTCCGAATGATCAGGATCAGGTTGAACTGGATAACTATTATACCG GTGAGAAAATTCTGATCACCCTGGACAAAGCACTGACCCCGAATCAGAAT GCACAGCGTTATTTCAAACGTTATCAGAAACTGAAAGAAGCCGTGAAACA TCTGACCAGCCTGATTGAAGAAACCCGTACCACCATTCTGTATCTGGAAA GCGTTGAAACCGCACTGGCACAGGCAAGCCTGACAGAAATTGCAGAAATT CGTGAAGAACTGATCCAGACCGGTTTTATTCGTCGTCGTCAGCGTGAAAA AATTCAGAAACGCAAAAAA
Fib1Pos2	GGTCTGGTTATTCATAGCATTGTTGATGAACTGAGCAGCAAACTGACCGG TGGTAAAATTGACAAAATTCATCAGCCGGAAGATGACGAAGTGATTTTTA ACATTCGCAACAACAAAGAGAACTTCCGTCTGGTTCTGAGCGCAAGCGCA AGCAATCCGCGTGTTTATCTGACCAGCAATTATCAGAAAGAAAACCCGCT GAAAGCACCGATGTTTTGTATGCTGCTGCGCAAATATATCCAGGGTGGTA ATATTGTTGAAATCAGCCAGATTGGCTTTGAACGCATTATCAAAATTAGC GTGGAAAGCCTGGATGAGCTGAAAGAAAAACCGTGAAAAACATCATGAT CGAGATTATGGGTCGCCATAGCAACATTATTATCACCCATGGTGAAGAGA ACAAAATCATCGATAGCATTAAACGTGTGCCGTTTAGCATTAGCCGTGTT CGTCAGGTTCTGCCTGGTCATGATTATAGCCTGCCTCCGGAACAGAATAA ACTGAATCCGCTGGATGATATTAGCAAAGACCTGTTCATTAAAAACCTGG AAGAACTGGAAGGTCCGATCTTCAAAAGCATTTATAGCCGTTTTCTGGGC ATTAGCCCGATTATTGCAAAAGAAATTTGTTATCGTGCCGGTGTGAATCA GAACGCCATCATTAAAGATATCAGCGACGAACAGTTTGACAGCCTGCATA AAGTTTTTTGCAACCTGTTCAACGACATCAACAGCAACAAATATAGCCCG TGCATTATCATTGATAAAAAAGTGGATAAAGTGGTGGATTTCAGCTGCAT TAATCTGACCCTGTTTAGCGATCTGAGCTACATTAACAAAGATAGCATGA GCCGTATCCTGGAAGATTTTTATCGCACCAAAGATATTAAAGATCGCATT AATCAGCGTAGCAGCGACCTGAAAAAAAGCATTAGCGTTAAACTGGATCG CCTGTATAACAAACTGAAAAAACAAGAAGAAGAACTGAGCGAAAGCGAAA ACGCCGATATCTATAAAATCAAAGGCGAACTGATCACCAGCTACATCTAT ATGGTTGAAAAAGGCATGGAAAGCATTGAAGTGGCCAACTTTTATGATGA GAATTGCAACGATGTGACCATCGAGCTGAACAAAAATCTGACACCGAGCG AAAATGCGCAGAAATACTTCAAAAAATACAACAAAATGAAACACGCCAAA GTGGAAATTAGCCATCAGATTAGCCTGAACAAAGAAGAGATCGATTACCT GGAAAACATTATCCTGAGCATCGAAAATTGTGAAAATCTGGCAGAACTGC AGGACATTAAAGAGGAACTGGCAAAAGTCGGTTATATCAAAACCCAGAAA AAAAACAGCAAAAAAGAT

Table S7.2 (continued) For *E.coli* optimized selected fibronectin binding domain sequences.

Domain name	For *E.coli* optimized domain DNA sequence
Fib2MG1	AATGGTAGCGCAAAAAATCCGGTTCTGCCTCCGGTTGAAAAACTGGGTAA AGGTCTGTATTTCTATGAAGTTGATCTGGCAGATACCCAGGGTAAAAGCG ATAAAGAACTGCTGGATCTGCTGAAACAGAATGGCACCCAGAGCTATAAA GCAACCATTAAAGTTTATGGTGCCAAAGATGGTAAACCGGATCTGACCAA TCTGGTTGCCACCAAAGATCTGACCGTTAATCTGAAT
Fib2MG2	AATGGCATGGATAAAAATCCGGCACTGCTGCCGCTGGAAGGTCTGGCAAA AGGTCAGTATTTCTATGAAGTTGATCTGAATGGTAACACCGTGGGTAAAG ATGGTCAGGCCCTGCTGGAACAGGTTCGTGCAAATGGCACCCATACCTAT CTGGCAACCGTTAAAGTTTATGGTGCCAAA
Fib2MG3	GATAGCGGTGCATATTTTGTTGAACTGAAACTGAGCGGTAACGATATGGG TAAAAAAGTTCAGGTGATTGTGAACGGCAAAAAATTCAATCAGAGCAATG CATATAGCGTGGCCGAAATTAGCAATTATGGTGTTGAA
Fib2MG4	AATGGCACCGCAAATAATCCGGCACTGCTGCCGGTTGAAGGTCTGGCAAA AGGTCAGTATTTCTATGAAGTTGATCTGAATGGTAACACCACCGGTAAAG AAGGTCAGGCCCTGCTGGATCAGCTGCGTGCAAATGGCACCCATACCTAT CAGGCAACCGTTAAAGTTTATGGTAGCAAAGATGGTAAACCGGATCTGAG CAATCTGGTTGCCACCCGTCAGGTGACCATTCGTCTG
Fib2Pos1	AATGGTGATGCAAAAAATCCGGCACTGAGTCCGCTGGGTGAAAATGTTAA AACCAAAGGCCAGTATTTCTATCAGGTTGCACTGGATGGTAATGTTGCCG GTAAAGAAAAACAGGCACTGATTGATCAGTTTCGTGCAAATGGCACCCAG ACCTATAGCGCAACCGTTAATGTTTATGGTAACAAAGATGGTAAACCGGA CCTGGATAACATTGTTGCAACCAAAAAAGTGACCATCAAAATCAAT
Fib2Pos2	AATGGCACCGCAAAAAATCCGGCACTGCCTCCGCTGGAAGGTCTGACCAA AGGTAAATACTTTTATGAGGTGGATCTGAATGGTAACACCGTTGGTAAAC AGGGTCAGGCACTGATTGATCAGCTGCGTGCAAATGGCACCCAGACCTAT AAAGCAACCGTTAAAGTGTATGGCAACAAAGATGGTAAAGCCGATCTGAC CAATCTGGTTGCCACCAAAAATGTGGATATCAACATCAAT

Table S7.3 Clustal Omega (v 1.2.1) multiple sequence alignments of selected domain sequences that were retrieved from a 3.9 M gut gene catalogue [26] with HMMer search employing known fibronectin binding PFAMs.

Multiple sequence alignment of Fib1-type domains (retrieved with PF05833)

```
FIB1MG4      AICLQGVVGELAPQLTGSRIEKIQQPARDQIILLLRGSR---RLFLNAGANQPRIHLTEQ
FIB1POS1     GFFLHHMTEELRHELVGGRIQKINQPFEQELVLQIRSNRKSLKLLLSAHSVFGRVQLTDT
FIB1MG1      GFFLHHMVQELKAELLSGRIQKINQPFEQELVLQIRGNRKNQKLLLSAHSVFGRIQRTQT
FIB1MG3      GITIANLVWEFKHTLEGGKIAKIAQPEKDELLITIKNNKENYRLQISASASLPLIYLTAN
FIB1POS2     GLVIHSIVDELSSKLTGGKIDKIHQPEDDEVIFNIRNNKENFRLVLSASASNPRVYLTSN
FIB1MG2      GLVIHSIVDELHKKLLGGKIDKVYQPENDEVVLHIRNNKENFKLVLSCSASNPRVYLASD
             .: :  :. *:   * ..:* *: **  ::::: :: .:   :* :.. :    :  :

FIB1MG4      L-RDNPSQPPMFCMLLRKHLSGGIIESVRQEPLERVVTLTVLASDEMGERSRFTLVWEGM
FIB1POS1     T-FENPAVPNTFIMVMRKYLQGAVIEAIQQVENDRILEISVSNKNEIGDSVAVTLVIEIM
FIB1MG1      N-FENPAFPNTFIMVMRKYLQGAVIEGIEQMENDRILEIRVSNKNEIGDAISVSLMIEIM
FIB1MG3      N-KTSPLTAPNFCMLLRKHIGSARIISVKQPGLERILEFELEHLDELGDLCRKRLIVEIM
FIB1POS2     YQKENPLKAPMFCMLLRKYIQGGNIVEISQIGFERIIKISVESLDELKEKTVKNIMIEIM
FIB1MG2      YKKENPINAPMFCMLFRKYIQGGNIVNVSQVDFERIIKISVESFDELKEKTTKDIIIEIM
                 .*     * *::**:: .. *  : *   :*:: : :   :*: :     :: * *

FIB1MG4      PRRANLILCD-RDGRIIDCLRRVDLEAEQDRQVLPGLFYRLPTRQDKRSPLSVTEEEFAA
FIB1POS1     GKHSNIILLDKASGKIIEAIKHVGFSQNSYRTILPGSTYVAPPQTGSLNPFTVGDEKLFE
FIB1MG1      GKHSNIILLDRTSNKIIEAIKHVGFSQNSYRTILPGSTYIAPPKTDAVNPFTIGDEALFA
FIB1MG3      GKHSNIIFCK-EDGTIIDSIKHVSASMSSVREVLPGREYFIPQTIAKENPLE-VTEDVFK
FIB1POS2     GRHSNIIITHGEENKIIDSIKRVPFSISRVRQVLPGHDYSLPPEQNKLNPLDDISKDLFI
FIB1MG2      GRHSNIILTHSSNNKIIDSAKRIPPSVSRVRQILPGQTYVLPPKQDKLNPITDISLNSFV
              :::*:*: .  .  **:. :::  . .  * :***  *  *      .*:

FIB1MG4      LLGRAAPDAPLDDWLLDTFTAISPLVARELTVRACGSTDAP-----ASQGNALWDVFSRW
FIB1POS1     ILQTE---EIEPKRLLQIFQGLGRDTATELSGRLTT---------------DRLKTFRAF
FIB1MG1      LLHKE---ELSPKNLQKCFQGLGRDTAQELAKRLETD--------------EKLKTFRAF
FIB1MG3      NCIST-SPTSVQKALYGHLTGISPIIAEELCHLASIDSDRSATELTEPELIHLYHTFRLM
FIB1POS2     KNLEE-LEGPIFKSIYSRFLGISPIIAKEICYRAGVNQNAIIKDISDEQFDSLHKVFCNL
FIB1MG2      DTLSS-FNGPIFKAIYSKFLGISPVIAKEICFRANIDENLLVSEISSDDISKIYREFHNL
                          . :   : .:.   * *:                          *

FIB1MG4      QKDVNENTFTPTLIKRNGS--LADFTYGLVTQYGTYA-ETEVYDSFSHLLDDFYEKREQA
FIB1POS1     FASPTQPSLT-----------EKSFSALLF---SDSK--T-QMSTLSELLDTFYKDKAER
FIB1MG1      FEAPSDPHLT-----------TKSFSAIPF---ADAT--SQTFETLSDLLDDYYRDKAER
FIB1MG3      MEDVKDGHFSPSVIYDGDT--PIEYASVPLSCYDSKGYCRKAYDSISALLENYYASRDTI
FIB1POS2     FNDINSNKYSPCIIIDKKVDKVVDFSCINLTLFSDLSYI--NKDSMSRILEDFYRTKDIK
FIB1MG2      FKYIKDNIYNPCMVIDTSIDKVLDFSCINLSLFSNLSII--NDDSISKILENYYATKDIK
                  ..   .              .::   .            .::* :*: :*  :

FIB1MG4      ERVKQKGRDLLKTATTARDRVRRKLAAQEKELAACLDRDHLRICGELITANLYRMERGQS
FIB1POS1     YRVNQQASELIRRVENELEKNRKKLGKQEDELLATEKAEEFRQKGELLTTFLHQVPNDQD
FIB1MG1      DRVQQQASELIRKVENDLEKNRKKLAKQEAELAATDNAEEFRQKGELLTTFLHQVP----
FIB1MG3      TRIRQKSSDLRRIVQTALERSCKKYDLQLKQLKDTEKREKYRIYGELLNTYGYELKGGEK
FIB1POS2     DRINQRSSDLKKSISVKLDRLYNKLKKQEEELSESENADIYKIKGELITSYIYMVEKGME
FIB1MG2      DRIHQRSSDLRKSISIKLDRLYNKLNKQEKELIESENADIYKIKGELITSYIYMIEKGME
              *:.*:. :* :      ::  .*   *  :*    . :  :  ***:.:  : :

FIB1MG4      RLTAQNYYDENCADVDIPLDVRLSPQENAARYFKQYAKAKTAEKYLTAQLQRGREELQYL
FIB1POS1     QVELDNYYTG--EKILITLDKALTPNQNAQRYFKRYQKLKEAVKHLTSLIEETRTTILYL
FIB1MG1      ------------------------------------------------------------
FIB1MG3      SFKCINYYDN--KEITIPLDPQLTARENAQKHFDKYNKLKRTYEALSQLTKETKAEVDHL
FIB1POS2     SIEVANFYDENCNDVTIELNKNLTPSENAQKYFKKYNKMKHAKVEISHQISLNKEEIDYL
FIB1MG2      SVEVANFYDPEYKNIKISLNTNFTPSENAQKYFKKYNKLKTAKKEITSQMEITKEEIDYL


FIB1MG4      ESVLQELAQAESEQDFNDIRTELTDGGYLRGRGKKQ------
FIB1POS1     ESVETALAQA-SLTEIAEIREELIQTGFIRRRQREKIQKRKK
FIB1MG1      ------------------------------------------
FIB1MG3      ESVSSALDIALEENDLVQIKEELMEFGYVKKRRANEKRPK--
FIB1POS2     ENIILSIENCENLAELQDIKEELAKVGYIKTQKKNSKKD---
FIB1MG2      ENIMLSIENCENLAELMDIREELGKVGYLRSKNNSKKETK--
```

Multiple sequence alignment of Fib1-type domains (retrieved with PF11966)

```
FIB2MG4      -----------------DSGAYFVELKLSGNDMGKKVQVIVNGKKFNQSNAYSVAEISNY
FIB2MG1      NGSAKNPVLPPVEK--LGKGLYFYEVDLAD-TQGKSDKELLDLLKQNGTQSYKA-TIKVY
FIB2POS1     NGDAKNPALSPLGENVKTKGQYFYQVALDGNVAGKEKQALIDQFRANGTQTYSA-TVNVY
FIB2POS2     NGTAKNPALPPLEG--LTKGKYFYEVDLNGNTVGKQGQALIDQLRANGTQTYKA-TVKVY
FIB2MG2      NGMDKNPALLPLEG--LAKGQYFYEVDLNGNTVGKDGQALLEQVRANGTHTYLA-TVKVY
FIB2MG3      NGTANNPALLPVEG--LAKGQYFYEVDLNGNTTGKEGQALLDQLRANGTHTYQA-TVKVY
                              .* ** :: *     **. : :::  : * :.:* .  :. *

FIB2MG4      GVE--------------------
FIB2MG1      GAKDGKPDLTNLVATKDLTVNLN
FIB2POS1     GNKDGKPDLDNIVATKKVTIKIN
FIB2POS2     GNKDGKADLTNLVATKNVDININ
FIB2MG2      GAK--------------------
FIB2MG3      GSKDGKPDLSNLVATRQVTIRL-
             * :
```

Table S7.4 Genes sequences of which the selected metagenomic domain sequences for fibronectin binding assay.

	Description	Max score	Query coverage (%)	E-value	Identity (%)	Accession
Fib1MG1	dihydroorotate dehydrogenase [*Streptococcus parasanguinis*]	674	100	0	100	WP_041818310.1
Fib1MG1	fibronectin-binding domain protein [*Streptococcus parasanguinis* ATCC 15912]	674	100	0	100	AEH55860.1
Fib1MG2	hypothetical protein [*Clostridium dakarense*]	942	99	0	79	WP_042271714.1
Fib1MG3	putative uncharacterized protein [*Clostridium* sp. CAG:7]	1138	100	0	93	CCY41046.1
Fib1MG3	hypothetical protein [[*Clostridium*] clostridioforme]	1006	99	0	81	WP_027643980.1
Fib1MG4	MULTISPECIES: hypothetical protein [*Oscillibacter*]	679	99	0	59	WP_036630417.1
Fib2MG1 & Fib2MG2	cell wall anchor protein [*Streptococcus* sp. SR4]	719	100	0	100	WP_037611629.1
Fib2MG3	YSIRK signal domain/LPXTG anchor domain surface protein [*Streptococcus vestibularis*]	522	100	6E-179	95	WP_003096693.1
Fib2MG4	hypothetical protein [*Coprococcus comes*]	496	94	9E-163	43	WP_008375477.1

Table S7.5 Protein coding sequences with (mainly predicted) FbpA domains (PF05833). Adapted from [http://pfam.xfam.org/].

FbpA copies	Total sequences	Architecture	sequences per architecture
1	1495	FbpA, DUF814	836
		FbpA, DUF814, DUF3441	374
		FbpA	246
		FbpA, DUF3441	12
		FbpA, DUF814, zf-CCHC, DUF3441	12
		HATPase_c, FbpA, Topo-VIb_trans	11
		FbpA, DUF814, DUF3441, RRM_1, RRM_5 x 2	2
		FF x 3, FbpA, DUF814, DUF3441	1
		FbpA, DUF814, DUF3441, Sua5_yciO_yrdC, SUA5	1
2	332	FbpA x 2, DUF814	266
		FbpA x 2, DUF814, zf-CCHC, DUF3441	28
		FbpA x 2, DUF814, DUF3441	23
		FbpA x 2	8
		FbpA x 2, DUF3441	3
		FbpA x 2, DUF814 x 2, zf-CCHC, DUF3441	2
		Clat_adaptor_s, FbpA x 2, DUF814, DUF3441	1
		FbpA x 2, DUF814, DUF3441, Sua5_yciO_yrdC, SUA5	1

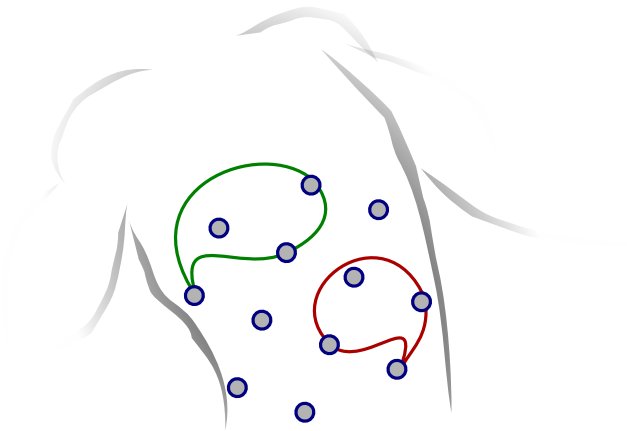
Table S7.6 Protein coding sequences with (mainly predicted) SSURE domains (PF11966). Most architectures contain predicted trans-membrane and/or cell wall anchoring domains, indicating that their subcellular localization is likely on the cell-surface. Architectures shaded in grey are currently without solid trans-membrane or cell wall anchoring domain predictions. Adapted from [http://pfam.xfam.org/].

SSURE copies	Total sequences	Architecture	sequences per architecture
1	213	SSURE	111
		YSIRK_signal, SSURE	96
		SSURE, Gram_pos_anchor	5
		YSIRK_signal, SSURE, Gram_pos_anchor	1
2	360	YSIRK_signal, SSURE x 2	296
		SSURE x 2	56
		YSIRK_signal, SSURE x 2, Gram_pos_anchor	5
		SSURE x 2, Gram_pos_anchor	3
3	112	YSIRK_signal, SSURE x 3	104
		SSURE x 3	7
		YSIRK_signal, SSURE x 3, Gram_pos_anchor	1
4	44	YSIRK_signal, SSURE x 4	27
		YSIRK_signal, SSURE x 4, Gram_pos_anchor	15
		SSURE x 4	2
5	18	YSIRK_signal, SSURE x 5	16
		SSURE x 5	2
6	7	YSIRK_signal, SSURE x 6	6
		SSURE x 6	1

References

1. **Savage DC.** Microbial ecology of the gastrointestinal tract. Ann. Rev. Microbiol. 1977;31:107-133.
2. **Amann RI, Ludwig W, and Schleifer KH.** Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 1995;59(1):143-169.
3. **Suau A, Bonnet R, Sutren M, et al.** Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. Appl. Environ. Microbiol. 1999;65(11):4799-4807.
4. **Sokol H and Seksik P.** The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. Curr Opin Gastroenterol 2010;26(4):327-31.
5. **Tannock GW.** Molecular assessment of intestinal microflora. The American Journal of Clinical Nutrition 2001;73(2):410s-414s.
6. **Rajilic-Stojanovic M and de Vos WM.** The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 2014;38(5):996-1047.
7. **Qin J, Li R, Raes J, et al.** A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59-65.
8. **The Human Microbiome Project C.** A framework for human microbiome research. Nature 2012;486(7402):215-221.
9. **Altschul SF, Madden TL, Schaffer AA, et al.** Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-402.
10. **Krogh A, Brown M, Mian IS, et al.** Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235(5):1501-31.
11. **Johnson LS, Eddy S, and Portugaly E.** Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 2010;11(1):431.
12. **Boesten RJ and de Vos WM.** Interactomics in the human intestine: Lactobacilli and Bifidobacteria make a difference. J Clin Gastroenterol 2008;42 Suppl 3 Pt 2:S163-7.
13. **Buck BL, Altermann E, Svingerud T, et al.** Functional analysis of putative adhesion factors in Lactobacillus acidophilus NCFM. Appl Environ Microbiol 2005;71(12):8344-51.
14. **Bisno AL, Brito MO, and Collins CM.** Molecular basis of group A streptococcal virulence. Lancet Infect Dis 2003;3(4):191-200.
15. **Farfan MJ, Inman KG, and Nataro JP.** The major pilin subunit of the AAF/II fimbriae from enteroaggregative Escherichia coli mediates binding to extracellular matrix proteins. Infect Immun 2008;76(10):4378-84.
16. **Leduc I, White CD, Nepluev I, et al.** Outer membrane protein DsrA is the major fibronectin-binding determinant of Haemophilus ducreyi. Infect Immun 2008;76(4):1608-16.
17. **Lorca G, Torino MI, Font de Valdez G, et al.** Lactobacilli express cell surface proteins which mediate binding of immobilized collagen and fibronectin. FEMS Microbiol Lett 2002;206(1):31-7.
18. **Munoz-Provencio D, Llopis M, Antolin M, et al.** Adhesion properties of Lactobacillus casei strains to resected intestinal fragments and components of the extracellular matrix. Arch Microbiol 2009;191(2):153-61.
19. **Nitsche-Schmitz DP, Rohde M, and Chhatwal GS.** Invasion mechanisms of Gram-positive pathogenic cocci. Thromb Haemost 2007;98(3):488-96.
20. **Schwarz-Linek U, Hook M, and Potts JR.** The molecular basis of fibronectin-mediated bacterial adherence to host cells. Mol Microbiol 2004;52(3):631-41.
21. **Pizarro-Cerda J and Cossart P.** Bacterial adhesion and entry into host cells. Cell 2006;124(4):715-27.
22. **Guarner F and Malagelada J-R.** Gut flora in health and disease. The Lancet 2003;361(9356):512-519.
23. **Bornstein P and Ash JF.** Cell surface-associated structural proteins in connective tissue cells. Proc Natl Acad Sci U S A 1977;74(6):2480-4.
24. **Han SW and Roman J.** Fibronectin induces cell proliferation and inhibits apoptosis in human bronchial epithelial cells: pro-oncogenic effects mediated by PI3-kinase and NF-[kappa]B. Oncogene 2006;25(31):4341-4349.
25. **Finn RD, Bateman A, Clements J, et al.** Pfam: the protein families database. Nucleic Acids Research 2014;42(D1):D222-D230.
26. **Nielsen HB, Almeida M, Juncker AS, et al.** Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 2014;32(8):822-8.
27. **Eddy SR.** Accelerated Profile HMM Searches. PLoS Comput Biol 2011;7(10):e1002195.
28. **Eddy S and Wheeler T.** HMMER User's Guide. Available from: ftp://selab.janelia.org/pub/software/hmmer3/3.1b2/Userguide.pdf.
29. **Sievers F, Wilm A, Dineen D, et al.** Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.
30. **McWilliam H, Li W, Uludag M, et al.** Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res 2013;41(Web Server issue):W597-600.

31. **Casadaban MJ and Cohen SN.** Analysis of gene control signals by DNA fusion and cloning in Escherichia coli. Journal of Molecular Biology 1980;138(2):179-207.
32. **Beckett D, Kovaleva E, and Schatz PJ.** A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Sci 1999;8(4):921-9.
33. **Chakravartty V and Cronan JE.** Altered regulation of Escherichia coli biotin biosynthesis in BirA superrepressor mutant strains. J Bacteriol 2012;194(5):1113-26.
34. **Zhao W, Schorey JS, Groger R, et al.** Characterization of the fibronectin binding motif for a unique mycobacterial fibronectin attachment protein, FAP. J Biol Chem 1999;274(8):4521-6.
35. **Roth RI, Owen RL, Keren DF, et al.** Intestinal infection with Mycobacterium avium in acquired immune deficiency syndrome (AIDS). Histological and clinical comparison with Whipple's disease. Dig Dis Sci 1985;30(5):497-504.
36. **Bumbaca D, Littlejohn JE, Nayakanti H, et al.** Sequence analysis and characterization of a novel fibronectin-binding repeat domain from the surface of Streptococcus pneumoniae. OMICS 2004;8(4):341-56.
37. **Courtney HS, Li Y, Dale JB, et al.** Cloning, sequencing, and expression of a fibronectin/fibrinogen-binding protein from group A streptococci. Infect Immun 1994;62(9):3937-46.
38. **Christie J, McNab R, and Jenkinson HF.** Expression of fibronectin-binding protein FbpA modulates adhesion in Streptococcus gordonii. Microbiology 2002;148(Pt 6):1615-25.
39. **Hennequin C, Janoir C, Barc M-C, et al.** Identification and characterization of a fibronectin-binding protein from Clostridium difficile. Microbiology 2003;149(10):2779-2787.
40. **Holmes AR, McNab R, Millsap KW, et al.** The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence. Molecular Microbiology 2001;41(6):1395-1408.
41. **Courtney HS, Dale JB, and Hasty DI.** Differential effects of the streptococcal fibronectin-binding protein, FBP54, on adhesion of group A streptococci to human buccal cells and HEp-2 tissue culture cells. Infect Immun 1996;64(7):2415-9.
42. **Lee YK, Lim CY, Teng WL, et al.** Quantitative approach in the study of adhesion of lactic acid bacteria to intestinal cells and their competition with enterobacteria. Appl Environ Microbiol 2000;66(9):3692-7.
43. **Mukai T, Asasaka T, Sato E, et al.** Inhibition of binding of Helicobacter pylori to the glycolipid receptors by probiotic Lactobacillus reuteri. FEMS Immunol Med Microbiol 2002;32(2):105-10.
44. **Todoriki K, Mukai T, Sato S, et al.** Inhibition of adhesion of food-borne pathogens to Caco-2 cells by Lactobacillus strains. J Appl Microbiol 2001;91(1):154-9.
45. **Neeser JR, Granato D, Rouvet M, et al.** Lactobacillus johnsonii La1 shares carbohydrate-binding specificities with several enteropathogenic bacteria. Glycobiology 2000;10(11):1193-9.
46. **Finegold SM, Howard RA, and Vera LS.** Effect of diet on human intestinal fecal flora: comparison of Japanese and American diets. Am. J. Clin. Nutr. 1974;27:1456-1469.
47. **Mevissen-Verhage EA, Marcelis JH, de Vos MN, et al.** Bifidobacterium, Bacteroides, and Clostridium spp. in fecal samples from breast-fed and bottle-fed infants with and without iron supplement. J Clin Microbiol 1987;25(2):285-9.
48. **Le Chatelier E, Nielsen T, Qin J, et al.** Richness of human gut microbiome correlates with metabolic markers. Nature 2013;500(7464):541-6.
49. **Garrity GM and Holt JG.** Bergey's Manual of Systematic Bacteriology, 2nd ed. "Taxonomic Outline of the Archaea and Bacteria.". ed. D.R. Boone and R.W. Castenholz. Vol. 1 (The Archaea and the deeply branching and phototrophic Bacteria). 2001. New York: Springer-Verlag. 155-166.
50. **Jensch I, Gamez G, Rothe M, et al.** PavB is a surface-exposed adhesin of Streptococcus pneumoniae contributing to nasopharyngeal colonization and airways infections. Mol Microbiol 2010;77(1):22-43.
51. **Jones BV, Begley M, Hill C, et al.** Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A 2008;105(36):13580-5.
52. **Cecchini DA, Laville E, Laguerre S, et al.** Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. PLoS One 2013;8(9):e72766.



Chapter 8

General Discussion
&
Summary

General discussion

Healthy human adults differ in their microbiota composition as they all have a personalized microbiome. This variability combined with the sheer amount of bacterial species present in healthy humans make the gut microbiota a complex multidimensional entity. Microbiota characterization and biomarker discovery requires high-resolution multidimensional and quantitative analyses, such as phylogenetic microarrays or community-wide DNA sequencing approaches, while as a consequence of the individual diversity large sample sets are commonly required to allow statistically significant observations for hypothesis construction and testing. Different approaches have been explored to address the huge subject-specific variability by stratifying microbial communities based on different types of stable patterns within the microbiota community structures, such as enterotypes as described below [1-3]. Stratification based on these structures could allow for a simplification of the resolution and depth required to determine the microbiota composition, and thereby lower the barriers encountered in development of personalized diets, medication and treatments for every (health related) area where the microbiota plays a role.

Table 8.1 Metagenome data sets generated by the MetaHIT project.

Data set name	Subjects	Comments	Ref.
3.3M catalogue*,I	- 85 Danish subjects (Inter99 cohort [5]) - 39 Spanish subjects adults with IBD	Illumina GA sequences were assembled and 3.3 million unique genes were predicted	2010 [6]
Global Sanger (GS) metagenome S	- 17 publicly available non-European data sets - 22 newly sequenced European subjects	2 of the publicly available sets were performed with pyrosequencing	2011 [1]
Inter99 metagenome I	- 292 Danish subjects	New Illumina GA reads mapping on the 3.3M gene catalogue without *de novo* assembly at the time of publication.	2013 [2]
3.9M catalogue I	- 177 Danish subjects - 141 Spanish subjects	Expansion of 3.3 million gene catalogue with new Illumnia GA reads from additional samples to generate a new non-redundant 3.9 million gene catalogue	2014 [7]

*) first milestone from the MetaHIT project; IBD) inflammatory bowel disease; I) Illumina Genome Analyser (GA) technology used as sequencing method; S) Sanger-sequencing technology (although two of the publicly available sets were performed with pyrosequencing)

This thesis aimed to describe the several factors that are known to shape or modulate the gut microbiota in composition and functioning: host genotype (and early life imprinting) and specific dietary components, such as prebiotics and probiotics. For this purpose, a high-throughput phylogenetic microarray platform (HITChip) was utilized in various trials and cohorts. Additionally, the metagenome data sets and expansion thereof that were generated by the MetaHIT (Metagenomics of the Human Intestinal Tract) project during the course of this thesis were investigated in more detail to gain more insight into the association between the gut microbiota and host energy homeostasis. For clarification of the content of the next sections, these metagenomic data sets are shortly described above (Table 8.1). DNA extracts from independent aliquots of the same fecal samples used in the 3.3M catalogue, GS metagenome (only for the European samples), the Inter99 metagenome, and the 3.9M catalogue, were

gathered and analyzed using the Human Intestinal Tract Chip (HITChip) microarray platform [4]. To complement these studies that addressed the associations between the gut microbiota, host genotype and host energy homeostasis, the development of functional predictions based on data generated from DNA-based profiling approaches was another focal point for this thesis.

Impact of the host genotype

The importance of the host genotype, which is known for about twenty years (reviewed in **Chapter 2**), was further illustrated in **Chapter 3** with a cohort of monozygotic twins were several bacterial groups appear to be consistently related to (structurally conserved) the host genotype. In a recent 16S rRNA gene amplicon sequencing study, which investigated the microbiota of 416 twin pairs from the TwinUK cohort showed that for the majority (63%) of the Operational Taxonomic Units (OTUs) the variation could best be explained by genetics [8]. Especially the members of the *Ruminococcaceae* and *Lachnospiraceae* families showed a high degree of hereditability, whereas the Bacteroidetes members were more associated to environmental factors. These findings match very well with the list of structurally conserved genus-like groups reported in **Chapter 3**. Alongside the influence of the host genotype twin siblings discordant for BMI allows the association between the gut microbiota and host energy homeostasis to be addressed as well. Both the work presented here and the later TwinUK cohort study assessed this association and both studies, although only the latter showed that the members of the novel *Christensenellaceae* family had the highest association to the BMI measurements, reported enrichment of the members of the *Oscillospira* taxon in lean(er) individuals.

Ecosystem structure recognition

The first community-wide stratification method originated from the analysis of global fecal metagenomes by Sanger sequencing and suggested that the human gut microbiota could be categorized in three more prevalent community structure types, designated enterotypes, which appeared to be unrelated to health status, body mass index, age, gender or demography [1]. Multidimensional cluster analysis followed by a between-class analysis had identified the first two enterotypes by a clear enrichment of *Bacteroides* and *Prevotella*, while the third enterotype remained less clearly defined but was mostly characterized by a relative increase in members of the *Ruminococcus* genus. This finding was based on 39 fecal samples from individuals from six different countries, including 22 samples that were newly sequenced and of European origin. Based on this sample set, it could be shown that the driving genera for the discrimination of the three enterotypes also discriminated the phylogenetic microarray (HITChip) datasets [1](Figure 8.1 - inset).

After the introduction of the enterotype concept, popular press and numerous subsequent publications jumped on the idea of distinct microbiota types. It is good to keep in mind that the original publication defined enterotypes as “densely populated areas in a multidimensional space of community composition” which are “not as sharply delimited as, for example, human blood groups”. Indeed most investigators currently favor the concept that the microbiota composition distribution among individuals resembles more a continuum, with a pronounced

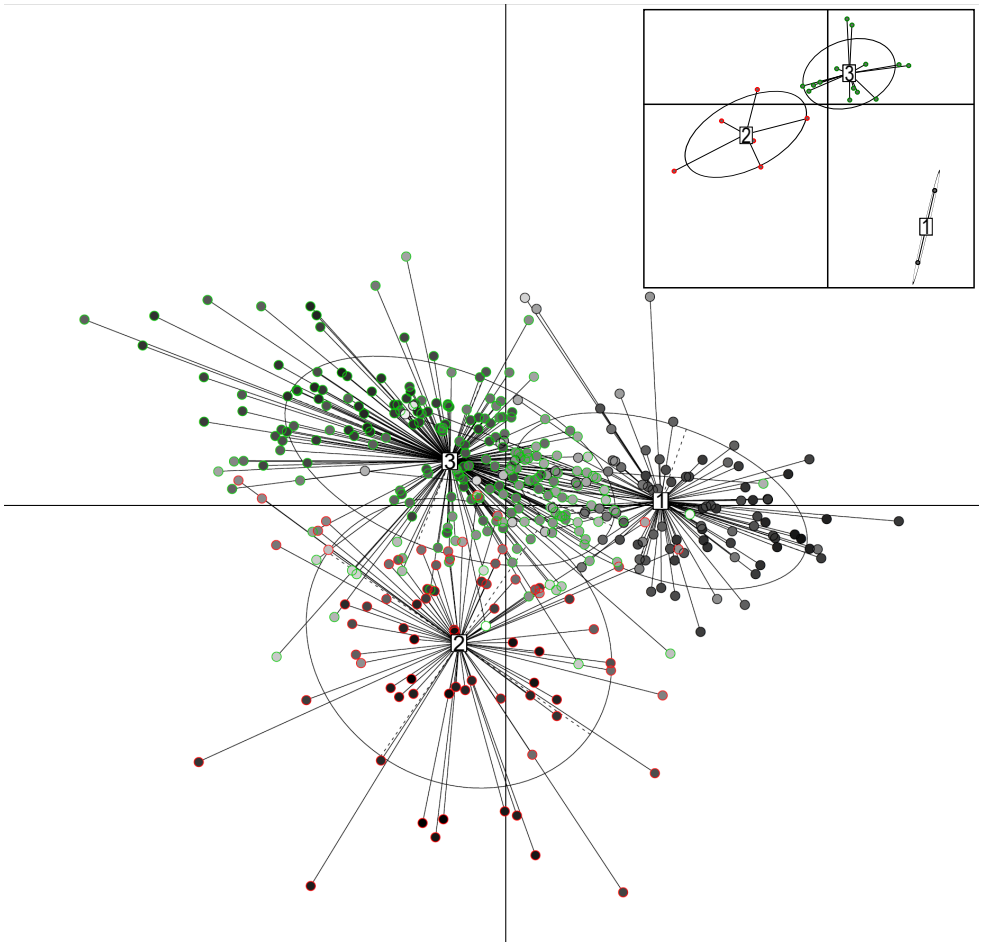


Figure 8.1 Between class analysis, a supervised method which incorporates the outcome of the enterotype clustering of the sequencing data, on HITChip data from 396 European individuals that were analyzed for the 3,9 million gene catalogue by Nielsen and co-workers [7]. Box inset figure top right displays the between class analysis on the HITChip data of 22 European individuals as reported by Arumugam and co-workers [1].

polarity along *Bacteroides* and *Prevotella* gradients [9-12], potentially driven by long-term dietary habits [13]. However, the putative *Ruminococcus*-dominated enterotype, which from the start appeared to have the least clear boundaries, seems much less pronounced [9-11, 13]. Enterotype determination on a more extensive sample set, which included the 396 Danish and Spanish subjects of the 3.9M catalogue [7], did not succeed to clearly discriminate distinct microbiota types. Similarly, an independent HITChip analysis of the same 396 European subjects, following the same approach as in the original publication, did not display clearly distinct microbiota types either although some clustering was observed (Figure 8.1). Moreover, based on the HITChip data of the monozygotic (MZ) twin study (**Chapter 3**) the concept of an imprinted structural core was introduced, which proposes that for several genera in our gut microbiota the ratios of the signals within these genera are conserved by host genotype and/or early life imprinting

[14]. This structural conservation does not necessarily imply that the genera are present at equal relative abundances in each MZ twin pair, since for example the abundance of bacteria related to *Oscillospira guillermondii* is significantly different between twin siblings that have discordant BMIs [14]. Notably, most bacterial groups that were identified as enterotype-drivers are among this imprinted structural core [14], implying that the driving genera for each of the enterotype are conserved and that their abundances may move along a gradient as well. Indeed, in a small scale (n=5) longitudinal study the subjects appeared to switch from one enterotype to another over within a decade [15].

The ability of the gut microbiota to switch between states was investigated in more detail by Lahti and co-workers in a recent large scale HITChip meta-analysis [3]. In this meta-analysis the microbiota composition in samples of more than 1,000 healthy western subjects was assessed and five distinct bacterial groups that display a robust and independent bimodal relative abundance distribution were identified: *Dialister* spp., relatives of *Bacteroides fragilis*, a *Prevotella* subgroup (consisting of the relatives of *Prevotella melaninogenica* and *P. oralis*), and two groups of uncultured *Clostridiales* (UCI and UCII) [3]. The bimodality of these groups implies that they are either present in relatively high abundance levels or in very low abundance, or nearly absent, in the vast majority of samples. Bimodality of these five microbial groups appeared to be independent of (short-term) dietary interventions, but their alternating abundance-scores were associated to several host characteristics such as age and health-related aspects like BMI [3]. Intriguingly, the bimodal groups UCI and UCII were also identified to be part of the bacterial co-occurrence network that was present at lower levels in the heavier sibling of the monozygotic twins discordant for BMI (**Chapter 3**). It was proposed that bi-stable bacterial groups can act like "tipping elements" that have significant diagnostic potential, and may even be considered as "bacterial dipswitches" that could be targeted in future therapeutic approaches [3].

As a side note, analogous to most investigations into robust microbiota markers for host health status, the work by Arumugam and co-workers that resulted in the recognition of the enterotypes did not take into account the possible variations in macro-physiological diversity of human gut functions among individuals [1]. Macro-physiological factors, and in particular gut-transit time, could substantially impact on the microbiota composition and is in most studies either ignored or not assessed due to practical limitations. Recent work by Vandeputte and co-workers used self-reported Bristol Stool Scale classification as a proxy for gut transit time [16], and showed that stool consistency was found to be strongly associated to all frequently used microbiota markers, including species richness, Bacteroidetes to Firmicutes ratios, and enterotypes, emphasizing the importance of gut-transit time as a determinant of microbiota composition. A later independent study in a larger cohort, covering a wide age range, reported strong associations between gut microbiota composition and stool consistency [17]. Even though no relation between bacterial gene richness and stool consistency was observed in this study [17], the results further emphasize the importance of gut-transit with respect to the gut microbiota profile.

Ecosystem richness and diversity

A quantitative metagenomics study by Le Chatelier and co-workers [2] presented another approach for stratification, or ecosystem structure recognition, by utilizing the bacterial richness that seems to be much more accepted as compared to the enterotype concept, probably due to its more obvious relevance in a microbial ecosystem context [18]. Bacterial richness of the human gut microbiome has already been proposed to be an important characteristic in health and disease, and low richness has been associated with inflammatory bowel disorder (IBD) [6, 19, 20], with elevated inflammatory status in elderly individuals [11] and reduced body weight control in obese individuals [21].

It was shown that the bacterial gene numbers in 292 Danish individuals displayed a bimodal distribution, which was confirmed with bacterial probe counts on the HITChip (Figure 8.2). Both methodologies showed a high level of agreement as shown by the high correlation between the logarithm of the gene counts and logarithm of the probe counts (Pearson's r=0.8; Figure 8.2), as well as by the high concordance of 88% that was observed for the classification to high or low richness of the samples [2]. While the abundance of 120,723 bacterial genes significantly differed when subjects were separated on basis of bacterial richness, only 15,894 genes differed when subjects were separated on their BMI. This remarkable result indicates that the gut microbiota of high- and low-gene count individuals is more distant than the gut microbiota of lean and obese individuals. Importantly, obese individuals with a low microbiota gene richness, irrespective of their precise BMI, also had an altered serum composition, including elevated levels of leptin, decreased levels of adiponectin, and displayed decreased insulin resistance, hyperinsulinaemia, increased blood-levels of both free fatty acids and triglycerides, decreased HDL-cholesterol and a more pronounced inflammatory phenotype (as determined by elevated C-reactive protein and higher white blood cell counts), in comparison with obese individuals with high bacterial gene richness [2]. Taken together, this implies that subjects low in bacterial richness are prone to encounter metabolic disturbances that increase their risk for the development of pre-diabetes, type-2-diabetes and ischemic cardiovascular disorders [22, 23]. Consequently, such gut metagenome alterations could be employed to recognize subsets of adults with different metabolic risk profiles, which may contribute to the stratification and resolution of some of the heterogeneity associated with adiposity-related phenotypes. Not included in the published study [2] is the observation that the low gene count subjects were found to be enriched in the *Bacteroides* enterotype, which is in line with the observation that a high animal fat intake is correlating with this enterotype in an US cohort [13].

Earlier studies like the one reporting the first relation between obesity and microbiome [24] did not focus on richness in relation to obesity or host energy homeostasis. However, obesity and high fat high sugar diets have been linked to a low diversity of the gut microbiota [25, 26]. In these studies diversity did not show a strong bimodal distribution like the bacterial gene richness and definitely not as many and strong correlations to the various metabolic markers. In the monozygotic twin cohort discordant for BMI overall microbiota richness could also not be linked to host energy homeostasis either, although the diversity of the members of *Clostridium* cluster IV was found to be lower in the heavier twin siblings (**Chapter 3**). Stratification methods

within obese individuals make sense, since despite their shared excessive body weight they are very different in terms of health status assessment and risk for co-morbidities. Correspondingly, the pathophysiology and therapeutic responsiveness in human individuals suffering from obesity is highly heterogeneous. Interestingly, an almost perfect stratification of subjects with high and low gene-richness can be achieved on basis of only very few bacterial species [2], most of which are detectable on the HITChip and their microarray signals show a high level of correlation to the sequence-based composition profiling. The latter implies that simpler molecular diagnostics targeting these bacterial species may provide a scenario for rapid identification of individuals that are at risk to develop common morbidities. Furthermore, the significance of the observed gene richness variation among individuals is supported by the notion that the gut microbiota appears rather stable over time during adulthood [27], indicating that microbiota richness may very well be an individual-associated characteristic with limited dynamics. Moreover, a dietary intervention resulted in an increased bacterial gene richness and an improved clinical phenotype overall, but was less effective, at least for inflammatory parameters, in subjects classified with low richness prior to the intervention [28]. Notably, the subjects with a low bacterial gene richness did report to consume a diet low in fruits, vegetables, and fish compared to the subjects with a high bacterial gene richness [28], which suggests that richness is affected by long-term dietary habits. Nevertheless, bacterial richness seems to provide some insight into the highly heterogeneous responsiveness of obese individuals with respect to treatment programs.

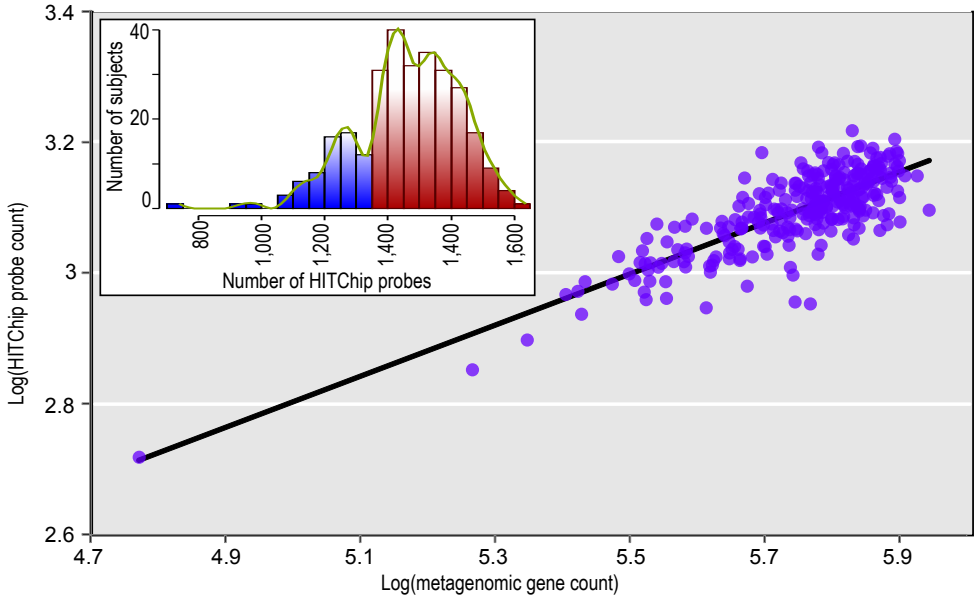


Figure 8.2 Correlation of (logarithmic) metagenomic gene counts and (logarithmic) HITChip probe counts in 292 Danish individuals, Pearson's r=0,8. DNA samples for HITChip and metagenomic analysis came from different aliquots of the same sample. The "low probe count" (blue) and "high probe count" (red) assignment has high concordance of 88.0% with the independent gene count assignment to the low and high bacterial richness groups. Inset: HITChip probe count distribution over 292 Danish individuals.

Subject-specific responses and bacterial co-occurrence networks

The subject specific responsiveness observed in obese individuals, both in terms of metabolic parameters and gut microbiota composition, is not a unique observation in the field. Most studies tend to indicate that there is only a small overlap in the gut microbiota composition between human adults [27, 29, 30]. Hence it is not surprising that human individuals display highly variable metabolic and gut microbiota composition responses to dietary interventions [31-33]. Although most dietary responses appear to be unpredictable, at least with current knowledge, some examples of predictable outcomes are known. For instance, it has been reported, and confirmed in this thesis in **Chapter 5**, that the effect of dietary FOS supplementation in adults is inversely correlated to the *Bifidobacterium* levels at baseline, i.e. individuals with low levels of *Bifidobacterium* show an high increase whereas individuals with high levels of *Bifidobacterium* show an decrease upon FOS consumption [34, 35]. A meta-analysis of three dietary interventions identified several bacterial taxa, mostly Firmicutes, of which the baseline levels could predict the responsiveness of the gut microbiota as well as serum cholesterol measurement following the interventions [36]. Since the gut microbiota composition is often described as stable in time [27, 29, 30], the baseline gut microbiota may be an essential criterion to achieve more successful nutritional and pharmaceutical treatments of metabolic disorders [31, 36, 37].

This thesis described two different dietary interventions, one of which was a supplemented multi-species probiotic in a clinical trial with IBS patients and showed no consistent drastic shifts in the bacterial composition yet showed an altered bacterial co-occurrence network that was unique for the probiotic consuming patients (**Chapter 4**). The other dietary intervention was the above mentioned FOS supplementation in healthy individuals and the consumption of this prebiotic did cause consistent shifts in the bacterial composition but also a strong re-orientation of the bacterial co-occurrence networks in the participants (**Chapter 5**). Moreover, in a recently published study where the diets of African Americans and African were "swapped" for (a mere) 2-weeks and it was shown that the microbiota in both groups displayed vast changes in their co-occurrence network structures and fecal metabolomes [38]. Additionally, these network changes were accompanied by significantly altered levels of mucosal biomarkers of colon cancer risk [38]. These three dietary intervention studies all strongly suggest an individual specific re-alignment of the microbiota was caused by the intervention, i.e. the microbiota composition changes appear minimal when assessed over all participants yet co-occurrence networks display a restructuring which is likely to have consequences for the activity and interactions between microbial groups. In case of the "diet swap study", the actual metabolome was shown to change drastically [38], supporting the host-organismal consequences of microbiota changes elicited by dietary interventions. For the prebiotic FOS consumption study the co-occurrence networks suggest that upon FOS consumption alternative butyrate production pathways becomes more dominant in the gut (**Chapter 5**). The benefit of such a co-occurrence network derived hypothesis was shown in **Chapter 3** where the sample group enriched for a "butyrate producing" network showed higher butyrate levels as compared to the sample group that was enriched for a "primary fiber degrading" network. Both co-occurrence network derived hypotheses (**Chapter 3** and **Chapter 5**) were based on knowledge of cultured representatives of the taxa involved in

the co-occurrence changes, illustrating the potential of extrapolated function approaches for generating new and testable hypotheses.

From structure to function

Extrapolating function based on microbiota composition is often employed in the field. Although such exercises currently yield enough testable hypotheses, there is much more to gain as the majority of the gut metagenome still encodes for unknown genes or represents uncultured and thus functionally non-characterized isolates [6]. The sections above already give two examples with remarkable findings that could still benefit from more knowledge of the unknown genes and genomes. In the study by Le Chatelier and co-workers 120,723 bacterial genes were significantly different in abundance between the subjects with a high and low bacterial richness [2]. Even with the very high resolution provided by this metagenome approach, the vast majority (>90%) of these differentially abundant genes could not be assigned to a known bacterial species with sequenced genomes. Similarly in the large meta-analysis by Lahti and co-workers the species-like groups corresponding to the genus-like groups displaying the most prominent bimodal distribution represented uncultured bacteria [3]. The potential use of these "tipping element" bacterial groups in diagnostics and therapy underpins the importance of their functional characterization to further our understanding of the possible causal relationship between the bimodal, alternate abundance states and host physiology. Moreover, the functional characterization of these (uncultured) species-like groups may also unravel the mechanism involved in their tipping-point status, and their conditional persistence and abundance in the gut system. A logical starting point for the genetic characterization of these groups may be found in mining gut metagenome datasets, which also contains genetic information of the uncultured isolates.

Two different strategies to improve the knowledge gain from uncultured microbes and/or uncharacterized genes were presented in **Chapter 6** and **Chapter 7**. Combining metagenomic-species (MGS), which are defined as co-abundance gene groups with more than 700 genes, from metagenomic data with 16S rRNA gene abundance profiling as presented in **Chapter 6** allowed assignment of 16S rRNA gene IDs on MGS and subsequently investigate the genetic potential of the genes within the MGS that correspond to 16S rRNA gene sequences of uncultured bacteria. The 13 most highly correlated MGS with an occurrence pattern among studied subjects that suggests their potential relevance in the context of host-health status, were further analyzed in **Chapter 6**. Interestingly these 13 MGS, assigned to uncultured isolates, are related to seven genus-like groups, five of which are part of the "primary fiber degrading" co-occurrence network identified in **Chapter 3**. The latter is perfectly in line with the general view that saccharolytic fermentation processes are an important core activity of a normally functioning gut microbiota. Despite these cultivation independent insgihts, *in vitro* culturing of representative isolates of the gut microbiota still remains a crucial step for functional characterization. However, the combinatory approach presented in **Chapter 6** could eventually increase the success rate of culturing approaches by predicting, or excluding, nutritional and environmental requirements of yet uncultured gut bacteria. Besides the large number of uncultured microbes that are present in any ecosystem, another lack in our current understanding of ecosystems derives from the

fact that the vast majority of identified genes have predicted functions based on comparative sequence analyses with genes encoding validated functions. When the cultured isolates are not available functional metagenomics provides currently the only cultivation circumvention that allows the functional characterization of the protein encoded by an unknown DNA sequence. The approach presented in **Chapter 7** provides a more direct functional characterization method for gene or domain sequences from (meta)genomic gene sets, making it a variation of the classic functional metagenomics approaches in which huge libraries are screened for functional properties of interest.

Functional metagenomics

Currently there are two options for functional characterization of metagenomes available, i.e. functional screening of all clones from metagenomic expression libraries or employing (DNA) enrichment strategies before or during the expression library (for more detail see **Chapter 1**). The work presented in this thesis offers a third option: targeted cloning on basis of metagenome sequence mining in conjunction with an analytical characterization protocol (**Chapter 7**). This new approach appears to be a valid approach for selection and characterization of metagenomic sequences without functional annotation, with an added advantage of being precisely targeted towards specific domains of interest. In three of the four metagenomic Fib1-type domains (i.e. the fibronectin-binding domain PF08533, FbpA) the best hits from a BLASTP [39] analysis may only yield putative proteins and automated prediction methods may fail, but the fibronectin binding would still be a possible prediction when (manually) taking more hits into account. Although completely novel functional genes or domains are unlikely to be uncovered by this approach, it can still be expected to allow the expansion of the sequence space associated with current domain families and could fuel the characterization of subfamilies within current domains. What remains unanswered is where the limit of this approach is: "how distant from the known sequences can this approach go?". Answers to this question are currently strictly speculative, and are probably varying per protein domain.

A variation of the approach presented in **Chapter 7** could be applied to focus on specific functionality, independent of an *a priori* domain selection, and potentially unique for an ecologic nice of interest. To mine for functions specific for the gut ecosystem, a metagenomic gene catalogue can first be filtered for genes that are unique or at least overrepresented in the gut ecosystem as compared to other ecosystems. Such a filtering was done on the first MetaHIT gene catalogue of 3.3 M genes by Qin and co-workers and yielded 150,475 rare core gut genes [6]. From this gene set approximately 67% were of unknown function as assessed with an orthologous group analysis performed with the eggNOG system [40]. These gut specific genes with unknown functionality were subsequently subjected to a LocateP [41] and an InterProScan analysis [42]. This resulted in a set of 1,925 genes encoding potentially surface expressed or secreted proteins that, due to one or more sequence signatures, were assigned to the following predicted functional classes: carbohydrate binding and/or breakdown (220 genes), protein binding and/or breakdown (266 genes), antibiotic/toxin resistance or production (202 genes), host-interaction (73 genes),

transporters (66 genes), lipid metabolism (94 genes), other (1,004 genes; oxi-reductases/ sporulation/transformation/signaling).

In contrast to the approach presented in **Chapter 7** the functionality of the obtained sequences per predicted functional class are likely much broader in scope, e.g. the carbohydrate binding genes are likely to bind various different type of carbohydrates. The carbohydrate binding and/or breakdown class was chosen to further investigate the use of the functional metagenome mining approach. From a carbohydrate perspective, diet-derived substrates contain complex carbohydrates and are depleted of simple sugars in the colon. In addition, the mucus barrier consisting of complex carbohydrate chains substituted on protein backbones is continuously produced by the host and after microbial degradation can serve as a source of carbohydrate substrates for growth for the gut microbiota. Consequently, carbohydrate binding and degradation is commonly considered to be a crucial capacity for microbes that persist and function in the colon. Whole-cell binding assays could allow identification, sorting and subsequent characterization of metagenomic clones containing bacterial genes encoding proteins which are involved in attachment to carbohydrates specific for the gastrointestinal tract.

Analogously to the approach presented in **Chapter 7** the genes encoding potential carbohydrate binding domains were synthetized with codon optimization for *E.coli*, and cloned in the pET24d-domain-AviHisC expression system. Following recombinant protein expression, cell-free extracts (CFE) were prepared as described in **Chapter 7**, and analysed by SDS-PAGE, revealing that four of the eight domains that were cloned were successfully produced to a detectable level. The recombinant protein containing CFEs of these clones were passed through a 0.45 µm filter and subsequently, utilizing the introduced His-tag, were purified with using immobilized metal-ion affinity chromatography (HIS-Select® Nickel affinity gel, Sigma-Aldrich). To identify the carbohydrate that could potentially be recognized by the expressed domains, the purified recombinant proteins were hybridized to the neoglycoconjugate (NGC) microarrays A and B, using standard conditions for binding and washing of the array (collaboration with the Glycoscience Group from the National University of Ireland in Galway). Fluorescently-labeled anti-HisTAG antibodies (monoclonal anti-6X His IgG-CF640R; Sigma SAB4600167) were employed to detect binding of the recombinant protein to specific carbohydrates displayed on the NGC microarrays, revealing that only at very high concentrations of the recombinant protein, low-intensity binding to Gal-b-BSA could be detected for three of the recombinant proteins, which was likely non-specific, reflecting an artifact of array overloading. Moreover, the non-specific binding to linker controls (4AP-BSA and 4AP-HSA) for all samples tested indicated that non-specific binding for these samples was a significant consideration in these experiments. The lack of significant and specific binding to the saccharides displayed by the array may also indicate that the appropriate ligand is not represented in the saccharide panel on the NGC arrays. Nevertheless, the low intensity binding to Gal-b-BSA observed for one of the recombinant proteins (SG2-4) could be effectively inhibited by the addition of 100 mM Galactose to the hybridization buffer, suggesting that the observed Gal-b-BSA binding of this recombinant protein could be specific (data not shown). Similarly, concentration-dependent binding to Gal-b-BSA was observed with one of the other recombinant proteins (SG3-4), which

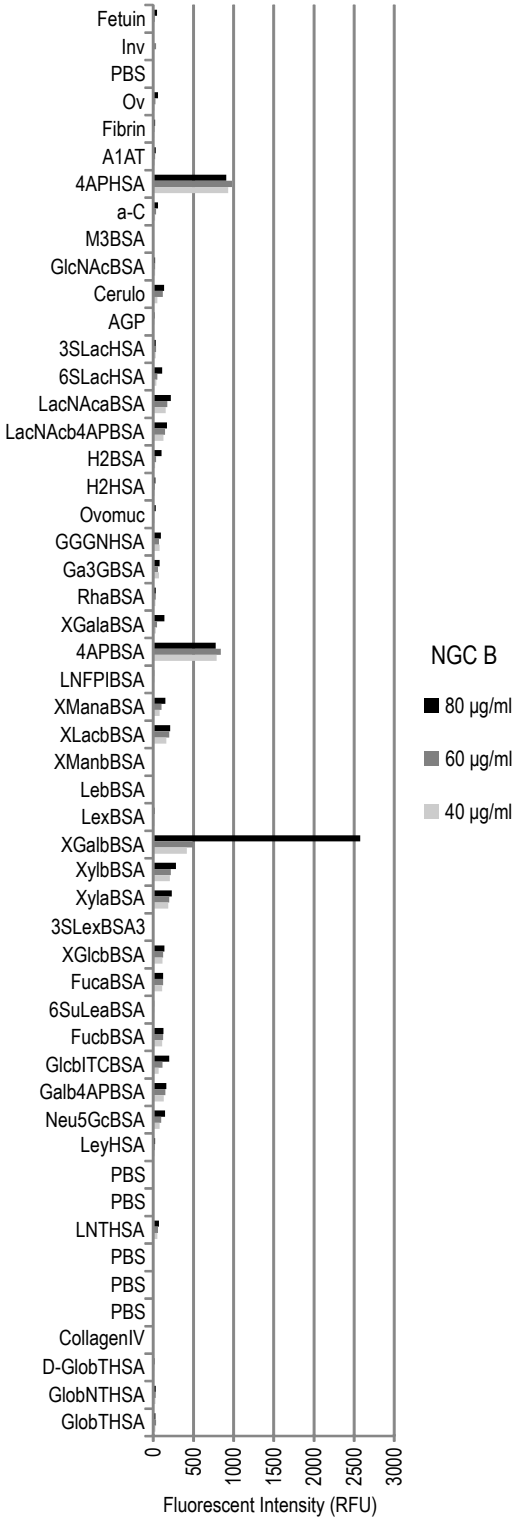


Figure 8.3 Recombinant protein of carbohydrate binding domain-construct SG3-4 (sequence mined from the human gut) using neoglycoconjugate microarray B. Low, yet concentration dependent binding observed at Gal-b-BSA.

also gave the highest binding intensity among the proteins tested, although this was still quite low intensity (see above; Figure 8.3). These most promising recombinant proteins deserve further testing of their carbohydrate-binding capacity to draw more definite conclusions. Analogous to the Fib2-type domain, the carbohydrate binding sequences may be too narrowly defined and could require more of the neighboring protein sequence they derive from, to achieve appropriate folding and function, or may require tandem repeated domain elements to allow a multi-domain binding pocket domain to be displayed. Further expansion of the domain-groups to be targeted and their corresponding functional assays is needed to determine the potential of the approach applied in this thesis to evaluate the functional properties of gut-enriched domains.

Additionally, by adapting the magnetic bead based "fishing" protocol for binding functionality (**Chapter 7**), another enrichment strategy for expression libraries, by whole cell binding of the expression clones that gained a specific binding functionality, could be realized. Such an enrichment strategy that takes place after generating a metagenomic expression library has not yet been reported. This procedure would both reduce the number of clones that need to be characterized and it would simplify the initial setup and storage as the cloning host cells do not need to be plated immediately after transformation with the cloning vector, i.e. the initial clone library could be directly used. Proof of principle for such an enrichment assay was obtained utilizing a known protein which confers a whole-cell binding function to its natural host, i.e. the mannose binding protein, encoded by *msa*, of *Lactobacillus plantarum* WCFS1 [43]. Wild-type (wt) *L. plantarum* WCFS1 adheres to mannose due to the protein encoded by *msa*, while the *msa* mutant derivative (*msa::cat*) has lost this adherence capacity [43]. Mixtures of overnight cultures of WCFS1 and its *msa::cat* derivative, in starting ratios of 1:9 and 1:99, were subjected to several rounds of an enrichment procedure. In each round of this procedure the culture mixtures were incubated with biotinylated oligo-6-mannose, according to the method of Grun and co-workers [44], coupled to Streptavidin coated magnetic beads (M280, Invitrogen, same procedure as for the coupling of biotinylated fibronectin in **Chapter 7**). Half of the magnetic beads were separated from the culture mixtures and were inoculated into fresh MRS medium and grown overnight. To assess if enrichment for the wild-type (mannose adhering) strain occurred, the other half of the magnetic bead associated culture fraction obtained in each enrichment round was plated on MRS agar plates (allowing growth of both the wild type and its *msa::cat* derivative). Per MRS plate 100 colonies were transferred to MRS plates supplemented with chloramphenicol on which only the mutant strain can grow, to assess the relative abundances within the population of the wild-type and the *msa* mutant. The percentage of wild-type colonies obtained after each round of enrichment is shown in Figure 8.4, and these data indicate that after 3 to 4 rounds approximately 90% enrichment was achieved. In this proof of principle experiment, the wild-type (WCFS1) cells could represent expression clones containing an insert that does encode for the function of interest, while the *msa::cat* mutant cells could represent the remaining metagenomic expression clones that exist within the complete library. Hence, such an enrichment strategy seems feasible for binding functions of proteins that are expressed at cell surface.

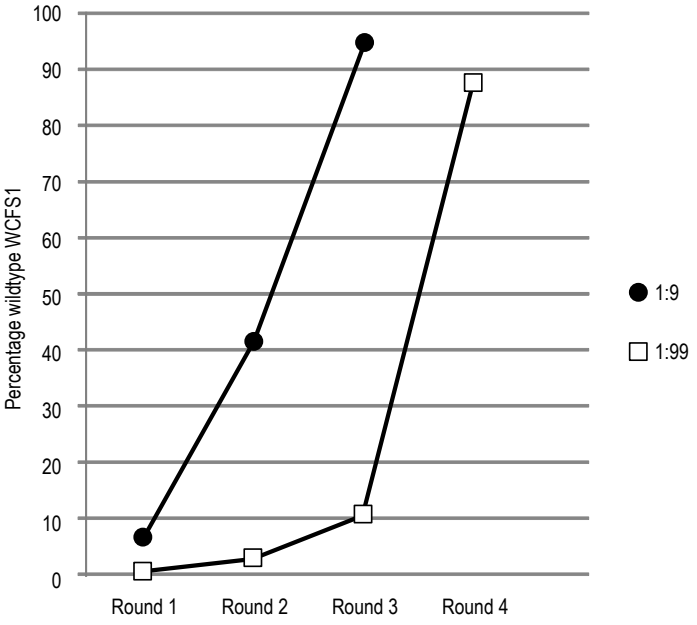


Figure 8.4 The percentage WCFS1 wildtype (wt) after several rounds of magnetic bead based enrichment on mixtures of overnight cultures of wt and msa::cat mutant. Starting ratios of wt to msa::cat were 1:9 and 1:99. In each round the mixtures were incubation with biotinylated oligo-6-mannose coupled to Streptavidin coated magnetic beads. The magnetic bead fraction was inoculated into fresh medium and grown overnight. Part of the magnetic bead fraction of each round was plated on MRS agar plates and per MRS plate 100 colonies were transferred to MRS plates supplemented with chloramphenicol on which only the msa::cat mutant can grow.

The three approaches for functional characterization of metagenomes (discussed above and in **Chapter 1**) can be applied in complementary setups, and have a different potential for novel function discovery versus sequence assignment to function or domain definition. With the targeted mining as is presented in **Chapter 7**, the use of synthetic genes avoids the requirement for specific PCR templates to obtain the desired sequences, which intrinsically also allows codon optimization to facilitate expression in the chosen expression host. Moreover, the tags that can be introduced via the cloning vector facilitate downstream characterization of the expressed protein or domain. For instance the C-terminal Avi-tag introduced to the fibronectin binding domains (**Chapter 7**) allows the recombinant protein to be biotinylated, which facilitates its fixation on Surface Plasmon Resonance based sensors that allow quantitative determination of its binding capacity. A drawback of the approach presented in **Chapter 7** may be that it is quite laborious and time consuming, and requires multiple controls in each subsequent step of the procedure, which makes the approach not readily compatible with high-throughput analyses. However, the alternative strategy that employs random cloning and functional screening of expression libraries requires high-throughput screening capacity to identify positive clones, which is in many cases not trivial to set-up. Still these random approaches have the potential to identify novel enzymes [45, 46], which is another aspect that is not achievable with the sequence mining-driven approach presented in this thesis. In conclusion, a combination, or sequential application, of random cloning combined with high-throughput screening, and a domain-function mining approach

could be a powerful scenario for the deciphering of relevant functional repertoires encoded within the microbiota of the intestinal tract.

Future perspectives

The ever-increasing association of the human gut microbiota and various aspects of human health and disease demands intelligent approaches to ultimately proof the biological causality of the reported associations. To be more precise in this "question on causality" framework it is crucial that this field proceeds towards understanding of particular bacterial functions in relation to aberrations of health and homeostasis. Microbiome derived molecular therapeutics is of interest for an ever-growing community of researchers spread over the globe. The latter can be questioned, or labeled as a "hype", but current literature is amassing quite some examples that encourage molecular therapeutics based on knowledge of the microbiome. For example, the bacterial metabolite trimethylamine N-oxide (TMAO), which has gained a bad reputation as TMAO blood plasma levels are associated with cardiovascular disease [47]. Both the dietary lipid phosphatidylcholine and the red meat component l-carnitine were shown, in mice models, to be converted by the gut microbiota into TMAO [47, 48]. Hence it makes sense to assess and manage the microbiota in patients with cardiovascular disease that have high levels of TMAO. The observation that in the previously discussed diet swap of Native Africans and African Americans, a high urine TMAO level is negatively correlated to the abundance of *Akkermansia muciniphila* is for instance a relavant observation providing new directions for follow up studies [38]. Another example, from the pharmaceutical area, is the inactivation of a drug for cardiac arrhythmia treatment, digotoxin, by some *Eggerthella lenta* strains which encode for the so-called cardiac glycoside reductase enzymes [49]. Interestingly, these enzymes can be inhibited by dietary arginine supplementation. Therefore, for the patients that rely on digotoxin, it would be of interest to determine whether they harbor these *Eggerthella lenta* strains, or other microbes that have genes encoding for cardiac glycoside reductase enzymes in their genomes, with the intention to adapt their diet to suppress the presence or drug-inactivation activity of these microbes, or place these patients on an alternative therapeutic regime. Such examples nicely illustrate that compounds ingested by humans, dietary or pharmaceutical, can be converted by the gut microbiota into metabolites which can exert different biological effects compared to the original compound and suggests that (personalized) microbiome derived molecular therapeutics indeed are a sensible avenue to pursue and expand in the (near) future. To progress our understanding of the gut microbiota community that can facilitate microbiome derived molecular therapeutics, several adaptations to frequently used research strategies and techniques would be very beneficial to this research field. Some recommendations are discussed in the following sections.

Other gut inhabitants

This thesis as well as the main bulk of the published studies has focused on the bacteria and their genetic content. While clearly being the most abundant cellular organisms in the gut, bacteria are not the only micro-organisms and representatives from the other kingdoms of life, i.e. *Eukarya* and *Archaea*, have also been detected in the gut [50]. To date, detection of archaeal species from

two different phyla have been reported. Members of the archaeal phylum Crenarchaeota have only been reported in one study, but members from the archaeal phylum Euryarchaeota are frequently encountered [50]. From the Euryarchaeota, the methanogenic archaea are most abundantly present in the human gut [51-53]. While not as abundant as various bacterial species the methanogenic archaea may still play a key role in the gut as their conversion of carbon dioxide and hydrogen into methane impacts on the thermodynamics of the system and leads to a net reduction of gas volume. This may be important in the context of gastrointestinal complaints such as distension, which is a symptom of IBS. In line with this reasoning, the IBS cohort studied in **Chapter 4** appeared to have four times less methanogens in their intestinal microbiota as compared to the healthy controls [54].

Compared to the *Archaea*, a more diverse set of organisms from the *Eukarya* kingdom have been reported to reside in the intestinal tract. Some of these eukaryote species, in particular *Candida* spp. [50], are well adapted to the conditions in the gut. Most eukaryote species in the gut, such as various yeast species [55] or filamentous fungi [56], are commonly only detected at low abundance levels and considering their most probable origin and/or instable temporal presence patterns, it is often assumed they are transient gut inhabitants that derive from normal or contaminated dietary constituents. However, for some individuals certain eukaryal intestinal parasites are suspected to be part of the normal gut microbiota repertoire [56]. Since infectious organisms may have pronounced and possibly long-term effects on the gut microbiota [57], future studies should not neglect eukaryal organisms with pathogenic potential even if they are not among the normal gut inhabitants.

Besides the cellular organisms in the intestinal tract, there seems to be another important biological entity which is frequently not investigated but likely the most abundant among the "micro-organisms", i.e., the viruses. The collection of all viral (derived) entities has been designated the "virome" and consists of both host viruses and viruses that predate on cellular gut micro-organisms, including bacteriophages. It has become evident that the full scope of viral entities has not yet been determined. Remote homology detection a human gut metagenome has shown an enrichment of viral sequences among the genes that have no homologues in the database [58], which indicates that not all viral DNA has been detected with standard data mining methods, but also implies that there is significant proportion of viral DNA with unknown functionality. However, it appears from sequencing studies that bacteriophages represent the most dominant viral particles in the human gut [59] and it is suggested that they may even outnumber the bacterial cell population by one order of magnitude [60]. Moreover, due to technological improvements a rising number of new eukaryotic RNA viruses are being detected, suggesting that the knowledge on this part of the human virome is currently still underrepresented [61]. This huge amount of viral particles in the gut is very likely to have a potent influence on the community structure of the gut microbiota with respect to predator-prey interactions and horizontal gene transfer [62]. In fact, it is therefore well possible that the virome plays an important role in host-microbe interactions. For instance, several lines of evidence have shown that the virome affects the microbiota composition and/or functioning that translates into an altered immune system status (for a review see [63]). Recently, an increase of the Caudovirales bacteriophage abundance

has been shown in IBD patients, further supporting a potentially important role of the virome in inflammatory disease states [64]. Additionally, evidence is increasing for viral presence in healthy subjects, which expands the conventional concept that the host merely responds to "fight the disease" [65, 66]. Hence, complete understanding of the gut microbiota in the context of human health cannot be completed without further exploring and understanding the role of the human virome.

Technological advances

Current community analyses could greatly benefit from a deeper taxonomic resolution, e.g. instead of only being able to detect *E.coli* it can of great importance to know if only the commensal strains [67] or also the Shiga toxin harboring strains [68] are detected. Both computational advances such as "oligotyping" [69] which identifies the most informative site in the 16S rRNA gene sequences to maximizes the resolution of the generated data, as well as methodological or technological advances that improve the quality and resolution of every sequence read obtained, like low-error amplicon sequencing [70], will improve 16S rRNA gene amplicon sequencing and allow discrimination of closely related taxa that ultimately could achieve single nucleotide resolution. However, random shotgun metagenomics is more preferable to investigate the gut community members as it could even allow the capture microbial evolution. In this community sequencing approach both computational approaches, such as co-abundance gene group analysis [7], and technological advancements, such as the newer generation of sequence technologies discussed below, will continue to improve microbiota profiling as well. Increasing sequence depth with the current generation of sequencing technologies may already improve single-nucleotide polymorphism detection and *de novo* assembly of contigs, but is costly and requires alternative sequencing strategies like the in depth sequencing of a subset of samples.

It should be emphasized that even if all microbial DNA can be sequenced confidently, now or in the near future, analysis of metagenomic gene sets has been frustrated by large amounts of genes that lack an annotation. The latter especially holds true for genes that have no homologues in the database, the so-called ORFans. These ORFans have been around since the first genomes were published and annotated [71] and do not seem to reduce by the exponential sequence increase of genomes and metagenomes [72]. Computational solutions are being developed and to date have provided two novel options: non-homology based or "guilty by association" methods and remote-homology detection. Various non-homology based functional annotation methods exist and some have already been applied to metagenomes [73, 74], although these methods work best on larger scaffolds or contigs (i.e., where more than one gene is present to allow genetic context associations), which is generally not achievable in metagenomic analysis of complex microbial communities when applying the current high-throughput, short-read sequencing methods. Remote-homology, often Hidden Markov Model (HMM) based, detection methods determine conservation profiles between families, rather than between single sequences [75, 76] and may boost automatic function prediction methods, but are computationally demanding. Nevertheless, one study analyzed over 35 million genes from three metagenomic data sets with remote homology detection and could infer functions onto approximately 15% of the ORFans

[58]. This is an impressive improvement, although it should be kept in mind that the predicted functionalities still require functional confirmation. Provided that such predicted functions can be experimentally tested, the approach presented in **Chapter 7** could be adapted to serve as a predicted-function confirmation approach. For the remaining unknown ORFans functional metagenomic laboratory screening methods or direct cloning of the sequence into an expression host, both in combination with various functional screening assays, may be the only options for characterization. However, the throughput of functional screening assays will need to be increased in order to make this a feasible methodology to pursue in the future.

To date the commonly used sequence technologies, such as 454 pyrosequencing, SOLiD, or Illumina sequencing, are from so-called second generation high-throughput sequencing (SG-HTS) methodologies. These SG-HTS technologies produce relatively short reads and are based on PCR amplification. Direct sequencing of the DNA molecule as they are isolated from *in situ* or *in vivo* samples, without PCR amplification and the corresponding bias risks, would greatly enhance the field of metagenomic research. With this in mind, the newest generation of sequencing technologies, the so-called third generation high-throughput sequencing (TG-HTS) technologies, focus on sequencing-by-synthesis [77], e.g., Helioscope Single Molecule Sequencer, Single Molecule Real-Time Sequencer, Single Molecule Real-Time (RNA Polymerase) Sequencer, Nanopore DNA Sequencer, and Multiplex Polony Technology, and do not require the initial PCR amplification step (for details see reviews: [77-79]). These technologies offer besides "low costs per base and increased sequence output" also significant conceptual advantages over the current generation sequencing technologies, including longer read lengths (1,000 bp and more), and do not require PCR amplification and can work with small amounts of starting material. The release of the PacBio RS II Single Molecule, Real-Time (SMRT®) DNA Sequencing System (Pacific Biosciences) enables long and high accuracy circular consensus sequencing (CCS), which has been compared with Illumina HiSeq with respect to random shotgun metagenomic analysis of a commercial biogas reactor community [80]. This comparison illustrates the potential of CSS by the genome assembly of two dominant phylotypes from the investigated community, which could not be achieved by HiSeq [80]. However, further improvements of the PacBio-like systems are needed to improve relative amount of high-quality reads over the currently large amount of poor-quality reads ("waste"), to ensure reliable application of the PacBio RS II system in metagenomic analyses. It is very likely that TG-HTS technologies will emerge that can improve contig assembly in metagenomic analysis, which will not only improve taxonomic resolution, but could also improve functional prediction due to its compatibility with the non-homology based, "guilty by association", annotation approaches.

Towards microbiome-derived molecular therapeutics

Although there is a clear need to characterize the gut microbiota at (metagenomic) DNA level even further, DNA itself is not representative for actual functioning. A more direct assessment of functionality can be achieved with community wide transcriptomics, proteomics or metabolomics. Moreover, metatranscriptomics is the only methodology by which the "genetic" content of free RNA viruses can be assessed [81]. Combinations of meta-omics approaches with physiological

host (cell) parameters can unveil novel mechanisms behind host-microbe interactions. As, for instance, was shown for the microbial metabolite TMAO where: TMAO production from dietary components was measured by metabolomics; TMAO modulation of gene expression in liver cells was detected by qRT-PCR [47]. Interesting dynamics relevant for the gut microbiota functioning can be identified by combinations of metagenomics with an additional omics approach. A study that combined metaproteomics and metagenomics provided more insight into Crohn's disease (CD), by not only confirming prior findings but also expanding towards new leads in the form of highly expressed, CD-specific, unknown proteins, that contribute to the further unraveling of the complex etiology in CD [82]. Proteomics allowed the identification of these CD-specific proteins, encoded by metagenomic genes of unknown function, and established their expression *in situ* and the potential importance in CD of these metagenomic genes that were previously annotated as "hypothetical proteins". These findings underpin the imporatnce of protein characterization efforts to clarify the specific role of such proteins in the CD patient's intestinal microbiota. Another study that combined metagenome and of the metatranscriptome analyses showed that biosynthesis of small molecules, such as tryptophan, was underrepresented among the transcribed genes relative to the encoded repertoire of genes, whereas methanogenesis genes of the archaea *Methanobrevibacter smithii* were highly enriched in the transcribed functions as compared to the metagenome [83]. Obviously, such selective expression of encoded functions would have been overlooked when only the metagenome would have been determined. Moreover, this combination of analyses provides insight in the specific activity profile of particular members of the microbiota, which for example, can confirm the microbe-microbe interaction hypotheses that come forward from co-occurrence network mining during dietary interventions (**Chapter 4**; **Chapter 5**; [38]). Further testing of such microbe-microbe interactions could then be performed in *in vitro* co-cultures, or employing *in vitro* synthetic communities. This kind of *in vitro* evaluation of bacterial interactions has already contributed to the understanding of the ecology within the gut, as for instance co-culturing of *Eubacterium hallii* and *Bifidobacterium adolescentis* in starch-containing media confirmed bacterial cross-feeding that allowed the conversion of the *B. adolescentis* metabolite lactate, into butyrate by the secondary fermentation by *E. hallii* [84]. A similar interaction is hypothesized to occur due to the re-arrangement of the microbiota structure in individuals consuming FOS (**Chapter 5**). Another example is the co-culturing of *Roseburia* spp. and *Bifidobacterium* species, which allows the *Roseburia* spp. to grow on FOS due to *Bifidobacterium* liberating a suitable growth substrate, i.e. fructose, during the breakdown of FOS and also results in butyrate formation [84, 85].

Finally, the ultimate way forward would be to have an integrative "ecosystems biology" approach that combines several omics technologies supplemented with physiological host parameters, such as the ecological model of small intestine proposed by Zoetendal and co-workers [86]. The previously mentioned example of *Eggerthella lenta* strains and their possible metabolism of the drug digoxin nicely illustrates the potential of a multi-omics approach [49]. In that study, the realization that cardiac glycoside reductase enzymes can inactivate digoxin, that the corresponding genes were only present in certain *Eggerthella lenta* strains and show an expression pattern that can be inhibited by arginine, could only come from the applied combination of

metabolomics, strain level profiling, and metatranscriptomics [49]. By adding gut macro-physiological parameters, such as gut transit time [15], and host cell responses in combination with a multi-omics approach will likely provide mechanisms for the majority of currently unclear associations of the host and its microbiota. Subsequently, well-defined interventions, taking into account the parameters mentioned above, may then allow scientists to address the causal relations between microbes and their host. It should be noted though, that our knowledge and insight in host-microbe interactions and our ability to modulate these interaction for the benefit of human health does require continuous improvement and development of computational tools in order to effectively mine the ever expanding data-sets that are generated in this field.

Acknowledgements

Special thanks to Dr. Marie Le Berre, Dr. Michelle Kilcoyne, Prof. Lokesh Joshi from the Glycoscience Group from the National University of Ireland Galway for their NGC microarray work on the recombinant carbohydrate binding domains.

References

1. **Arumugam M, Raes J, Pelletier E, et al.** Enterotypes of the human gut microbiome. Nature 2011;473(7346):174-80.
2. **Le Chatelier E, Nielsen T, Qin J, et al.** Richness of human gut microbiome correlates with metabolic markers. Nature 2013;500(7464):541-6.
3. **Lahti L, Salojärvi J, Salonen A, et al.** Tipping elements in the human intestinal ecosystem. Nat Commun 2014;5.
4. **Rajilic-Stojanovic M, Heilig HG, Molenaar D, et al.** Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 2009;11(7):1736-51.
5. **Toft U, Kristoffersen L, Ladelund S, et al.** The impact of a population-based multi-factorial lifestyle intervention on changes in long-term dietary habits: the Inter99 study. Prev Med 2008;47(4):378-83.
6. **Qin J, Li R, Raes J, et al.** A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59-65.
7. **Nielsen HB, Almeida M, Juncker AS, et al.** Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 2014;32(8):822-8.
8. **Goodrich JK, Waters JL, Poole AC, et al.** Human genetics shape the gut microbiome. Cell 2014;159(4):789-99.
9. **Huse SM, Ye Y, Zhou Y, et al.** A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One 2012;7(6):e34242.
10. **Jeffery IB, Claesson MJ, O'Toole PW, et al.** Categorization of the gut microbiota: enterotypes or gradients? Nat Rev Microbiol 2012;10(9):591-2.
11. **Claesson MJ, Jeffery IB, Conde S, et al.** Gut microbiota composition correlates with diet and health in the elderly. Nature 2012;488(7410):178-84.
12. **Knights D, Ward TL, McKinlay CE, et al.** Rethinking "Enterotypes". Cell Host & Microbe 2014;16(4):433-437.
13. **Wu GD, Chen J, Hoffmann C, et al.** Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science 2011.
14. **Tims S, Derom C, Jonkers DM, et al.** Microbiota conservation and BMI signatures in adult monozygotic twins. ISME J 2013;7(4):707-17.
15. **Rajilic-Stojanovic M, Heilig HG, Tims S, et al.** Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 2012.
16. **Vandeputte D, Falony G, Vieira-Silva S, et al.** Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut 2015.
17. **Tigchelaar EF, Bonder MJ, Jankipersadsing SA, et al.** Gut microbiota composition associated with stool consistency. Gut 2015.
18. **Lozupone CA, Stombaugh JI, Gordon JI, et al.** Diversity, stability and resilience of the human gut microbiota. Nature 2012;489(7415):220-30.
19. **Lepage P, Hasler R, Spehlmann ME, et al.** Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. Gastroenterology 2011;141(1):227-36.
20. **Manichanh C, Rigottier-Gois L, Bonnaud E, et al.** Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006;55(2):205-11.
21. **Turnbaugh PJ, Hamady M, Yatsunenko T, et al.** A core gut microbiome in obese and lean twins. Nature 2009;457(7228):480-484.
22. **Ouchi N, Parker JL, Lugus JJ, et al.** Adipokines in inflammation and metabolic disease. Nat Rev Immunol 2011;11(2):85-97.
23. **Shoelson SE, Lee J, and Goldfine AB.** Inflammation and insulin resistance. J Clin Invest 2006;116(7):1793-801.
24. **Ley RE, Backhed F, Turnbaugh P, et al.** Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 2005;102(31):11070-5.
25. **Yatsunenko T, Rey FE, Manary MJ, et al.** Human gut microbiome viewed across age and geography. Nature 2012;486(7402):222-227.
26. **Turnbaugh PJ, Backhed F, Fulton L, et al.** Marked alterations in the distal gut microbiome linked to diet-induced obesity. Cell host & microbe 2008;3(4):213-223.
27. **Costello EK, Lauber CL, Hamady M, et al.** Bacterial community variation in human body habitats across space and time. Science 2009;326(5960):1694-7.
28. **Cotillard A, Kennedy SP, Kong LC, et al.** Dietary intervention impact on gut microbial gene richness. Nature 2013;500(7464):585-588.
29. **Caporaso JG, Lauber CL, Costello EK, et al.** Moving pictures of the human microbiome. Genome Biol 2011;12(5):R50.

30. **Jalanka-Tuovinen J, Salonen A, Nikkila J, et al.** Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One 2011;6(7):e23035.
31. **Lampe JW, Navarro SL, Hullar MA, et al.** Inter-individual differences in response to dietary intervention: integrating omics platforms towards personalised dietary recommendations. Proc Nutr Soc 2013;72(2):207-18.
32. **Walker AW, Ince J, Duncan SH, et al.** Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J 2011;5(2):220-30.
33. **McOrist AL, Miller RB, Bird AR, et al.** Fecal butyrate levels vary widely among individuals but are usually increased by a diet high in resistant starch. J Nutr 2011;141(5):883-9.
34. **Benjamin JL, Hedin CR, Koutsoumpas A, et al.** Randomised, double-blind, placebo-controlled trial of fructo-oligosaccharides in active Crohn's disease. Gut 2011;60(7):923-9.
35. **Whelan K, Judd PA, Preedy VR, et al.** Fructooligosaccharides and fiber partially prevent the alterations in fecal microbiota and short-chain fatty acid concentrations caused by standard enteral formula in healthy humans. J Nutr 2005;135(8):1896-902.
36. **Korpela K, Flint HJ, Johnstone AM, et al.** Gut Microbiota Signatures Predict Host and Microbiota Responses to Dietary Interventions in Obese Individuals. PLoS ONE 2014;9(3):e90702.
37. **Haiser HJ and Turnbaugh PJ.** Is it time for a metagenomic basis of therapeutics? Science 2012;336(6086):1253-5.
38. **O'Keefe SJ, Li JV, Lahti L, et al.** Fat, fibre and cancer risk in African Americans and rural Africans. Nat Commun 2015;6:6342.
39. **Altschul SF, Madden TL, Schaffer AA, et al.** Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-402.
40. **Powell S, Szklarczyk D, Trachana K, et al.** eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res 2012;40(Database issue):D284-9.
41. **Zhou M, Boekhorst J, Francke C, et al.** LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics 2008;9:173.
42. **Zdobnov EM and Apweiler R.** InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001;17(9):847-8.
43. **Pretzer G, Snel J, Molenaar D, et al.** Biodiversity-based identification and functional characterization of the mannose-specific adhesin of Lactobacillus plantarum. J Bacteriol 2005;187(17):6128-36.
44. **Grun CH, van Vliet SJ, Schiphorst WE, et al.** One-step biotinylation procedure for carbohydrates to study carbohydrate-protein interactions. Anal Biochem 2006;354(1):54-63.
45. **Jones BV, Begley M, Hill C, et al.** Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A 2008;105(36):13580-5.
46. **Cecchini DA, Laville E, Laguerre S, et al.** Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. PLoS One 2013;8(9):e72766.
47. **Wang Z, Klipfell E, Bennett BJ, et al.** Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 2011;472(7341):57-63.
48. **Koeth RA, Wang Z, Levison BS, et al.** Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 2013;19(5):576-85.
49. **Haiser HJ, Gootenberg DB, Chatman K, et al.** Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science 2013;341(6143):295-8.
50. **Rajilic-Stojanovic M and de Vos WM.** The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 2014;38(5):996-1047.
51. **Dridi B, Henry M, El Khechine A, et al.** High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS One 2009;4(9):e7063.
52. **Miller TL and Wolin MJ.** Methanosphaera stadtmaniae gen. nov., sp. nov.: a species that forms methane by reducing methanol with hydrogen. Arch Microbiol 1985;141(2):116-22.
53. **Miller TL, Wolin MJ, Zhao HX, et al.** Characteristics of methanogens isolated from bovine rumen. Appl Environ Microbiol 1986;51(1):201-2.
54. **Rajilic-Stojanovic M, Biagi E, Heilig HG, et al.** Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology 2011;141(5):1792-801.
55. **David LA, Maurice CF, Carmody RN, et al.** Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505(7484):559-563.
56. **Scanlan PD and Marchesi JR.** Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J 2008;2(12):1183-93.
57. **Jalanka-Tuovinen J, Salojärvi J, Salonen A, et al.** Faecal microbiota composition and host–microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. Gut 2014;63(11):1737-1745.

58. **Lobb B, Kurtz DA, Moreno-Hagelsieb G, et al.** Remote homology and the functions of metagenomic dark matter. Front Genet 2015;6:234.
59. **Breitbart M, Haynes M, Kelley S, et al.** Viral diversity and dynamics in an infant gut. Res Microbiol 2008;159(5):367-73.
60. **Minot S, Bryson A, Chehoud C, et al.** Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 2013;110(30):12450-5.
61. **Acevedo A and Andino R.** Library preparation for highly accurate population sequencing of RNA viruses. Nat Protoc 2014;9(7):1760-9.
62. **Brussow H, Canchaya C, and Hardt WD.** Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 2004;68(3):560-602, table of contents.
63. **De Paepe M, Leclerc M, Tinsley CR, et al.** Bacteriophages : an underestimated role in human and animal health ? Frontiers in Cellular and Infection Microbiology 2014;4.
64. **Norman Jason M, Handley Scott A, Baldridge Megan T, et al.** Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. Cell 2015;160(3):447-460.
65. **Virgin HW, Wherry EJ, and Ahmed R.** Redefining Chronic Viral Infection. Cell 2009;138(1):30-50.
66. **Focà A, Liberto MC, Quirino A, et al.** Gut Inflammation and Immunity: What Is the Role of the Human Gut Virome? Mediators of Inflammation 2015;2015:7.
67. **Eckburg PB, Bik EM, Bernstein CN, et al.** Diversity of the human intestinal microbial flora. Science 2005;308(5728):1635-1638.
68. **Karch H, Tarr PI, and Bielaszewska M.** Enterohaemorrhagic Escherichia coli in human medicine. Int J Med Microbiol 2005;295(6-7):405-18.
69. **Eren AM, Maignien L, Sul WJ, et al.** Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol 2013;4(12).
70. **Faith JJ, Guruge JL, Charbonneau M, et al.** The long-term stability of the human gut microbiota. Science 2013;341(6141):1237439.
71. **Dujon B.** The yeast genome project: what did we learn? Trends Genet 1996;12(7):263-70.
72. **Tautz D and Domazet-Loso T.** The evolutionary origin of orphan genes. Nat Rev Genet 2011;12(10):692-702.
73. **Harrington ED, Singh AH, Doerks T, et al.** Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A 2007;104(35):13913-8.
74. **Vey G and Moreno-Hagelsieb G.** Beyond the bounds of orthology: functional inference from metagenomic context. Mol Biosyst 2010;6(7):1247-54.
75. **Sadreyev RI, Baker D, and Grishin NV.** Profile-profile comparisons by COMPASS predict intricate homologies between protein families. Protein Sci 2003;12(10):2262-72.
76. **Sanchez-Flores A, Perez-Rueda E, and Segovia L.** Protein homology detection and fold inference through multiple alignment entropy profiles. Proteins 2008;70(1):248-56.
77. **Schadt EE, Turner S, and Kasarskis A.** A window into third-generation sequencing. Hum Mol Genet 2010;19(R2):R227-40.
78. **Brockhurst MA, Colegrave N, and Rozen DE.** Next-generation sequencing as a tool to study microbial evolution. Mol Ecol 2011;20(5):972-80.
79. **Pareek CS, Smoczynski R, and Tretyn A.** Sequencing technologies and genome sequencing. J Appl Genet 2011;52(4):413-35.
80. **Frank JA, Pan Y, Tooming-Klunderud A, et al.** Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. bioRxiv 2015.
81. **Zhang T, Breitbart M, Lee WH, et al.** RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 2006;4(1).
82. **Erickson AR, Cantarel BL, Lamendella R, et al.** Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. PLoS ONE 2012;7(11):e49138.
83. **Franzosa EA, Morgan XC, Segata N, et al.** Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A 2014;111(22):E2329-38.
84. **Belenguer A, Duncan SH, Calder AG, et al.** Two routes of metabolic cross-feeding between Bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut. Appl Environ Microbiol 2006;72(5):3593-9.
85. **Falony G, Vlachou A, Verbrugghe K, et al.** Cross-feeding between Bifidobacterium longum BB536 and acetate-converting, butyrate-producing colon bacteria during growth on oligofructose. Appl Environ Microbiol 2006;72(12):7835-41.
86. **Zoetendal EG, Raes J, van den Bogert B, et al.** The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. ISME J 2012;6(7):1415-26.

Summary

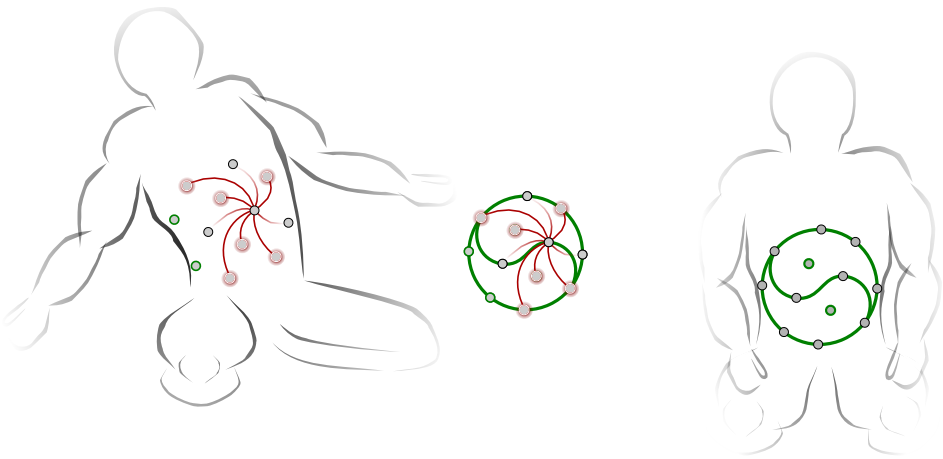
The gut microbiota is shaped by host genotype, early life imprinting, lifestyle and diet. Moreover, specific dietary components, such as prebiotics and probiotics, can have pronounced effects on the microbiota composition and functioning. The work presented in this thesis focused on studying the human gut microbiota, its association with host genotype (and early life imprinting) and host energy homeostasis, and how it is affected by prebiotic and probiotic consumption using 16S rRNA gene and metagenomics-based approaches.

Starting from the host, **Chapter 2** aimed to describe the influence of our own human genotype. The importance of the host genotype was illustrated in **Chapter 3** with a cohort of monozygotic twins, including a group of twins, the siblings of which were discordant for their Body Mass Index (BMI). An ever-increasing body of evidence that associates the gut microbiota with the energy homeostasis of its host has been reported. Results obtained from mice studies are clear though not massively reproduced, whereas in human studies the relation between the gut microbiota composition and energy homeostasis are contradictory, which is likely due to the heterogeneous study populations. In this work, the relation of the gut microbiota and host energy homeostasis was addressed with this setup that eliminated the confounding factors host genotype and early life imprinting, i.e. with the monozygotic twin siblings discordant for BMI. Relative abundances of genus-like groups could vary between co-twins but the signals within no less than 27 taxonomic groups, were significantly more similar between co-twin siblings compared to random unrelated subjects within this cohort. These results led to the hypothesis that several bacterial groups are structurally conserved by the host genotype, although the impact of early life (dietary) influences cannot be excluded as these twins were not separated at birth.

Different types of dietary intervention can be used to modulate the gut microbiota. Treatment of Irritable Bowel Syndrome (IBS) with a multi-species probiotic consisting of *Lactobacillus rhamnosus* GG, *Lactobacillus rhamnosus* Lc705, *Propionibacterium freudenreichi* PAJ, and *Bifidobacterium animalis* BB12, was successful in reducing IBS symptoms. **Chapter 4** described how the multispecies probiotic did not induce consistent drastic shifts in the bacterial composition, but altered the co-occurrences between bacterial taxa, suggesting an individual-specific re-alignment of the microbiota. Consumption of the prebiotic FOS did cause consistent shifts in the bacterial composition, described in **Chapter 5**. Moreover, also re-orientation of the bacterial co-occurrence networks was detected in FOS-consuming participants, which suggested that an alternative butyrate production pathway becomes more prominent in the colonic microbiota upon the consumption of FOS. The latter interpretation was based on knowledge obtained from cultured representatives of the taxa involved in the co-occurrence changes, illustrating the potential of extrapolated function approaches for generating new and testable hypotheses.

The importance of the knowledge gained with cultivation and characterization of isolates in laboratory settings, better known as “wet-lab” approaches, has always been important to understand the biology of those specific microbes and will likely remain so. However, most microbial groups in any ecosystem, including the gut, have only been identified based on a (partial) 16S rRNA gene sequence and are not represented by a cultured representative, which

hampers interpreting their role in the ecosystem. **Chapter 6** and **Chapter 7** aimed to provide starting points for combining molecular or sequence-based technologies with cultivation and characterization by wet-lab approaches. By combining 16S rRNA gene profiling and untargeted shotgun metagenomics **Chapter 6** highlighted various uncultured organisms with an interesting occurrence pattern among studied subjects and provided several leads for media development that could enable future isolation and cultivation attempts, as well as targeting various uncultured organisms due to their interesting genetic potential. Using fibronectin-binding domains as a proof of principle, **Chapter 7** illustrated a novel targeted approach to mine metagenomic gene catalogues and extract target genes of interest as well as a way to validate genes encoding proteins with predicted binding functions *in vitro*.



Appendices

Co-author affiliations

Michiel Kleerebezem [a,b,c]
Willem M. de Vos [a,d,e]
Erwin G. Zoetendal [a,f]

Jos Boekhorst [b]
Sandra J.M. ten Bruggencate [b]
Catherine Derom [g]
Sacha A.F.T. van Hijum [b,f,h]
Daisy M. Jonkers [i]
Victor C.L. de Jager [a,h]
Kajsa Kajander [j,k]
Riina A. Kekkonen [j]
Mirjana Rajilić–Stojanović [a,l]
Wim H. Saris [m]
Saskia van Schalkwijk [a,b]
Johannes Snel [a,b]
Harro Timmerman [b]
Robert Vlietinck [g]

[a] Laboratory of Microbiology Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands
[b] NIZO food research, Division of Health & Safety, P.O. Box 20, 6710 BA Ede, The Netherlands
[c] Host Microbe Interactomics Group, Wageningen University, De Elst 1, 6708 WD, The Netherlands
[d] Department of Veterinary Biosciences, Division of Microbiology and Epidemiology, 00014, University of Helsinki, Helsinki, Finland
[e] RPU Immunobiology, Department of Bacteriology and Immunology, and Department of Veterinary Biosicences, 00014, University of Helsinki, Finland
[f] TI Food and Nutrition, PO Box 557, 6700 AN, Wageningen, The Netherlands

[g] Department of Human Genetics, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium
[h] CMBI Bacterial Genomics, Radboud University Medical Centre, CMBI 260, NCMLS, PO Box 9101, 6500 HB, Nijmegen
[i] Department of Gastroenterology and Hepatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands
[j] Valio Ltd, R&D, PO Box 30, 00039 Helsinki, Finland
[k] Oy Verman Ab, PO Box 146, 04201, Kerava, Finland
[l] Department of Biochemical Engineering and Biotechnology, Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, Beograd 11000, Belgrade, Serbia
[m] Department of Human Biology, Maastricht University Medical centre MUCM+, Maastricht University, PO box 818, 6200 MD Maastricht The Netherlands

Nederlandse samenvatting

Binnenin ons lichaam, in onze darmen, zit een enorme hoeveelheid aan eencellige organismen die bij elkaar een complex ecosysteem vormen. Dit ecosysteem wordt een "microbiota" genoemd en de laatste jaren is er veel wetenschappelijk onderzoek geweest dat de samenstelling en het functioneren van de microbiota in de darm veelvuldig linkt aan onze gezondheid. Hoewel de wetenschap de microbiota nog lang niet geheel in kaart heeft, is het duidelijk dat de microbiota word gevormd door verscheidene factoren zoals het genotype van de gastheer (oftewel onze eigen humane set van DNA en de daarin ge-encodeerde genen), maar ook door de dagelijkse levensstijl en dieet gewoontes. Een kort overzicht van deze factoren is te vinden in **hoofdstuk 1**.

Het eerste deel van dit proefschrift richt zich dan ook op het bestuderen van een aantal van deze factoren die de samenstelling van de darm microbiota beïnvloeden. Hiervoor word voornamelijk gebruikt gemaakt van een universeel onderdeel van levende cellen, namelijk het "ribosoom". Elke cel heeft zijn eigen eiwitten en enzymen nodig om te kunnen (over)leven en deze worden via ribosomen in elkaar gezet. Bij bacteriën is één van de onderdelen van zo'n ribosoom vastgelegd in het 16S rRNA gen. Dit gen hebben ze dus allemaal en dit gen bestaat uit stukken die sterk geconserveerd zijn (de zogenaamde "conserved regions"), maar ook uit delen waarin veel variatie mogelijk is (de zogenaamde "variable regions"). Deze combinatie van "conserved" en "variable regions" in dit 16S rRNA gen kunnen we dan gebruiken om verschillende bacterie soorten uit elkaar te houden.

In **hoofdstuk 2** worden de verschillende bevindingen uit de wetenschappelijke literatuur op een rijtje gezet over de invloed van ons eigen genotype op de microbiota. Daarna wordt in **hoofdstuk 3** door middel van een cohort van eeneiige tweelingen de invloed van het genotype van de gastheer verder geïllustreerd. In dit tweelingen cohort had de helft van de tweeling paren een aanzienlijk verschil in Body Mass Index (BMI) tussen de broers of zussen. Deze zogenaamde discordante tweelingen stelde ons in staat om ook naar de relatie tussen de energie huishouding (homeostase) van de gastheer en de microbiota te kijken. De laatste 10 jaar is er namelijk ook steeds meer bewijsmateriaal verzameld die een belangrijke rol toeschrijft aan de microbiota in relatie tot onze energie homestase. Vooral in muizen studies zijn duidelijke, hoewel niet veelvuldig gereproduceerde, resultaten gevonden met betrekking tot specifieke bacteriële families. Menselijke studies laten echter nogal tegenstrijdige resultaten zien, wat waarschijnlijk te wijten is aan de heterogene, qua dieet en genotype betreft, studie populaties. De rekrutering van de discordante tweelingen in het cohort van **hoofdstuk 3** brengt een studie opzet met zich mee die het toelaat de impact van het genotype van de gastheer en tot zekere hoogte ook de "inprenting" uit de eerste levensjaren (zoals bijvoorbeeld geboorte middels een keizersnede of antibiotica gebruik) te scheiden van de relatie tussen gastheer energie homeostase en de microbiota. Naast een drietal bacteriële groepen die significant verschillen tussen de discordante tweelingen, zijn ook maar liefst 27 bacterie groepen die een aanzienlijk meer verglijkbaar profiel hebben tussen tweelingen dan tussen willekeurig niet-verwante personen. Deze resultaten suggereren dat, hoewel de totale hoeveelheid van deze bacteriën dus zelfs tussen tweelingen kunnen verschillen (en mogelijk in relatie staat met de gastheer energie homeostase), de verhoudingen van de onderlinge groepsleden vastgelegd word door het genotype van de gastheer.

Van de factoren die de microbiota beïnvloeden bied het dieet een interessante en relatief makkelijk te gebruiken methode om onze microbiota te manipuleren. Er zijn specifieke bestanddelen die aan onze voeding toegevoegd kunnen worden waarvan het bekend is dat ze de samenstelling en/of het functioneren van de microbiota drastisch kunnen veranderen, zoals bijvoorbeeld probiotica (levende bacteriën) en prebiotica (voedsel voor bacteriën; vaak zijn dit vezels). In dit proefschrift komen ook de effecten van dieet interventies op de microbiota aan bod in zowel een probiotische trial en een prebiotische trial.

In de probiotische trial (**hoofdstuk 4**) is een zogenaamd multi-species probiotica succesvol ingezet voor het verminderen van de symptomen van prikkelbare darm syndroom (IBS). Dit multi-species probiotica bestaat uit de vier bacterie stammen *Lactobacillus rhamnosus* GG, *Lactobacillus rhamnosus* LC705, *Propionibacterium freudenreichi* PAJ en *Bifidobacterium animalis* BB12. In **hoofdstuk 4** word de microbiota van de fecale monsters van de personen uit deze trial geanalyseerd en lijkt de inname van de multi-species probiotica geen drastische verschuivingen in de bacteriële samenstelling ten gevolge te hebben. Wel waren er per patiënt op het eerste gezicht willekeurige veranderingen terug te vinden in bacteriële populaties. Echter als we een coïncidentie analyse uitvoeren, waarbij we kijken of de bacterie groepen samen voorkomen of gelijke patronen hebben, blijken er in de loop van de trial een uitvoerige herstructurering van de bacteriële patronen plaats te vinden. Dit kan gezien worden als een verandering in de bacteriële netwerken. Deze netwerken omvatten bacteriële groepen waartoe soorten behoren waarvan bekent is dat ze specifieke rollen vervullen in de voornaamste fermentaties die plaats vinden in de dikke darm en lijken er soorten een rol te spelen die eigenlijk voornamelijk in de dunne darm actief zijn. Bovendien gebeurt dit alleen in de IBS patiënten die de probiotica gebruiken. Dit lijkt er dus op te wijzen dat de individu-specifieke wijzigingen van de microbiota niet zo willekeurig zijn en inderdaad een gevolg van de interventie zijn (en dus gepaard met de verbetering in symptomen).

In de prebiotische trial (**hoofdstuk 5**) kregen gezonde volwassen mannen middels een drankje extra FOS (wat voor menselijke enzymen onverteerbaar is) aan het dieet toegevoegd. Hier waren wel een aantal consistente veranderingen in de bacteriële samenstelling te zien in de fecale monsters van de meeste deelnemers: voornamelijk een toename van bifidobacteriën en een algemene afname van Bacteroidetes groepen. Echter, ook hier was een re-oriëntatie van de (andere) bacteriële groepen te vinden. De bacteriële groepen die betrokken zijn in deze netwerk verschuivingen suggereren dat er alternatieve routes voor butyraat productie in gang worden gezet. Butyraat is zogenaamd kortketig vetzuur dat wij zelf niet aan kunnen maken en wat een primaire energie bron is voor onze lichaamseigen darmwandcellen. Bovendien hebben vele voorgaande studies butyraat ook nog eens gelinkt aan darmweefsel fysiologie, ontstekings mechanismen en de opslag van vetten en suiker moleculen.

Dit soort analyses op het niveau van de microbiota samenstelling kan dus nieuwe en toetsbare hypothesen genereren. Het is echter goed om te beseffen dat deze biologische interpretaties alleen mogelijk zijn door eerdere studies die representatieve soorten uit de darm hebben geïsoleerd en deze “isolaten” functioneel hebben gekarakteriseerd. Uiteraard maken de bacteriën gebruik van vele andere genen die niet zo algemeen aanwezig en geconserveerd zijn als

het 16S rRNA gen. Deze genen bepalen uiteindelijk wel waartoe de bacterie functioneel toe in staat is. Als we bacterieel DNA isoleren uit een fecaal monster dan maken we in eerste instantie alle bacterie cellen kapot en krijgen we dus één grote poel van genen, ook wel een metagenoom genoemd. In de laatste fase van dit proefschrift verschuift de focus van 16S rRNA gen werk ("wie zijn er aanwezig?") naar metagenoom ("waartoe zijn ze in staat?") georiënteerd onderzoek.

De analyses van microbiële ecosystemen hebben in ieder geval nog te kampen met twee obstakels die het complete begrip van zo'n ecosysteem in de weg staan. Het eerste obstakel wordt veroorzaakt door het feit dat de meeste microbiële groepen in een ecosysteem, zo ook in de darm, slechts zijn "waargenomen" op basis van een (gedeeltelijke) 16S rRNA gen sequentie. Deze isolaten hebben dus geen gekweekte vertegenwoordigers, wat ons belemmert bij het interpreteren van hun rol in het ecosysteem. Dit zorgt er namelijk onder andere voor dat we niet precies weten welk genetisch materiaal deze isolaten inbrengen. Het gebruik van technieken die een metagenoom van een ecosysteem in kaart brengen zou een oplossing kunnen bieden. Immers pikken dit soort technieken in theorie alle bacteriële genen op. Echter, behalve dat we in de praktijk waarschijnlijk niet alles te pakken krijgen, blijkt ook dat voor het merendeel van de genen in de huidige metagenomen geen goede functionele voorspellingen mogelijk zijn en dit vormt dan ook het tweede obstakel. Het belang van het kweken en volledig karakteriseren van isolaten uit de darm middels de zogenaamde "natte laboratoriumtechnieken" is dus essentieel voor de biologische interpretatie van ecosystemen en dit zal voorlopig nog wel zo blijven.

In **hoofdstuk 6** en **hoofdstuk 7** hebben we gekeken naar oplossingen die mogelijk bij kunnen dragen bij het oplossen van de hiervoor genoemde obstakels. Door 16S rRNA gen gebaseerde analyse resultaten, die antwoord geven op "wie zijn er aanwezig?", te combineren met groepen genen die samen voorkomen in metagenomische datasets is het mogelijk deze genen toe te wijzen aan een aantal tot nu toe ongekweekte organismen (**hoofdstuk 6**). Dit stelt ons in staat gerichter te kunnen zoeken en betere isolatie en kweek methoden te ontwerpen. Maar ook door te kijken naar wat we kunnen afleiden uit het potentiaal van de gen sets kan er prioriteit worden gegeven aan welke organismen het interessants zijn om als eerste te isoleren, te kweken en te karakteriseren. Met de voorgestelde methode in **hoofdstuk 7** proberen we uiteindelijk dieper in te gaan op de genen met onbekende functies. Soms zijn kleinere stukken binnenin een gen, ook wel domeinen genoemd, nog wel te herkennen middels voorspellende modellen uit geavanceerdere bio-informatica tools. Aan de hand van fibronectin (een onderdeel van de buitenste laag om onze weefsels waar bacteriën als eerst mee in contact kunnen komen) bindende domeinen illustreert **hoofdstuk 7** een proof-of-principle voor een nieuwe gerichte aanpak om aan meer genen toch een functionele voorspelling te hangen en daarna deze gen sequenties ook in een laboratorium setting te testen.

Dankwoord

Eindelijk is het er dan toch van gekomen: mijn boekje is af. Het koste wat moeite… Dat leien dakje was soms ver te zoeken. Hoewel ik qua onderzoek niet mag klagen: voor- en tegenspoed waren redelijk in balans. Wel wemelde het van de "side projects" die overigens vaak niet eens verplicht waren, maar ondergetekende had (en heeft dit nog steeds enigszins) moeite om "nee" te zeggen tegen het verkennen van nieuwe wetenschappelijk interessante invalshoeken. Daar hebben genoeg mensen vast baat bij gehad, maar anderen zullen geduld moeten hebben gehad … al deze mensen wil ik hierbij graag bedanken.

Michiel, dank voor je toegewijde promotor taken. Ik heb veel kunnen leren van je diepe moleculaire kennis en jouw pragmatische aanpak heb ik enorm kunnen waarderen. Dankzij jou kon ik dan ook bij NIZO terecht op het moment dat het cloneer werk niet wilde vlotten bij Microbiologie en zoals jij al voorspeld had lag dat dan altijd aan één klein (lullig) detail. Alles werd altijd uitgebreid besproken en doorgenomen. Zelfs als ik dacht dat er niets nieuws te melden was dan liep onze meeting toch vaak uit. Ook voor zaken buitenom het werk had je veel begrip, toonde je altijd interesse en kon ik altijd bij je terecht. Dank daarvoor. Sowieso waren de tripjes met jou (en Erwin) naast het werk gedeelte ook altijd erg gezellig. Je houd enorm van lekker eten en had daar zo je "principes" in … Het mooiste was het diner op de luchthaven van Hong Kong waar je per se nog een maaltijd moest bestellen omdat de ober niet verder vroeg nadat jij, Erwin en ik ieder al een gezinsmaaltijd(!) hadden bestelt. Vrij logisch vanuit de ober gezien, maar "zo hoort dat natuurlijk niet". Een mooie herinnering die me altijd zal bijblijven (en bij die ober vast ook …).

Erwin, bedankt voor de praktische adviezen in het lab en daarbuiten. Ook jouw fylogenetische voorkennis was erg behulpzaam voor de biologische kant van de darm ecologie. Tijdens de vele tripjes voor het MetaHIT project was er ook ruimte voor flauwe grappen en andere ongein/ontspanning. Dankzij jou heb ik George Carlin ontdekt en kan ik nu niet meer zonder glimlach door een luchthaven lopen of aan boord van een vliegtuig gaan. Jouw muzieksmaak gaat me net wat te ver, al lijkt het gitaar gehalte in mijn gemiddelde playlist wel behoorlijk te zijn toegenomen de laatste jaren. Mijn vrouw is je daar overigens wel wat minder dankbaar voor…

"Imagine that: a plane full of grown humans—many of them partially educated—and someone is actually taking the time to describe the intricate workings of a belt buckle."

-- George Carlin

Uiteraard wil ik ook Willem bedanken. Als groen MSc studentje had ik geen idee dat het niet vaak voor kwam dat je een student tijdens zijn MSc thesis wilde spreken. Toen had je het meteen al over een PhD positie maar ik wilde eerst nog eens de boel in Rotterdam bekijken. Jij voorspelde al dat ik daar niet zou eindigen, maar je pushte nergens voor en liet de deur open… Daar ben grofweg een halfjaar later dan weer dankbaar doorheen gelopen. Hauke, fearless group leader (denk dat jij deze woord keuze wel begrijpt), dank voor je interesse en voor het leiden van de alsmaar uitdijende MolEco groep (kan niet makkelijk zijn). Ook waardeer ik het in bijzonder dat jij en Willem nog een extra jaar voor me hebben geregeld. Dit was deels wegens aspecten uit

het privé leven en deels wegens de vele niet (direct) thesis gerelateerde extra werkzaamheden en side projects. Hoe onvoorzien deze vertragende factoren soms ook waren, deze zaken vallen in principe buiten de verplichtingen van de werkgever om en dat jullie dan toch de moeite namen om iets te regelen is grandioos.

Vincent, kerel, wij kennen elkaar al sinds 2003, sinds de start van de BSc Biotechnology. Hoe mooi kan het zijn we via ietwat verschillende paden tijdens de MSc elkaar gewoon weer tegen kwamen bij Microbiology om een PhD te doen. Jouw relativerende humor en uiterst pragmatische manier van handelen kan ik zeer waarderen! Daar waar ik moeilijk kon stoppen totdat alle details boven water waren, kon jij me afremmen zodra de hoofdzaak duidelijk was. Ook dank dat jij en Marieke die oudste van mij af en toe een dagje wilde vermaken zodat ik toch nog iets kon doen aan dit boekje tijdens de drukke begin periode bij Danone.

Tom, we zijn op zelfde moment begonnen met dezelfde begeleiders. We hebben vooral aan het begin heel wat zelf moeten uitvogelen. Ik heb groot respect voor de ontwikkeling (ook op persoonlijk vlak!) die jij hebt laten zien tijdens jouw PhD traject. Je creativiteit (figuren en animaties in presentaties) en doortastendheid (doorgaan tot in de late avonduren en het oefenen en timen van je presentaties) zijn bewonderenswaardig. We hebben ook met een klein groepje een aantal diner + film combinaties gedaan en dat was altijd gezellig. Ik denk dat je goede stap hebt gedaan in je carrière en wens je nog veel succes toe. Tom en Vincent, dank dat jullie mijn paranimfen willen zijn, erg leuk en een mooie afsluiting van het geheel.

Voordat ik verdere microbiologen en aanverwanten ga bedanken wil ik graag als eerst een speciaal bedankje uitbrengen naar de mensen die de boel op de achtergrond laten draaien: de "supporting staff" - oftewel de mensen van het secretariaat, de technicians en de research assistenten). Anja, Carolien, Gosse, Hans H, Ineke, Jannie, Jorn, Nees, Philippe, Sjon, Steven, Ton, Tom S, Tom vd W, Wilma, Wim, bedankt! Anja en Carolien, dank voor alle hulp bij het aanhouden van mijn account. Hans H., dank voor het bijbrengen van de praktische kneepjes in het lab (al tijdens mijn MSc tijd), voor het verzorgen van de eindeloze lijst van bestellingen en voor je vriendschap, zelfs nu je in Singapore zit. Ton, dank voor je kalme maar daadkrachtige organisatie van alles wat met BHV te maken heeft, je uitleg en begeleiding op de HPLC apparatuur. Maar ook voor het bestellen van een set synthetische genen, wat lastiger was dan we dachten (vlak na het vertrek van Nees...), waarmee uiteindelijk **hoofdstuk 7** mogelijk werd.

Thanks to my (other) office mates for the first half of my time at Microbiobiolgy: Milkha and Tian. Milkha, I sure do miss your humor and your cooking! In the second phase I shared an office with Coline, Noora and Kyle. The latter two had to face the cynical attitude of the former and myself, so thanks for enduring this. Coline, dank voor de "muzikale" input en het blijven herhalen van jouw mantra: "Alles komt goed".

Mirjana, thanks for getting me started in the analyses of the fascinating world of gut ecology and for linking me to Willem and Michiel. Furthermore I'd like to thank all other MolEco's, especially Carien, Carmen, Clara, Corina, Dennis, Detmer, Farai, Floor, Gerben,

Heissa, Janneke, Javi, Jing, Kyle, Leo, Maria, Martha, Mauricio, Muriel, Naïm, Odette, Oylum, Peter, Petia, Sahar, Siavash, Susana, Sven, Teresita, Thomas and of course my students Eric S and Stefanie S, for the great atmosphere in the workgroup.

Sowieso heb ik het erg naar mijn zin gehad bij het Laboratory of Microbiology. Het internationale karakter van de groep bracht zowel leven in de brouwerij als interessante en soms, nou ja vaak, flauwe discussies met zich mee … met name op de vrijdag. Hereby I would like to thank every Microbiologist(-like) or other associated person that contributed to the open and friendly ecosystem that is called the Laboratory of Microbiology with a special thanks to Amos, Bart, Bas, Bram, Carel, Caroline, Christos, Daan, Edze, Elleke, Fabian, Farrakh, John R, Laura, Marcel, Marco, Mark D, Mark L, Mark M, Mark vP, Matthijs, Melvin, Raymond, Rienk, Sjoerd, Stan, Teunke, Tessa, Tijn, and Vicente for your contributions (conscious or otherwise) to this thesis.

Ruud H. --collega uit de Zeeuwse klei-- en Sjon, het organiseren van het lab uitje was me een waar genoegen. Of het uitje ook echt zo Legen-(wait-for-it)-dary is geworden zullen we maar in het midden laten, maar de voorpret was er niet minder om. WE-day teams (ja we hebben een foto op het erepodium… maar jammer dat de echte uitslag dan niet vanaf onderaan de lijst begon met tellen) en Veluwe loop-ers (2010 & 2012) bedankt voor de sportieve afleiding!

Niemand kan alles maar alleen doen en tijdens het tot stand komen van dit werk is er ook veel input van buiten het Micro lab geweest. Dus hierbij wil mensen uit de vele samenwerkingen bedanken. Helemaal Saskia en Roger (toen nog NIZO), voor jullie hulp bij (q)PCRs en het cloneren. Ook dank voor en natuurlijk ook jullie collega's Harro, Iris, Jos, Onne, Peter, Patrick en Maaike voor analyses, culturen, ideeën, discussies of gewoon voor de gezellige lunches wanneer ik een paar dagen op het NIZO bivakkeerde. I would like to thank all members of the MetaHIT consortium for the nice meetings, fruitful discussions and scientific opportunities. Henk Schols, dank voor je advies. Special thanks to Dr. Marie Le Berre, Dr. Michelle Kilcoyne, Prof. Lokesh Joshi from the Glycoscience Group from the National University of Ireland Galway for their efforts in charcterizing the binding capacity of the recombinant carbohydrate binding domains. Furthermore I'd like to thank the thesis committee for their efforts in the realization of this thesis.

Ook wil ik mijn nieuwe collega's bedanken. Ik zou de mensen van de qPCR club willen bedanken, maar ja… First rule of qPCR club is: you don't talk about qPCR club. So I'd like to thanks my fellow (former) platform members: Amra, Andrea, Annemarie B, Annemarie O, Bernadet, Charmaine, Chee Guan, Chee Young, Christophe, Claudia vdB, Claudia L, Desiree, Eefje, Evan, Gabriël, Guus, Harm, Ingrid, Jan, Jolanda, Joost, June, Kaouther, Kees, Lidewij, Marko, Mona, Paul, Philipp, Raish, Rob, Rocio, Roger (nu Danone), Sam, Shugui, Tiemen, Yvonne but also the rest of "purple 4" for your moral support. Danoners, jullie hebben niet veel bijgedragen aan de tekst in dit werk maar hebben mij een fijne werkplek bezorgd waar ik, mede dankzij jullie begrip, dit boekje nog heb kunnen afronden.

Als PhD student heb je ook zeker mensen buiten je werk om nodig. Hierbij wil ik dan ook Mark vR en mijn goede jeugd vrienden Joost en Stefan bedanken. Na al die jaren nog echte vrienden, ondanks dat we het allemaal erg druk hebben en bepaald niet meer bij elkaar om de hoek wonen. Ook wil ik studievrienden Petrouschka, Petra, Anja & Joël bedanken. Leuk dat we elkaar af en toe nog wel eens zien. Nu ik in een dorp woon begrijp ik ook de waarde van goede buren, nooit te beroerd om iets te ondernemen met die ukken van ons, dus straat-genoten jullie ook bedankt.

Michel en Evelyn, dank voor de gezellige etentjes, verjaardagen en ook voor het oppassen tijdens de laatste lootjes. Dat we nog maar vaak bij elkaar over de vloer mogen komen en dat onze kinderen nog lang bevriend mogen blijven. En Michel nu heb ik geen smoesjes meer om niet regelmatig weer in de bios te zitten.

Alle (oud) leden van de karate groep(en) in Wageningen, Herveld en Zetten wil ik bedanken voor de sportieve afleiding en de gezellige trainings-dagen en weekenden! Mensen van Living Healthy ook jullie bedankt voor de nodige afleiding. Maar ik wil graag nog een speciaal bedankje richten aan Richard. Zonder jou was en is (!) het echt niet mogelijk om de karate groepen aan de gang te houden. Dank voor al je belangeloze inzet, begrip en uitzonderlijk geduld!

Ook mijn familie, die mij en mijn gezin de laatste jaren niet veel gezien hebben, wil ik graag bedanken en een aantal mensen in het bijzonder. Broeder en natuurlijk Samantha, we spreken elkaar weinig maar verjaardagen, oud & nieuw, weekendje weg, weekje Ardennen en de andere logeer partijen waren me een groot genoegen. Ook dank dat jullie mij mezelf liet afzonderen achter de laptop om toch nog wat te tikken. Pa en ma, dank voor jullie onvoorwaardelijke steun. Hopelijk hebben jullie nu een beetje een idee van wat ik zoal gedaan heb en het veld waarin ik nog steeds bezig ben. Er zijn natuurlijk meer (schoon)familie leden die interesse hebben getoond en/of op de kinderen hebben gepast (Meta, Rini, Judith, Roos, om er zo maar een paar te noemen), allemaal bedankt!

Ilse, we hebben het niet makkelijk gehad. Murphy's law leek soms speciaal voor ons geschreven. Toch zijn we er nog steeds en onze boys ook! Dat ik er vaak genoeg niet was met mijn hoofd (en dat er in mijn DNA niet veel lijkt te encoderen voor emotionele functies… de darm microbiota kan veel, maar blijkbaar vult het dit specifieke tekort niet aan) is niet makkelijk voor je geweest. Dank dat je er nog altijd bent!

Oma, opa & oma, dit boekje is ook voor jullie.

&

En natuurlijk ben ik ook jou niet vergeten, maar ik ben wel even je naam kwijt… excuses en alsnog bedankt! And off course, I not forget you, but your name slipped my mind… apologies for that and thanks to you too!

"*There is nothing*
outside of yourself
that can ever enable you
to get better, stronger, richer,
quicker, or smarter.
Everything is within.
Everything exists.
Seek nothing outside of yourself."

-- Miyamoto Musashi,
The Book of Five Rings

About the author

Sebastian Tims was born on the 22nd of July 1985 in Goes, The Netherlands. After completing his secondary education at Buys Ballot College in Goes in 2003, he continued his education with the multidisciplinary Biotechnology study-program at the Wageningen University and Research Centre. In the BSc phase he specialized in Cellular and Molecular Biotechnology, which led to an extensive training on "Putative protein interactors with Like Heterochromatin protein 1 in Arabidopsis" at the Laboratory of Molecular Biology. Sebastian obtained his BSc degree, cum laude, in August 2006. In the MSc phase he developed an interest in the application of molecular technologies for human wellbeing and specialized in Medical Biotechnology. His MSc thesis focused on the "Phylogenetic microarray analysis of the faecal microbiota of irritable bowel syndrome patients during a probiotic trial" and was performed at Laboratory of Microbiology. He continued to work on microbial communities in or on the human body at the Erasmus MC, Rotterdam, in a collaboration between the Forensic Molecular Biology and the Clinical Microbiology departments during his internship. From this internship the work on the fingertip microbiota, titled "Dynamic colonization of fingertip microbiota for forensic purposes", became his first first-author publication. Sebastian obtained his MSc degree, cum laude, in August 2008 after which he returned to the Laboratory of Microbiology to continue working on the gut microbiota and the associated large data sets, which became his prime interest. This research was part of the European Community's 7th Framework Program "Metagenomics of the human intestinal tract" (MetaHIT). His project aimed to uncover links between phylogenetic composition in the human intestinal microbiota and host health states, as well as developing novel methodologies for mining the genetic potential of uncultured microbes. In November 2013 he started at Danone Nutricia Research as a scientist to work on the gut microbiota of infants and patients with inherited metabolic disorders.

List of publications

Fructo-oligosaccharides induce a shift from Bacteroidetes to *Bifidobacterium* and restructure gut microbiota in healthy subjects. **Tims S**, Snel J, Bruggencate SJM ten, Schalkwijk S van, Timmerman H, Boekhorst J, Zoetendal EG, Vos WM de, Kleerebezem M. *Submitted*

Multispecies probiotic intervention induces microbial network changes in line with symptom reduction in patients with irritable bowel syndrome. **Tims S**, Rajilić–Stojanović M, Kajander K, Kekkonen RA, Zoetendal EG, Kleerebezem M, Vos WM de. *Submitted*

Partly fermented infant formulae with specific oligosaccharides support adequate infant growth and are well-tolerated. Huet F, Abrahamse-Berkeveld M, **Tims S**, Simeoni U, Beley G, Savagner C, Vandenplas Y, Hourihane J O'B *Submitted*

Application of the Human Intestinal Tract Chip to the non-human primate gut microbiota. Bello Gonzalez TDG, Passel MWJ van, Tims, S, Fuentes S, Vos WM de, Smidt H, Belzer C (2015). *Beneficial Microbes* 6:271-276.

Fat, fibre and cancer risk in African Americans and rural Africans. O'Keefe SJ, Li JV, Lahti LM, Ou J, Carbonero F, Mohammed K, Posma JM, Kinross J, Wahl E, Ruder E, Vipperla K, Naidoo V, Mtshali L, **Tims S**, Puylaert PGB, DeLany J, Krasinskas A, Benefiel AC, Kaseb HO, Newton K, Nicholson JK, Vos WM de, Gaskins HR, Zoetendal EG (2015). *Nature Communications* 6:6342.

Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent *Clostridium difficile* infection. Fuentes Enriquez de Salamanca S, Nood E. van, **Tims S**, Heikamp-de Jong I, Braak C.J.F. ter, Keller J.J, Zoetendal E.G, Vos W.M. de (2014). *ISME Journal* 8:1621-1633.

Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E; Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto JM, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, **Tims S**, Ussery DW, Yamada T, MetaHIT Consortium, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD (2014). *Nature Biotechnology* 32:822-828.

Faecal Microbiota Composition in Adults Is Associated with the FUT2 Gene Determining the Secretor Status. Wacklin P, Tuimala J, Nikkilä J, **Tims S**, Mäkivuokko H, Alakulppi N, Laine P, Rajilic-Stojanovic M, Paulin L, Vos WM de, Mättö J (2014). *PLoS One* 9:4.

Richness of human gut microbiome correlates with metabolic markers. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, **Tims S**, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD, Raes J, Hansen T, MetaHIT consortium, Bork P, Wang J, Ehrlich SD, Pedersen O (2013). *Nature* 500:541-546.

Long-term monitoring of the human intestinal microbiota composition. Rajilic-Stojanovic M, Heilig GHJ, **Tims S**, Zoetendal EG, Vos WM de (2013). *Environmental Microbiology* 15:1146-1159.

Microbiota conservation and BMI signatures in adult monozygotic twins. **Tims S**, Derom C, Jonkers DMAE, Vlietinck R, Saris WH, Kleerebezem M, Vos WM de, Zoetendal EG (2013). *ISME Journal* 7:707-717.

Enterotypes of the human gut microbiome. 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, MetaHIT Consortium, Weissenbach J, Ehrlich SD, Bork P (2011). *Nature* 473:174-180.

Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Rajilic-Stojanovic M, Heilig GHJ, Kajander K, Kekkonen RA, **Tims S**, Vos WM de (2011). *Gastroenterology* 141:1792-1801.

Microbial DNA fingerprinting of human fingerprints: dynamic colonization of fingertip microflora challenges human host inferences for forensic purposes. **Tims S**, Wamel W van, Endtz HP, Belkum A van, Kayser M (2010). *International Journal of Legal Medicine* 124:477-481.

As a member of the MetaHIT consortium

Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Krogh Pedersen H, Arumugam M, Kristiansen K, Voigt AY, Vestergaard H, Hercog R, Igor Costea P, Kultima JR, Li J, Jørgensen T, Levenez F, Dore J, **MetaHIT Consortium**, Nielsen HB, Brunak S, Raes J, Hansen T, Wang J, Ehrlich SD, Bork P, Pedersen O (2015). *Nature* 528:262-266.

An integrated catalog of reference genes in the human gut microbiome. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, **MetaHIT Consortium**, Bork P, Wang J (2014). *Nature Biotechnology* 32:834-841.

A human gut microbial gene catalogue established by metagenomic sequencing. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, **MetaHIT Consortium**, Bork P, Ehrlich SD, Wang J (2010). *Nature* 464:59-65.

Overview of completed training activities

Discipline specific activities

Systems biology course: Statistics of -omics data analysis	Wageningen, 2008
Light in the intestinal tract tunnel	Helsinki, Finland, 2009
Multivariate Analysis	Wageningen, 2009
Microbial Genomics & Metagenomics Workshop	Walnut Creek, CA, USA, 2009
Functional Metagenomics of the Intestinal Tract and Food-Related Microbes †	Helsinki, Finland, 2011
International Human Microbiome Consortium Kick-off	Heidelberg, Germany, 2008
Gutday	Utrecht, 2008
Next generation sequencing	Utrecht, 2009
Gutday	Vlaardingen, 2009
MetaHIT – International conference on metagenomics	Shenzhen, China, 2010
American Society of Microbiology †	Miami, FL, USA, 2010
Lactic Acid Bacteria 10 †	Egmond aan Zee, 2011
Gutday †	Wageningen, 2011
MetaHIT – International Human Microbiome Congress †	Paris, France, 2012
KNVM-Microbial Ecology Fall meeting ‡	Wageningen, 2012

General courses

VLAG PhD week	Bilthoven, 2008
Teaching and Supervising Thesis Students	Wageningen, 2010
Networking – NWO talent class	Den-Haag, 2010
Techniques for Writing and Presenting Scientific Papers	Wageningen, 2011
Communication with the media and the general public	Wageningen, 2011
AcKlas - Meer academici voor de Klas	Wageningen, 2012
Career Perspectives	Wageningen, 2013

Optional courses and activities

Preparing PhD research proposal	Wageningen, 2008
General laboratory meetings	Wageningen, 2008-13
Advanced Bioinformatics (MSc course, BIF-30806)	Wageningen, 2009
MetaHIT meeting	Heidelberg, Germany, 2008
MetaHIT meeting ‡	Barcelona, Spain, 2009
MetaHIT meeting ‡	Copenhagen, Denmark, 2009
MetaHIT meeting ‡	Shenzhen, China, 2010
MetaHIT meeting ‡	Noordwijkerhout, 2010
MetaHIT meeting ‡	Annecy, France, 2011

† Poster; ‡ Presentation

The research described in this thesis was financially supported by the European Community's Seventh Framework Seventh Framework Programme (FP7/2007-2013) under the grant agreement HEALTH-F4-2007-201052 (MetaHIT), coordinated by S Dusko Ehrlich (Institut National de la Recherche Agronomique, France).

Colophon

Cover design: Sebastian Tims
Inkscape 0.91, http://www.inkscape.org
Font cover:

Layout design: Sebastian Tims

Typeset: InDesign® CS6
Font Inner Works: Adobe Garamond Pro

Printed by: Gildeprint Drukkerijen – The Netherlands

Printing of this thesis was financially supported by:
Danone Nutricia Research