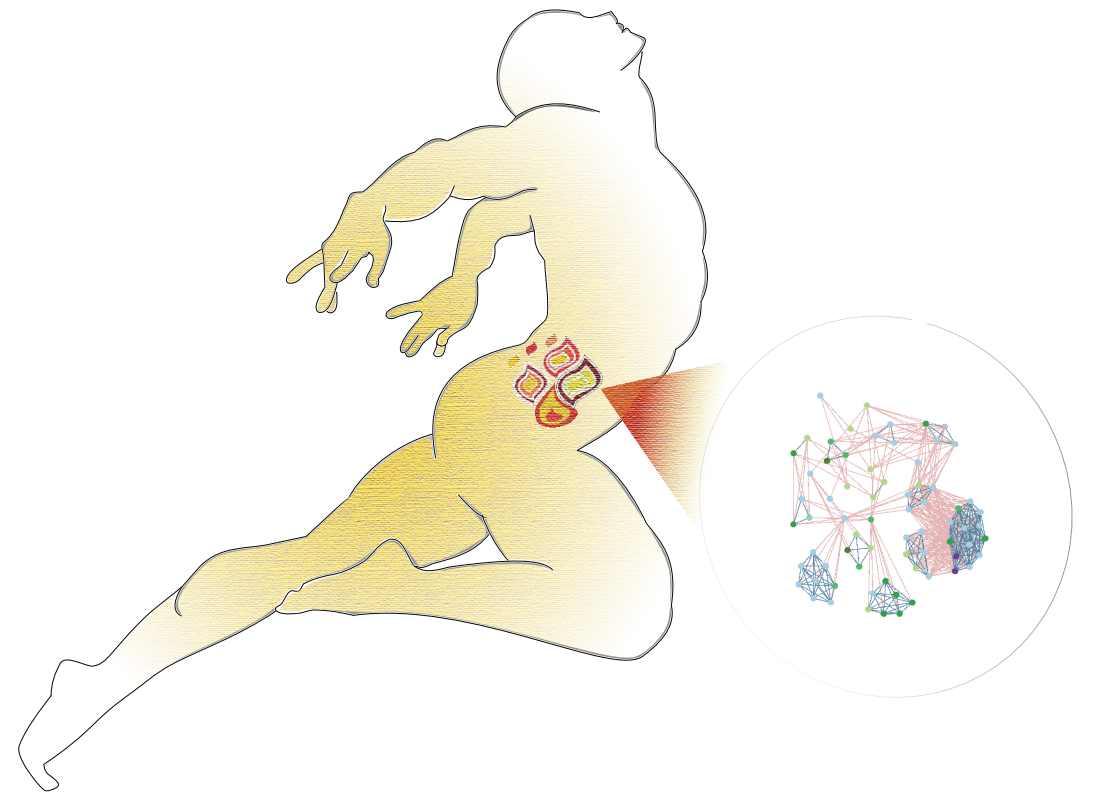# Ecophysiological insights into the human intestinal microbiota: from single strains to defined consortia

Sudarshan Anand Shetty

Propositions

1. Intestinal bacterial communities should be viewed along the r-K continuum that is governed by top–down and bottom–up selection processes.
   (this thesis)

2. Hyperdominance is the norm rather than the exception in microbial communities.
   (this thesis)

3. It is crucial to create sustainable funding models for long term accessibility and usability of bioinformatics research software.

4. In contrast with natural predation, human activities have a deleterious impact on wildlife.
   (Darimont CT, et al., 2009).

5. Identifying problems is good, working towards solutions is better.

6. People who have a healthy self-esteem, rarely indulge in selfies.

Belonging to the thesis entitled:

"Ecophysiological insights into the human intestinal microbiota: from single strains to defined consortia"

Sudarshan A. Shetty

Wageningen, March 20th 2019

Ecophysiological insights into the human intestinal microbiota: from single strains to defined consortia

Sudarshan Anand Shetty

Thesis committee

Promotors
Prof. Dr Willem M. de Vos
Distinguished Professor
Wageningen University & Research

Prof. Dr Hauke Smidt
Personal chair at the Laboratory of Microbiology
Wageningen University & Research

Co-promotor
Dr Leo Lahti
D.Sc. & Adjunct Professor
Department of Mathematics and Statistics,
University of Turku, Finland

Other members
Prof. Dr T. Abee, Wageningen University & Research
Dr R. Kleerebezem, Delft University of Technology
Dr P. Wilmes, University of Luxembourg
Dr A. Salonen, University of Helsinki, Finland

Ecophysiological insights into the human intestinal microbiota: from single strains to defined consortia

Sudarshan Anand Shetty

Thesis
submitted in fulfilment 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 Wednesday 20 March 2019
at 4 p.m. in the Aula,

Table of contents

Chapter 1

General introduction and thesis outline
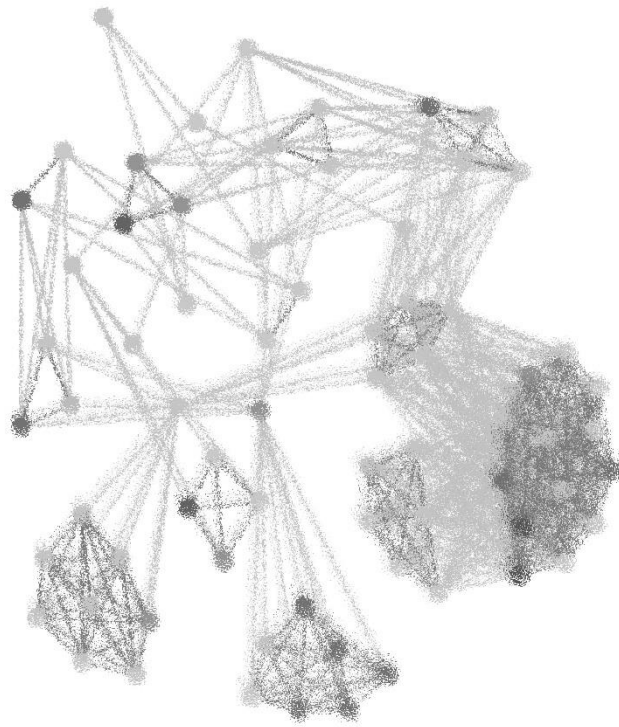
Introduction

Planet earth is colonised by over 10^{30} prokaryotic microbes with an estimated 1 trillion (10^{12}) microbial species (Green and Bohannan, 2006). These include free-living microorganisms as well as those associated with hosts such as plants and animals. The microbes associated with the human body colonise all outer surfaces, such as the skin as well as along the entire digestive tract ranging from the oral cavity, esophagus, stomach, small intestine and colon to the anus. A collective term, human microbiome, is used to describe these microbial communities and their function (Marchesi and Ravel, 2015). The human colon harbours anywhere between an estimated 1000-2500 microbial species accounting for trillions of microbial cells (Locey and Lennon, 2016; Qin *et al*., 2010; Ritari *et al*., 2015; Sender *et al*., 2016). The microbes in the colon, are in a continuous state of dynamic interactions with the host as well as other microbes (Shetty *et al*., 2017). Consequently, microbes play a major role in balancing human health while the human host also has an impact on the microbes (Flint, 2011; Shetty *et al*., 2017; Sonnenburg *et al*., 2016; de Vos and De Vos, 2012). For the past two decades, researchers from diverse fields of biology, sociology, and engineering have increased their focus on unravelling the influence of microbes on human health and well-being. Combined efforts have led to unprecedented insights into the complex interplay between microbes and the human body.

Molecular tools for profiling of microbial composition and function

The development and application of molecular tools has played a pivotal role in driving human microbiome research. Initially, molecular studies were focused on using 16S ribosomal RNA (rRNA) gene sequences PCR-amplified from total community genomic DNA extracted from faeces (Zoetendal *et al*., 2008). The 16S rRNA gene based identification of bacteria and archaea provided insights into the diverse community members, most of which had been undetected using culture dependent approaches (Eckburg *et al*., 2005). Comparative microbial community surveys of healthy and diseased individuals revealed differences in microbial community composition suggesting possible associations between the microbiome and host disease (Rajilić-Stojanović *et al*., 2011; Zoetendal *et al*., 2008). Later on, with the development of next

generation sequencing technology, large scale surveys of the human microbiome became feasible, providing unprecedented insights into diversity, structure and function (Larsen and Dai, 2015; Qin *et al.*, 2010). Potential functions were identified using shotgun metagenomics approaches by sequencing short fragments of DNA rather than just sequencing the PCR-amplified 16S rRNA gene. This approach led to further advances in comparative analysis of functional content between healthy and diseased individuals. For example, comparative metagenomic analysis of diabetic and healthy subjects revealed that patients with type 2 diabetes had enriched microbial functions related to sulphate reduction and oxidative stress resistance (Qin *et al.*, 2012). Application of metatranscriptomics, i.e. sequencing of cDNA obtained from total mRNA, has added to the information on actively transcribed genes in the human microbiome (David *et al.*, 2014; Zoetendal *et al.*, 2012). Going further from transcription level information to the translation of mRNA to proteins using metaproteomics, it is possible to gain information on microbiome function dynamics at the protein level (Klaassens *et al.*, 2007; Kolmeder *et al.*, 2012). For example, metaproteomics analysis aided in identifying 1,000 proteins that were part of the stable common core proteome and that were mainly involved in degradation and transport of carbohydrates (Kolmeder *et al.*, 2012). Integrating, meta- genomics, -transcriptomics and -proteomics provides information at various levels to better understand microbe-microbe and host-microbe interactions (Heintz-Buschart *et al.*, 2016). These interactions are mainly established *via* the variety of small molecules produced by microbes (Levy *et al.*, 2016; Nicholson *et al.*, 2012). To this end, recent advances in high throughput measurements of metabolites has created considerable interest in the field of human intestinal microbiome research. Microbial metabolites play an important role in modulating the immune system and identification of crucial metabolites that have beneficial effects may lead to novel drug discovery (Dorrestein *et al.*, 2014; Larsen and Dai, 2015).

Bioinformatics tools for analysing microbial composition and function

Analysis of molecular data obtained from high-throughput sequencing is an important part of unravelling composition, function and interaction within the microbiome.

Table 1: A list of various bioinformatics tools that are used for microbiome research.

Tools	Feature (s)	References
Pre-processing and quality control		
cutadapt	Remove adaptor sequences from raw reads	(Martin 2011)
AmpliSAT	A suite of tools for processing raw reads	https://tinyurl.com/yajnzodx
FastQC	Widely used tool for quality control checks on raw sequence data obtained from NGS	(Andrews 2010)
MultiQC	Generates reports for quality of data	(Ewels, *et al.* 2016)
Trimmomatic	Raw read processing obtained from Illumina	(Bolger, *et al.* 2014)
Marker gene sequencing data analysis		
QIIME1 and 2	Pipeline with multiple OTU picking strategies, taxonomic assignments and data visualisations	(Caporaso, *et al.* 2010)
Mothur	Widely used software tool for analysing marker-gene sequencing data	(Schloss, *et al.* 2009)
DADA2	Identifies unique sequence variants by incorporating quality-aware model for Illumina amplicon errors	(Callahan, *et al.* 2016)
NG-Tax	New tool for improve OTU identification with denosing of error reads	(Ramiro-Garcia, *et al.* 2016)
LotuS	Demultiplex, denoise, and construction of OTU tables	(Hildebrand, *et al.* 2014)
PANGAE	A pipeline/workflow for amplicon data analysis	(Giongo, *et al.* 2010)
phyloseq	R based tool for downstream analysis and visualisation of OTU tables generated by a variety of OTU/ASV identifying tools	(McMurdie and Holmes 2013)
microbiome	R based tool for downstream analysis and visualisation of OTU tables generated by a variety of OTU/ASV identifying tools. This tool depends on the phyloseq data structure	(Lahti and Shetty 2018)
Ampvis2	downstream analysis and visualisation of OTU tables generated by a variety of OTU/ASV identifying tools. Improved visualisation of multivariate data	(Andersen KSS, *et al.* 2018)
Calypso	Webserver for analysis and visualisation of OTU tables generated by a variety of OTU/ASV identifying tools	(Zakrzewski, *et al.* 2016)
MicrobiomeAnalyst	Webserver for analysis and visualisation of OTU tables generated by a variety of OTU/ASV identifying tools	(Dhariwal, *et al.* 2017)
Shotgun metagenomic data analysis		
IDBA-UD	Assembly of reads based on de Bruijn graph approach	(Peng, *et al.* 2012)
MetAMOS	Provides an integrated environment for metagenomic assembly and analysis	(Treangen, *et al.* 2013)

Ray meta	Scalable for big data, it uses a *de novo* assembly approach in parallel with a de Bruijn graph	(Boisvert*, et al.* 2012b)
MEGAHIT	Assembly based on improved de Bruijn graph assembly approach	(Li*, et al.* 2015)
MOCAT	Metagenomics assembly and gene prediction	(Kultima*, et al.* 2012)
metaSPAdes	Metagenomic assembler	(Nurk*, et al.* 2017)
mmgenome	Genome extraction from metagenomes for R software environment	(Karst*, et al.* 2016)
HUMAnN2	Analysis of presence/absence and abundance of microbial pathways from metagenomes, remove human DNA contamination from metagenomes	(Abubucker*, et al.* 2012)
MetaPhlAn2	Obtain taxonomic profiles from metagenomics data	(Truong*, et al.* 2015)
MG-RAST	Web based tool for analysis of metagenomics data	(Meyer*, et al.* 2008)
PanPhlAn	Strain-level microbial profiling using metagenomics data	(Scholz*, et al.* 2016a)
IMG M/ER	Web based curated, interactive data analysis of metagenomes, genomes, metatranscriptomics	(Markowitz*, et al.* 2012)
Kraken-HLL	Fast taxonomic classifications of metagenomics data	(Wood and Salzberg 2014)
DIAMOND	Ultra-fast alignment of reads to protein database	(Buchfink*, et al.* 2014)
BLAST+	Read alignment tool	(Camacho*, et al.* 2009)
MEGAN6	MEGAN 6 for interactive exploration and analysis of metagenomics data	(Huson*, et al.* 2016)
Centrifuge	Classification of metagenomic data	(Kim*, et al.* 2016)
CONCOCT	Identifies genome bins from metagenomics data	(Alneberg*, et al.* 2014)
metaSNV	Analysis of single nucleotide level variation in metagenomics data	(Costea*, et al.* 2017a)
Metatranscriptomics data analysis		
MetaTrans	Pipeline for quality control, taxonomic and gene expression analysis of RNA-Seq data	(Martinez*, et al.* 2016)
MG-RAST	Web based tool for analysis of RNA-Seq data	(Meyer*, et al.* 2008)
COMAN	Web based pipeline for metaranscriptomics data analysis	(Ni*, et al.* 2016)
IMG M/ER	Web based curated, interactive data analysis of metatranscriptomics	(Markowitz*, et al.* 2012)
Metaproteomics data analysis		
MetaProteomeAnalyzer	Software suite for analysis of label free quantification of proteins obtained from LC-MS	(Muth*, et al.* 2015)
iMetaLab 1.0	Web based automated pipeline for metaproteomic mass spectrometry data	(Liao*, et al.* 2018)

As mentioned previously, microbiomes can be studied at various levels of detail, from microbial community surveys to identifying functions and metabolic networks. Integration of such multi-level information obtained by multi-omic approaches has been pivotal for understanding the importance of the microbiome in human health (Erickson*, et al.* 2012, Heintz-Buschart*, et al.* 2017, Lagkouvardos*, et al.* 2014). Numerous bioinformatics tools play a major role in such integrated multi-omics studies. In Table 1, a list of key software tools that are widely used for microbiome research is provided. The research and development of software is a constantly developing field in microbiome research. The focus is mostly on improving accuracy and predictive power for both taxonomic classifications and protein annotations. Most attention has been towards the development of tools for analysing marker-gene and shotgun metagenomics sequencing data. There are few alternatives for analysis of metatranscriptomics and metaproteomics data. All the tools have primarily been developed for short read sequences obtained by the widely used Illumina® sequencing platform. Recently, long read sequencing facilitated by PacBio® Single Molecule, Real-Time (SMRT) and Nanopore® sequencing are becoming increasingly popular (Schadt*, et al.* 2010, Clarke*, et al.* 2009). The long-read technology is expected to improve the marker gene survey and metagenomic assemblies with longer read length. However, these techniques currently have higher error rates (Weirather*, et al.* 2017). Recently, millions of full length 16S rRNA gene sequences were obtained using the nanopore sequencing technology (Karst*, et al.* 2018). These sequences are an important addition to the sequence databases which are dominated by shorter marker gene sequences. However, the same study also identified an error rate of 0.17%, suggesting scope for improvement. In addition to improvement in sequencing chemistry, there is scope for development of user-friendly interactive tools to simplify data analysis by end-users who are not trained in bioinformatics. Future research and development of innovative bioinformatics pipelines and work flows incorporating data provenance and reproducibility will play a major role in understanding the microbial metabolic interactions and their implications for human health.

Interplay between microbes and between microbes and host

The complex interplay between the microbes and the host takes place at both physical and biochemical level (Curtis and Sperandio, 2011; Dorrestein *et al.*, 2014). The physical attachment (adhesion) of bacteria such as *Lactobacillus*, to the host cell surface is mediated *via* pilli (LPXTG, Fimbriae type) (Kankainen *et al.*, 2009; von Ossowski *et al.*, 2011; Turner *et al.*, 2003). In addition to adhesion, bacterial surface molecules, when in contact with immune cells can lead to their activation (Atarashi *et al.*, 2013; Kosiewicz *et al.*, 2014). The host immune cells interact with microbes *via* pattern recognition receptor (PRR) families, for example Toll-like receptors (Kelly *et al.*, 2005). This also gives the bacteria an opportunity to modulate the immune system by increasing or decreasing the expression of cytokines (Ottman *et al.*, 2017). In the colon, the anaerobic fermentative lifestyle is a major characteristic of bacteria. Fermentation end products such as short-chain fatty acids (SCFAs), namely, acetate, lactate, butyrate, propionate, formate and valerate, medium chain fatty acids like caproate, and branched chain fatty acids such as iso-butyrate and isovalerate, play a crucial role in normal host physiology (Flint *et al.*, 2012). Especially the SCFAs butyrate and propionate serve as energy sources for colonocytes, regulators of gene expression *via* histone deacetylase inhibition and cell differentiation, and anti-inflammatory agents (Davie, 2003; Donohoe *et al.*, 2011; Hamer *et al.*, 2008; Hosseini *et al.*, 2011; Lin *et al.*, 2012; Reichardt *et al.*, 2014). Apart from SCFAs, numerous other small molecule metabolites are actively interacting with the host cells, affecting host physiology. Some key metabolites are listed in Table 2 (Kinross *et al.*, 2011; Nicholson *et al.*, 2012; Sung *et al.*, 2017).

Disentangling the complex interactions between host and microbes and between different microbes is difficult. Investigation of the interaction between the host and microbes has to be done at different levels of complexity ranging from a single strain to entire community level interactions. Such studies will require cultivation of representative bacteria from the human intestinal microbiome in laboratory conditions. In a recent example of identifying interactions between intestinal bacteria

and the host immune system, a collection of cultivated clostridia were shown to interact with T-regulatory cells (Atarashi *et al.*, 2013).

Table 2: Small molecules in the human intestinal tract. The original table is from the review titled "Small molecules from the human microbiota" by Donia MS and Fischbach MA (Donia and Fischbach 2015). Here, we show only those small molecules that are present in the intestinal tract.

Class	Small molecule compound	Producer (example)	Known/predicted activity
RiPP (lantibiotic)	ruminococcin A	*R. gnavus*	antibiotic
RiPP (bacteriocin)	ruminococcin C	*R. gnavus*	antibiotic
RiPP (microcin)	microcin C7/C51	*E. coli*	antibiotic
RiPP (microcin)	microcin B17	*E. coli*	antibiotic
RiPP (microcin)	microcin J25	*E. coli*	antibiotic
RiPP (microcin)	microcin H47	*E. coli*	antibiotic
RiPP (TOMM)	clostridiolysin S	*C. sporogenes*	unknown
RiPP (TOMM)	listeriolysin S	*L. monocytogenes*	unknown
RiPP	heat-stable enterotoxin	*E. coli*	GI motility (guanylate cyclase 2C)
Amino acid metabolite	indolepropionic acid	*C. sporogenes*	immunomodulatory
Amino acid metabolite	indole	unknown	converted to indoxyl sulfate
Amino acid metabolite	skatole	*Clostridium* spp.	unknown
Amino acid metabolite	tryptamine	*R. gnavus*	neurotransmitter
Amino acid metabolite	phenyllactic acid	*Bifidobacterium* spp.	unknown
Amino acid metabolite	phenethylamine	*Lactobacillus* spp.	neurotransmitter
Amino acid metabolite	δ-aminovaleric acid	*Clostridium* spp.	unknown
Amino acid metabolite	GABA	unknown	unknown
Amino acid metabolite	α-aminobutyric acid	unknown	unknown
Amino acid metabolite	3-aminoisobutyric acid	*Clostridium* spp.	unknown
Amino acid metabolite	p-cresol	*Clostridium* spp.	unknown
Acid (short-chain)	propionic acid	*Bacteroides* spp.	immunomodulatory (GPR43)
Oligosaccharide	polysaccharide A	*B. fragilis*	immunomodulatory (TLR2)
Glycolipid	α-galactosylceramide	*B. fragilis*	immunomodulatory (CD1d)
Glycolipid	corynomycolic acid	*Corynebacterium* spp.	unknown
Terpenoid	bile acids (e.g., deoxycholic acid)	*Clostridium* spp.	metabomodulatory [TGR5, farnesoid X receptor (FXR), VDR]
NRP-PK	colibactin	*E. coli*	cytotoxic
Amine oxide	Trimethylamine N-oxide	*Anaerococcus hydrogenalis, Clostridium*	atherogenic

		asparagiforme, C. hathewayi, C. sporogenes, Escherichia fergusonii, Proteus penneri, Providencia rettgeri, and *Edwardsiella tarda*	
Sulfur compounds	H_2S	*Desulfovibrio*	colonic inflammation and cancer

In another recent study, mono-colonisation of mice with 53 commonly observed and phylogenetically diverse intestinal bacteria was used to investigate their immunomodulatory effect (Geva-Zatorsky *et al.*, 2017). These studies highlight the need for cultivation efforts to capture all the representative bacteria form the human host and investigate their influence on host health.

Characterizing the physiology of microbes

Traditional microbiology is largely dependent on culturing bacteria in laboratory conditions. The approaches most widely used included enrichment cultivation under selective conditions, and dilution to extinction to obtain isolated colonies by spread plating on agar and/-or gelrite plates with artificial media. Several of these approaches have been used to culture bacteria from the human colon. The human colon is anoxic (devoid of oxygen), except close to the mucosal layer where there can be diffusion of small amounts of oxygen (Espey, 2013; Khan *et al.*, 2012; Ouwerkerk *et al.*, 2016). The natural anoxic environment in the colon is challenging to replicate under laboratory conditions. Pioneering studies in late 1970s and 80s by Holdeman and Moore resulted in isolation of many anaerobic bacteria from the human intestinal tract, which remain a major resource for present day research (Holdeman and Moore, 1974; Holdeman *et al.*, 1984; Moore and Holdeman, 1974). In one of their studies, a total of 1,147 bacteria were isolated and classified based on morphology into 113 "distinct types". These included bacteria belonging to the genera *Bacteroides* (some now reclassified as *Prevotella* and *Dorea*), *Eubacterium*, *Ruminococcus*, and *Coprococcus*. However, since the description of these bacteria commonly observed in the human intestinal tract, only few have been further characterized in detail with respect to their physiology. For

example, *Rumminoccus bromii* is now known as a prominent resistant starch degrader and *Coprococcus eutactus* is known as butyrate and propionate producer (Holdeman and Moore, 1974; Reichardt *et al.*, 2014; Ze *et al.*, 2012). More than 30 years later, a study focused on novel butyrate producing bacteria, some of which were similar phylogenetically to those isolated by Holdeman and Moore (1974) from human faeces (Barcenilla *et al.*, 2000). Recent large-scale targeted and non-targeted cultivation approaches, often referred to as culturomics, have succeeded in culturing a high diversity of novel bacteria from human faeces (Browne *et al.*, 2016; Lagier *et al.*, 2012, 2016; Rettedal *et al.*, 2014).

Integrating multi-disciplinary approaches is a crucial aspect of improving our knowledge about the physiology, interaction networks and role of intestinal bacteria in human health (Figure 1).
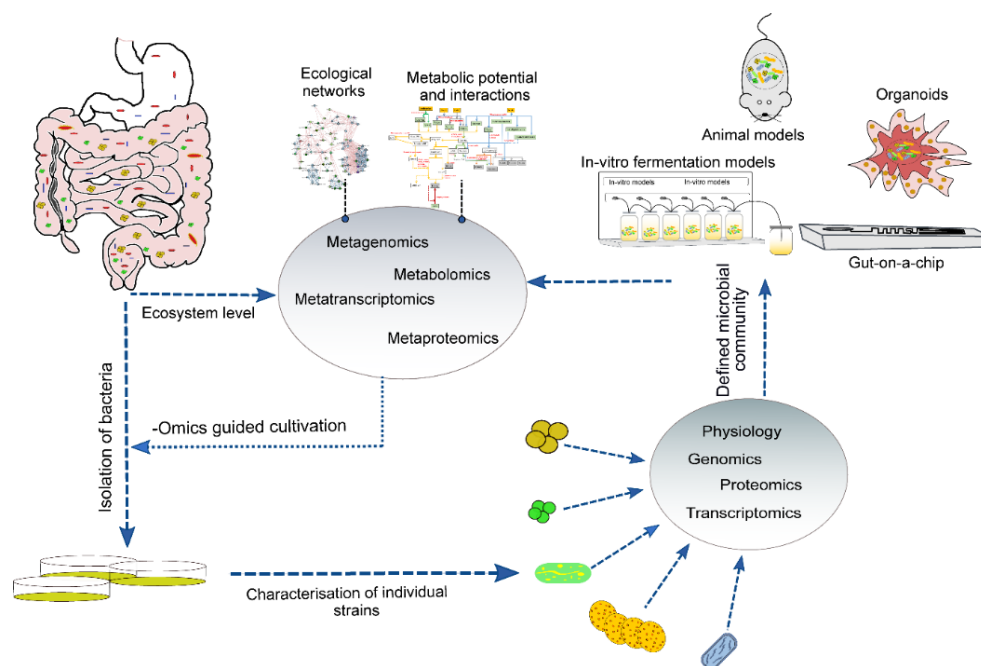
Figure 1: Synergistic approach to understanding individual to ecosystem level microbial interactions and their impact on the host.

The availability of large population level data on microbial profiling employing standard methodology provides an important avenue to search for recurring patterns and common members of the intestinal microbiome. Re-analysis of microbial community profiles from the human intestinal tract atlas of >1000 western adults revealed 34 genus-like groups that are part of the core microbiota in healthy Western adults (Shetty

et al., 2017). However, many of these have not been studied in detail for understanding their physiology, and some microorganisms like *Oscillospira* remain elusive in culture (Shetty *et al.*, 2017). Nowadays, attempts to culture novel bacteria can benefit from available multi-omics approaches (Gutleben*, et al.* 2018). Assembly of genomes of yet to be cultured bacteria from shotgun metagenomic data sets can give insights into the potential metabolic features aiding in designing targeted cultivation stratergies. Additionally, metatranscriptomics can give insights into the active functions and has been successfully used to culture novel bacteria.For example, a novel *Rikenella*-like bacterium was isolated using metatranscriptomics-derived evidence for the requirement of sulfated- and sialated-mucin glycans for growth of the bacterium (Bomar *et al.*, 2011). In addition, genomic and proteomic analysis of already cultured bacteria can provide novel insights into their physiology. For example, a novel pathway for breakdown of fructolysine to produce butyrate was uncovered in *Intestinimonas butyriciproducens* using a combination of classical microbiology, genomics, proteomics and metabolomics (Bui *et al.*, 2015). Such multifaceted approaches have allowed to gain comprehensive knowledge of a bacterium's life style.

Towards a representative minimal microbiome of the human intestinal tract

It is important to acknowledge that intestinal bacteria are not independently growing free-living entities. Information obtained from investigation of a given bacterium in isolation may not represent its natural lifestyle. Therefore, it will be crucial to study bacteria as communities by growing multiple bacterial species together. One approach is to grow bacteria in combination as co-, tri- or even more complex cultures, building up to create a consortium of bacteria that could be representative of a functioning minimal microbial community of the human intestinal tract. Such a minimal microbiome can be used to study several aspects of the host-microbiome and within microbiome interactions (Brugiroux *et al.*, 2016; Desai *et al.*, 2016; Stecher *et al.*, 2007). When designing a minimal microbiome, it is important to consider major factors influencing the intestinal microbiome. For example, diet is a major source of carbon and energy and other nutrients required for growth for intestinal bacteria, alongside host derived compounds such as mucus. Diet, especially components that cannot be

digested by the host, influence the composition and metabolic (fermentation) activity of the microbiome (Salonen *et al.,* 2014; Sonnenburg *et al*., 2016; Voreades *et al.,* 2014). There are several steps involved in degradation of various carbon sources, and different bacteria carry out different steps (Scott*, et al.* 2014, Flint*, et al.* 2012c). There is high functional redundancy, especially with regards to the butyrate producers using monosaccharides. This leads to competition for resources in the intestinal microbiome (Coyte *et al.*, 2015). A general overview of the ecophysiological interactions in the human intestinal ecosystem are depicted in Figure 2. Specialist bacteria capable of degrading complex dietary fibres and mucus are key players in the community. Known examples of such bacteria are *Ruminococcus bromii*, *Eubacterium rectale* and *Bacteroides thetaiotaomicron* capable of degrading complex polysaccharides and *Akkermansia muciniphila*, *Barnesiella intestinihominis* and *Bacteroides caccae* that are capable of degrading mucus (Derrien *et al*., 2004; Desai *et al*., 2016; Ravcheev *et al.*, 2013; Xu *et al*., 2003; Ze *et al*., 2012). These bacteria break down complex energy sources and provide simple carbohydrates for other members in the community. An experimentally proven metabolic interaction network is that comprising *A. muciniphila* and butyrate producers *Anaerostipes caccae, Eubacterium hallii,* and *Faecalibacterium prausnitzii* (Belzer*, et al.* 2017). The butyrate producers benefitted from simple sugars released from mucus by *A*. *muciniphila*, and in return *A*. *muciniphila* benefitted from the *E. hallii*-mediated production of Vitamin B12, an important co-factor. This example demonstrates the applicability of knowledge regarding fermentation and physiology of bacteria in designing a minimal bacterial consortium to give insights into syntrophic interactions. Ecological concepts such as stability, competition, co-existence, alternative stable states, as well as within host evolution under internal and external selective pressures can be studied (Großkopf and Soyer, 2014; Mee *et al*., 2014; De Roy *et al*., 2014). Using reductionist approaches, concepts of pathogen colonisation resistance, phage–bacterial host dynamics, bacteria-archaea-host dynamics, spatial organization and small molecule interactions have been investigated in *in-vivo* mouse models and *in-vitro* models (Brugiroux *et al*., 2016; Desai *et al*., 2016; Hamer *et al*., 2008; Reyes *et al*., 2013; Samuel and Gordon, 2006).
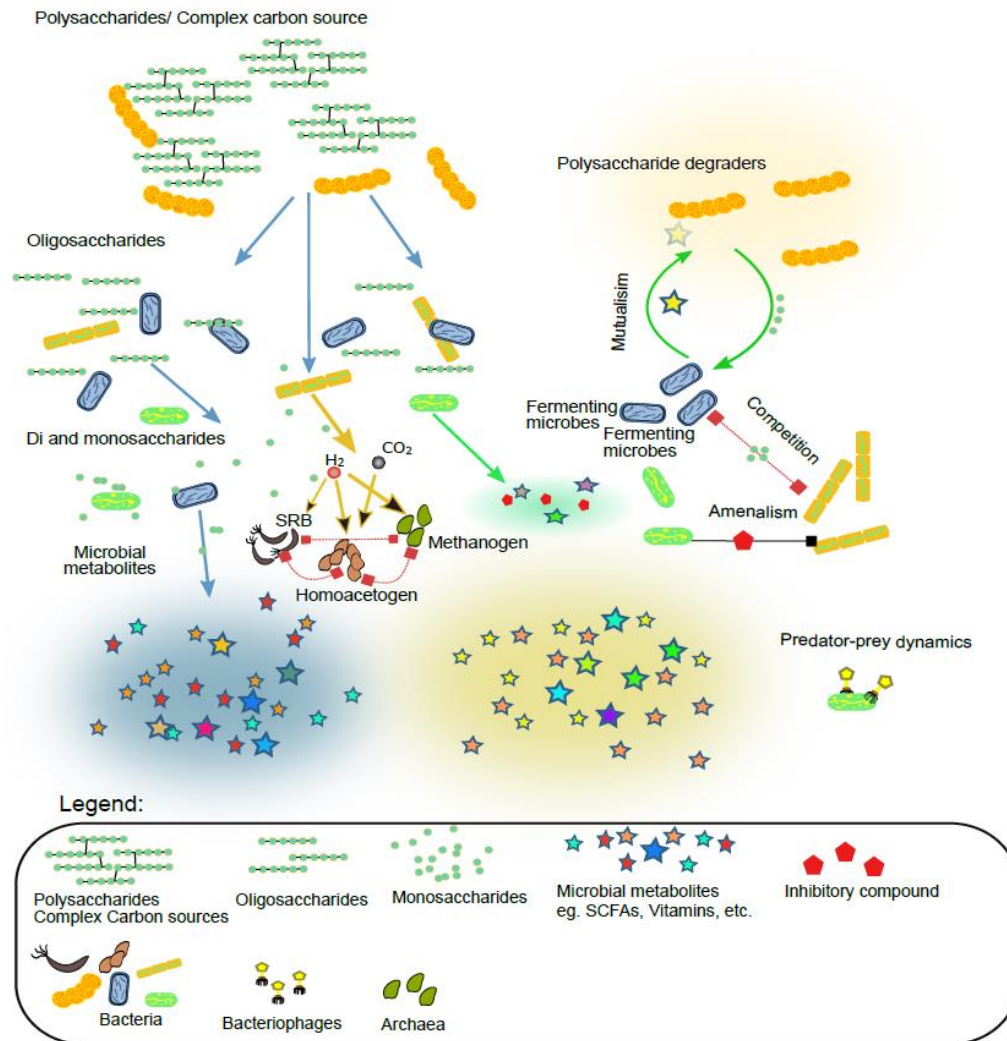
Figure 2: A simplified view of the ecophysiological interactions in the human gut ecosystem. The trophic interactions resulting from degradation of complex polysaccharides by primary degraders are an important aspect of the commensalism or mutualism observed in the microbiome. Several other interactions such as competition for breakdown products, vitamins and co-factors can be observed. Inhibition of growth by bacterial primary and secondary metabolites such as organic acids and antibiotics can lead to amensalism. Bacteriophages also influence the dynamics of bacterial communities (predatory-prey dynamics).

A common theme in the majority of the studies with respect to the selection of bacteria is a focus on prevalence of the bacteria in large populations. However, novel approaches such as genome-guided selection of bacteria representative of the complex human intestinal microbiota has also been reported (Brugiroux *et al.*, 2016). Using metagenomic approaches to identify potential functions of a complex microbiome and then analysing genomes of representative bacteria from the murine intestine, a selection of 12 bacteria was made. This included bacteria from five phyla, Firmicutes, Bacteroidetes, Actinobacteria, Verrucomicrobia and Proteobacteria and covered 66.6%

of the KEGG modules of the conventional murine intestinal microbiome (Brugiroux *et al.*, 2016).

Another approach is to use an ecophysiology guided design of a stable, resilient and functioning minimal microbiome. Ecophysiology guided design can be defined as an approach which integrates ecological awareness of co-occurrence, co-exclusion, cooperation, the concept of key stone species and the knowledge of physiology of diverse bacteria and their metabolic complementarity and competition. For example, using an ecological concept that employs network theory to infer co-occurrence and co-exclusion can give insights regarding the common partners for a specific bacterium (Faust and Raes, 2012; Leibold *et al.*, 2004). Also, key-species in the interaction network can be identified, and modularity of the ecosystem network can give insights into sub-networks in the metacommunity (Faust and Raes, 2012; Faust *et al.*, 2012; Fisher and Mehta, 2014; Trosvik and de Muinck, 2015). Incorporating genome level information and evidence for physiology from laboratory experiments will further aid in developing a potentially stable and resilient minimal microbiome.

In summary, molecular tools have empowered us with large amounts of data regarding compositional and functional attributes of intestinal microbiomes in health and disease. These provide an opportunity for narrowing the microbial targets for further mechanistic studies in order to generate causal information regarding their role in health and disease associations. Several bacteria from the intestinal tract have been cultivated in laboratory conditions. However, there is a need to investigate their physiology to better understand their role in the human intestinal microbiome. Reducing the complexity of a community with trillions of microbes consisting of more than 2000 species like groups is crucial for investigating various aspects of microbiome functionality. Design of representative minimal intestinal microbiomes using ecophysiology guided approaches can be a useful resource for better understanding the features of a complex microbiome.

Research aims and thesis outline

The research described in this thesis aims to integrate community level ecological insights and physiology of individual bacteria to have a better understanding of the human intestinal microbiome. To this end, this thesis provides insights into the ecosystem level understanding of the human intestinal microbiome, develops tools for analysis of microbial community profiling data, uncovers the physiology of key intestinal bacteria and presents data on the design of a minimal microbial community as a resource to understand microbial interaction networks.

Chapter 1 briefly introduces the readers to the key aspects of the human intestinal microbiome, with specific focus on the topics and the motivation for the work carried out during this thesis. **Chapter 2** highlights key ecological concepts pertaining to the intestinal microbiome. Meta-analysis of microbial profiling data of more than 1000 Western adults identifies the core microbiota, and genomic analysis identifies similarities and differences between the representative core bacteria. We discuss the ecological concepts of alternative stable states, community types, diversity and their potential for developing microbiome based diagnostics and therapeutics. **Chapter 3** reports an intriguing insight into the microbial community structure where hyperdominance seems to be norm rather than an exception. Using Earth Microbiome Project data, we show that less than 4% of the bacteria in a given ecosystem account for 50% of the entire community. We discuss the possible mechanisms giving rise to the observed hyperdominance in human intestinal microbial communities. The ecosystem level understanding of characteristic features of microbial communities is important when assembling a minimal microbial community. **Chapter 4** addresses an important technical aspect of investigating microbial communities, *i.e.* data analysis. We created a suite of R software environment-based packages, including a web-based user-friendly shiny app. The microbiome and microbiomeutilities R packages extend and complement the widely used phyloseq R package. The microbiomeutilities R package provides functionality for preliminary quality control, pre-processing and generates and HTML report for non-experienced users. We also address the challenge of learning R as beginners, by providing a web-based tool for doing basic analysis of

microbial community profiles. These are particularly useful for analysing defined bacterial cultures or enrichments where the user simply wishes to determine the microbial composition and associated changes. One observation we made in the early chapters of this thesis, was the lack of physiological understanding of many bacteria that are part of the core microbiota. In **chapter 5**, we move from ecological approaches to ecophysiological approaches where we integrate multiple techniques to have a better understanding of some key bacteria in the human intestinal tract. In a collaborative work, we elucidate the ecophysiology of an important lysine degrading, butyrate producing bacterium, *Intestinimonas butyriciproducens*. We employ phylogenetics and genomics guided physiological comparison of two *I. butyriciproducens* strains of mouse and human origin. In **chapter 6**, we focus on a potential next-generation biotherapeutic bacterium, *Eubacterium hallii*, which we observed to be a part of the core microbiota. The 16S rRNA gene and genome comparison of *E. hallii* DSM3353^T and *E. hallii* L2-7 (DSM17630) revealed misclassification of this medically important bacterium. Genome analysis supported the notion that L2-7 is potentially a novel species. This encouraged us to characterize the two strains using a polyphasic approach to taxonomically re-classify these strains as members of a new genus, *Anaerobutyricum*, with the strain DSM3353^T as the type species *A. hallii*, and strain L2-7 as novel species named *Anaerobutyricum soehngenii*. In **chapter 7**, we study in detail the ecophysiology of *A. soehngenii* using genomics, proteomics and *in-vitro* testing of substrate preferences. We identify the active genes involved in the pathways for butyrate synthesis and 1,2 propanediol utilization and *de novo* and salvage pathway for vitamin B12 biosynthesis. We identified a unique organisation of genes involved in DL- lactate utilization that is found in few intestinal bacteria. In **chapter 8**, we use an ecophysiology guided design of a representative human intestinal minimal microbiome. We use a combination of *in-silico* and *in-vitro* experiments to elucidate the metabolic interactions between key bacteria in the human intestinal microbiome. In **chapter 9**, I summarize and discuss the observations of the research data presented in this thesis and provide potential implications of these findings for human intestinal microbiome research.

Chapter 2

Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies

Sudarshan A. Shetty, Floor Hugenholtz, Leo Lahti, Hauke Smidt and

Willem M. de Vos
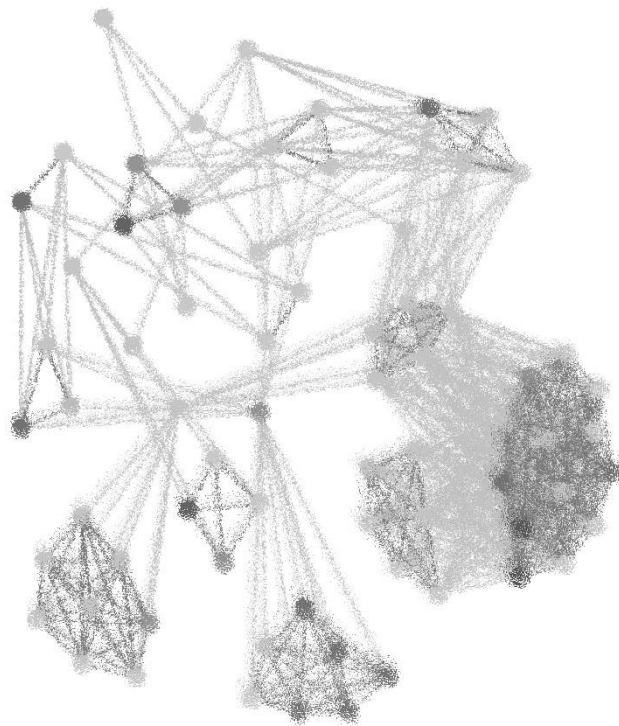
Abstract

High individuality, large complexity and limited understanding of the mechanisms underlying human intestinal microbiome function remain the major challenges for designing beneficial modulation strategies. Exemplified by the analysis of intestinal bacteria in a thousand Western adults, we discuss key concepts of the human intestinal microbiome landscape, i.e. the compositional and functional 'core', the presence of community types and the existence of alternative stable states. Genomic investigation of core taxa revealed functional redundancy, which is expected to stabilize the ecosystem, as well as taxa with specialized functions that have the potential to shape the microbiome landscape. The contrast between Prevotella- and Bacteroides-dominated systems has been well described. However, less known is the effect of not so abundant bacteria, for example, Dialister spp. that have been proposed to exhibit distinct bistable dynamics. Studies employing time-series analysis have highlighted the dynamical variation in the microbiome landscape with and without the effect of defined perturbations, such as the use of antibiotics or dietary changes. We incorporate ecosystem-level observations of the human intestinal microbiota and its keystone species to suggest avenues for designing microbiome modulation strategies to improve host health.

Keywords: Alternative stable states, Bistability, Core microbiota, Early warning signals, Landscape model, Tipping elements

Introduction

Microbial cells may outnumber our own cells in the body, presenting an example of nature's intricate evolutionary scheme of host–microbe symbiosis (Frank and Pace 2008, Ley*, et al.* 2008, Sender*, et al.* 2016). This unique setup of human– microbe interactions and its resilience is of paramount importance for health. Starting from birth, various factors shape the composition and function of the human intestinal tract micro- biome, defined as the entire habitat including all microbes, their genomes and surrounding environment (see Table 1) (Sekirov*, et al.* 2010, Ottman*, et al.* 2012).

The human intestinal tract microbiota consists mainly of Bacteria, but also includes Archaea, microeukaryotes and many viruses, mainly bacteriophages. The dominant bacterial phyla are Bacteroidetes and Firmicutes, while Actinobacteria, Proteobacteria and Verrucomicrobia constitute minor phyla (Li*, et al.* 2014, Falony*, et al.* 2016). Archaeal representatives consist largely of methanogens, among which *Methanobrevibacter* and *Methanosphaera* are the most prevalent genera (Dridi*, et al.* 2009). The microeukaryotes commonly observed in the human intestinal tract include *Blastocystis* spp. and fungal genera, such as *Candida* spp., *Gloetenia*/Paecilomyces and *Galactomyces*. Finally, various reports have detailed the presence of bacteriophages in the human intestinal tract (Minot*, et al.* 2013, Waller*, et al.* 2014). A recent deep genomic analysis of virus-like particles present in human fecal samples revealed that most bacteriophages be- long to the Caudovirales order (Podovridae, Siphiridae and Myoviridae), some to the Microviridae and no eukaryotic viral sequences were detected (Manrique*, et al.* 2016). The latter study revealed 64 healthy subjects to harbor a total of 44 bacteriophage groups that included a set of core bacteriophages shared among over half of the subjects, a common set that was present in 20%– 50% of the subjects, and a set of bacteriophages that are either rarely shared or unique to an individual. While there are indications that bacteriophages may be involved in controlling the intestinal microbiota in health and disease, additional deep studies are needed to assess how these phages are involved in the host–microbe interactions (Reyes*, et al.* 2010, Reyes*, et al.* 2015, Manrique*, et al.* 2016).

Interactions between the host, the prokaryotic and eukaryotic microbes and their viruses take place at various levels. While there are marked differences in microbial composition and function along the entire length of the intestinal tract, most research has focused on the fecal microbiota as fecal sampling is a cost-efficient and non-invasive method. However, the com- positional and functional attributes of the microbiota vary along the intestinal tract—most notably, the small intestine harbors a unique microbial ecosystem that has been characterized extensively (Zoetendal*, et al.* 2012, Zoetendal and de Vos 2014). More- over, the fecal microbiota may differ from that in the colonic mucosa but systematic studies are lacking because of the invasiveness of the sampling (Zoetendal*, et al.* 2002, Donaldson*, et al.* 2016). As bacteria are the dominant group in this ecosystem, they have received most attention in the recent studies (Lozupone*, et al.* 2012, Li*, et al.* 2014, Falony*, et al.* 2016). After birth, the infant intestinal tract becomes colonized by bacteria largely gained from the mother (Favier*, et al.* 2003, Koren*, et al.* 2012). The programmed outgrowth, manifested as succession, and the composition of the adult intestinal microbiota depend on a variety of deterministic host factors, including age, diet, genetic-make-up, immune and health status, and gender and geographical location (Biagi*, et al.* 2010, De Filippo*, et al.* 2010, Claesson*, et al.* 2012, Yatsunenko*, et al.* 2012).

Furthermore, stochastic factors such as colonization order or antibiotic exposure can also affect microbial colonization (Lederberg and McCray 2001, Harris*, et al.* 2014, Korpela*, et al.* 2016). Overall, this process leads to the development of a dynamic and complex web of interactions between microorganisms, the host and the environment (Costello*, et al.* 2012, Ottman*, et al.* 2012). Knowledge of the deterministic and stochastic factors affecting the host–microbiota homeostasis could help in defining the factors responsible for evolutionary selection and establishment of the symbiosis between human and microbes.

Bacteria form the dominant domain in the large intestinal microbiota and hence we focus here on the accumulated knowledge of this group. For convenience, the definition of technical terminologies used in this review is given in Table 1. We discuss the concepts of the bacterial compositional and functional core, community types and alternative stable states of the microbial ecosystem. We will further discuss how

improved understanding of these features could be instrumental in microbial diagnostics, predicting changes or homeostasis and determining the efficacy of intervention strategies for microbiome modulation.

The Common Core Microbiome in The Human Large Intestine

A major goal in human microbiome studies is to identify and characterize the bacterial taxa and functions that are shared by the majority of individuals. Viewing the bacterial community from a taxonomic perspective reveals an increasing level of complexity toward the lower taxonomic ranks from phylum level to species level. Moreover, the intestinal bacteria interact with each other and the host at various levels from single cells to populations of specific bacteria to the entire community. The complexity of such multi-scale interactions presents a major challenge in understanding community functioning and its impact on host health (Greenblum*, et al.* 2012, Greenblum*, et al.* 2013). One way to approach this challenge is to focus on the most abundant bacterial species and specifically the prevalent ones that are shared by many individuals.

Coevolution of host and microbiome can be explained by ecological selection processes (Pianka 1970, Ley*, et al.* 2006). Host factors such as genetics, immune regulation and age represent top–down selection that plays a role in shaping the microbial community. In contrast, bottom–up selection results in the growth of microbes that have specialized functions, such as those associated with degradation of specific dietary components. Another concept is the r/K selection where organisms that have adapted to maximize the rate of growth are so-called r-strategists, whereas those adapted to compete and survive when resources are limited are known as K-strategists (Pianka 1970).

However, applying these idealized concepts to the highly complex, dense and dynamical intestinal ecosystem is challenging (Freilich*, et al.* 2010). The large intestine is known to have spatial heterogeneity in terms of bacterial species which occupy microhabitats (Donaldson*, et al.* 2016). For example, the outer mucus layer is densely populated compared to the inner mucus layer (Johansson*, et al.* 2008). The presence of micro- habitats with variable population densities suggests that, rather than being

present at one of the two endpoints, the bacterial community should be viewed along the r-K continuum that is governed by top–down and bottom–up selection processes.

Table 1. Terms and definitions used in this review.

Terminologies	Definitions	Reference
Active functional core	The common functions that are translated and translated actively across human population and have an effect on the whole ecosystem	Kolmeder et al. (2012)
Alternative stable states	Resilient community states that are stable and resist change over ecologically relevant timescales	Fukami and Nakajima (2011)
Alternative transient states	The community states that are not necessarily stable and vary in structure and/or function	Fukami and Nakajima (2011)
Bacterial phylogenetic core	The assemblage of phylogenetically related bacterial taxa in the human intestinal microbiome	Rajilić-Stojanović et al. (2009); Tap et al. (2009)
Bimodal bacteria	Bacteria observed to be present in either high or low abundances within a given population of individuals. The states may be driven by external factors and are not necessarily resilient to changes	Lahti et al. (2014)
Bistable bacteria	Bacteria that have two alternative stable states of low and high abundance, and unlike bimodal bacteria, the states are able to resist change	Faith et al. (2013); Lahti et al. (2014)
Common core microbiota	Community of microbes and their functions that are shared in majority of humans	Qin et al. (2010)
Deterministic effect	Assembly of communities in a deterministic manner which is a result of environmental factors and interactions between community members	Fukami and Nakajima (2011)
Human intestinal microbiome	The entire habitat (all microbes, their genomes and surrounding environment) in the human	Marchesi and Ravel (2015)
Microbiota	The assemblage (collection and combination) of microorganisms present in a defined environment	Marchesi and Ravel (2015)
Minimal intestinal metagenome	The catalog of functions (coding capacity as assessed by metagenomics) involved in the homeostasis of the whole ecosystem, encoded across many species	Qin et al. (2010)
Resilience	The capacity of a community to return to the original state after perturbation	Paine, Tegner and Johnson (1998); Walker et al. (2004)

Stability	Tendency of the community to maintain a state of homeostasis and resist disturbances or to show resilience after disturbance	Scheffer et al. (2001)
Stochastic effect	Assembly of communities as a result of unpredictable disturbance, dispersal or birth–death events	Chase and Myers (2011)

The aforementioned factors result in the selection of bacteria belonging to specific phylogenetic clusters with adaptive and functional features characteristic for the ecosystem. This coexisting community of microbes and their functions, which are characterized by the minimal intestinal metagenome, have co-evolved with the host for mutual benefits and are prevalent in the population, constitutes the common core (Jalanka-Tuovinen*, et al.* 2011, Li*, et al.* 2014). The core microbiota may include specific keystone species that are important for maintaining an efficiently functioning ecosystem, and whose gain or loss may have profound influence on ecosystem structure and function through their effect on other community members (Paine 1966). In the intestinal tract ecosystem, examples of keystone species include bacteria that are part of the common core. Their functional capabilities include for instance breakdown of complex carbon sources to support the growth of the other core members (Ze*, et al.* 2012, Trosvik and Muinck 2015). In the following section, we will discuss both the phylogenetic core and the minimal intestinal metagenome.

Phylogenetic Core

Several studies focused on defining a human intestinal core microbiota in terms of phylogenetic composition (Hamady and Knight 2009). Longitudinal surveys of fecal bacterial 16S ribosomal RNA (rRNA) sequences in individuals for extended periods of time of up to more than 10 years have revealed that a significant fraction of bacterial phylotypes is continuously present in the large intestine indicating that it comprises a stable individual core in healthy adults (Faith*, et al.* 2013, Rajilić-Stojanović*, et al.* 2013). Cross-sectional comparisons of the fecal microbiota in a large population cohort have indicated the presence of a prevalent core microbiota, which comprises those bacterial taxa that can be detected in the majority of individuals (Zoetendal*, et al.* 2008). In contrast, the individual core microbiota includes those taxa that are frequently detected within a particular individual; this can overlap with the common core, but also include individual-specific features.

Several studies have aimed to characterize the composition of the common core but their comparison is difficult since different criteria were used (Table 2).

Table 2. Overview of studies that aimed at identifying the compositional and functional core.

Sr. No.	Core Compositional	Core Functional[¶]	Number of subjects	Criterion—other comments	Study/ year
1	No	Yes	154	Presence of OTU in all individuals—both obese and lean subjects included	(Turnbaugh, *et al.* 2009)
2	66 OTUs	No	35	Presence of OTU in more than 50% individuals	(Tap, *et al.* 2009)
3	15 genus-like groups	No	10	The study used phylogenetic microarray for the first time to demonstrate the common core	(Rajilić-Stojanović, *et al.* 2009)
4	19	No	4	Low number of subjects	(Claesson, *et al.* 2009)
5	406 in elderly subjects (at least 65 yr). 18 in subjects between 18 and 58 yr	No	127	Criterion was presence of OTU in more than 50% individuals. Included ageing subjects	(Claesson, *et al.* 2011)
6	6% (V1–V3) to 8% (V3–V5)	No	327	Criterion was presence of OTU in at least 95% of individuals	(Huse, *et al.* 2012)
7	Not studied	Yes	3	Criterion was proteome features present in all subjects	(Kolmeder, *et al.* 2012)
8	Variable	No	104	Depending on the cut-offs selected for ubiquity and abundance	(Li, *et al.* 2013)
9	75 species	294 110 genes	124	The most common species were those present in ≥90% of individuals with genome coverage >1%	(Qin, *et al.* 2010)

				Common genes are those present in >50%	
10	290 phylotypes	No	115	Core size depends on depth of analysis. Present in all individuals (phylogenetic microarray technology)	(Salonen, *et al.* 2012)
10	288 phylotypes	No	9	Present in all individuals at a detection threshold of 0.03% (phylogenetic microarray technology)	(Jalanka-Tuovinen, *et al.* 2011)
11	22 OTUs	No	64	Observed at an average frequency of occurrence higher than 90%	(Zhang, *et al.* 2014)
12	43 OTUs	No	20	Present in 15 out of 20 individuals	(Nam, *et al.* 2011)
13	9 genera, 30 OTUs	No	314	9 genera present in all and at OTU level shared by at least 90%	(Zhang, *et al.* 2015)
14	17 genera	No	4000 (Dutch, Belgian, UK and USA)	Prevalence threshold was 95%. At varying low mean abundance, the number of core genera detected was higher	(Falony, *et al.* 2016)
	14 genera		4000 plus 308 samples from Papua New Guinea Peru, and Tanzania		
	35 genera		In 1106 individuals from the Belgian Flemish Gut Flora Project cohort		

The discrepancy in categorizing the common core microbiota could be partially attributed to methodological differences in determining microbiota composition and defining the core. Importantly, the depth of sequencing and analysis, number of samples and taxonomic resolution are the factors that affect the analysis. The aforementioned issues pertaining to the analytical definition of core microbiota have been addressed by using the HITChip (Human Intestinal Tract Chip), a phylogenetic microarray targeting 16S rRNA gene sequences of bacterial taxa that are reported to occur in the human intestinal tract (Salonen*, et al.* 2012). This phylogenetic microarray represents a measurement platform with high reproducibility, robustness and dynamic range, allowing the incorporation of both abundant and rare taxa in microbiota profiling.

However, while capturing the vast majority of intestinal microbial species, these and other phylogenetic microarrays lack the ability to detect novel taxa that can be assessed by next-generation sequencing techniques (Hazen*, et al.* 2010, Roh*, et al.* 2010, Tottey*, et al.* 2013). Previously, the HITChip-based analysis of >1000 rare and abundant bacterial phylotypes from 130 genus-level groups in 115 adults revealed a common core of 290 phylotypes (Jalanka-Tuovinen*, et al.* 2011, Salonen*, et al.* 2012). We extended this methodology to analyze the common core in a cohort of 1006 non-compromised Western adults (Lahti*, et al.* 2014). This provided a global view of the common core that revealed bacteria related to *Faecalibacterium prausnitzii*, *Oscillospira guillermondii* and *Ruminococcus obeum* as the top three taxa shared by all adults (Figure S1, Supporting Information). Previously, we reported that the method used for DNA extraction affected the microbiome analysis and that mechanical extraction is to be preferred, as further confirmed by the International Human Microbiome Standards analysis (Salonen*, et al.* 2010, Santiago*, et al.* 2014). Hence, we separately analyzed the samples processed using a mechanical or enzymatic DNA extraction protocol. While no major difference in the inclusion or exclusion of major core bacteria was observed, differences in their ranked cumulative fractional abundances were observed (Figure S2, Supporting Information). Limiting our analysis to the subset of 401 samples obtained by mechanical extraction using repeated bead-beating revealed that bacteria related to *F. prausnitzii*, *R. obeum* and *Subdoligranulum*

variabile to be the top three taxa shared by all adults. Moreover, we found that 34 out of the 130 genus-like groups targeted by the HITChip are shared in 95% of the subjects at the minimum detection threshold of 0.1% relative abundance, with the highest prevalence being observed for bacteria belonging to the order Clostridiales (Figure 1; see Supporting Information for methods). This estimate is similar in number to the recently published data by the Belgian Flemish Gut Flora Project (FGFP) on 1106 individuals at 95% prevalence (Falony*, et al.* 2016). However, in the same study, additional inclusion of the subjects from the LifeLines-DEEP study cohort (Dutch nationals), as well as from UK and US studies, resulted in a reduced core consisting of 17 genera that could be observed in both cohorts and there was no specific threshold on relative abundance for identifying core bacteria. In comparison, our study cohort included subjects from across mainland Europe, UK/Ireland and the USA (see Supporting Information). The lower estimates of the core size in the FGFP cohort could be partially explained by technical issues, including differences in DNA extraction methods, variations in sequencing depth or 16S rRNA gene primer sequences. However, at genus level, the reported core taxa from these two studies are largely consistent (Table S1, Supporting Information). Furthermore, reducing the stringency in the prevalence threshold in our data set to 80% and 50% with a constant detection threshold of 0.1% resulted in an increased size of the common core from 34 to 44 and 55 genera at 80% and 50% prevalence, respectively (Figure 1 and Table S1, Supporting Information). Thus, half the population studied here shared 42% of the 130 genus-like groups investigated. This further supports previous observations that the definition of the core microbiota is sensitive to the abundance and prevalence thresholds (Salonen*, et al.* 2012). In addition, there is a large amount of shotgun metagenomics data available in public databases which in some cases can identify species and led to the metagenome species concept defining not yet cultured and genomically sequenced taxa (Sunagawa*, et al.* 2013). This metagenomic approach is not affected by primer or PCR bias but to some extent is limited by the depth of sequencing and choice of DNA extraction method. Therefore, we further investigated the largest available metagenome catalog of 1267 human fecal samples to identify the core bacterial genera (Li*, et al.* 2014). At a detection threshold set at 0.1%, only Bacteroides spp. were

detected to be present in >90% of the samples (Table S1, Supporting Information). In order to overcome the influence of lower sequencing depth, we reduced the abundance threshold by 10-fold to 0.01% (still a conservative threshold) and identified eight genera namely *Bacteroides*, *Eubacterium*, *Faecalibacterium*, *Alistipes*, *Ruminococcus*, *Clostridium*, *Roseburia* and *Blautia* in more than half of the studied population (Table S1, Supporting Information). While a comparison of technologies is not the aim of the current review, we would like to highlight that the core microbiota includes a limited number of taxa that are commonly observed, irrespective of the technology used. As sequencing costs decline and new bioinformatics tools are being developed, shotgun metagenomics can be expected to become a standard in microbiome research and will lead to more uniformity between studies for comparisons. Hence, it will be imperative to revisit the core microbiota analysis in the future for more conclusive evidence. Studies incorporating strain-level comparison and identification of genomes without a reference genome will provide an even more detailed picture of the members, dynamics and functional aspects of the core microbiota (Qin*, et al.* 2010, Li*, et al.* 2014, Scholz*, et al.* 2016a). We hypothesize that different phylogenetic groups, which share important core functions in the intestinal ecosystem, co-exist and are part of the core. Analysis of the cumulative fractional abundance indicated that the common core bacteria are among the most dominant taxa in the Western adult population, including bacteria related to *F. prausnitzii*, *R. obeum*, *S. variabile*, *Prevotella melaninogenica* and *Oscillospira guillermondii* (Figure S2, Supporting Information). Cultured representatives for the vast majority of the common core bacteria are available and have sequenced genomes with information regarding their general metabolic traits and their association with human health. To this end, an extensive survey of literature was done to identify the general metabolic traits of the common core genus-like groups, and information for a subset discussed in more detail in this review is highlighted in Table 3. For information on other common core genus-like groups, see Table S2 (Supporting Information). We observed that there is a lack of in-depth understanding of the physiology of most of the core bacteria and their potential role in improving the stability of the intestinal microbiome.
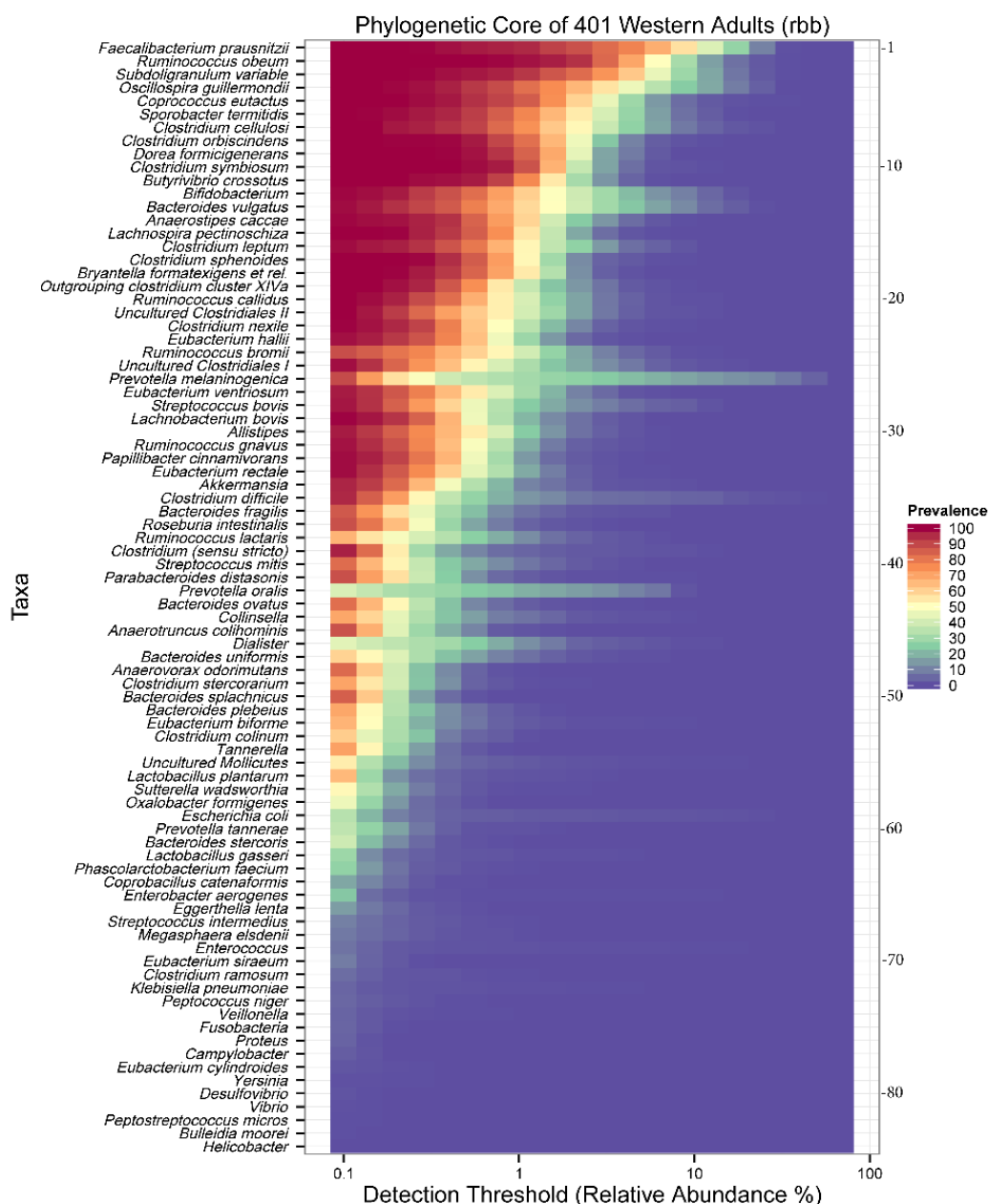
Figure 1. Bacterial phylogenetic core in the human large intestinal microbiome. We use data and analysis methodology from our previous studies (Jalanka-Tuovinen, *et al.* 2011, Salonen, *et al.* 2012). The data set was filtered based on DNA extraction method i.e. we included only the samples processed with the repeated bead beating method (rbb) which has been shown to outperform other methods (Salonen, *et al.* 2010).

Minimal Intestinal Metagenome

The human intestinal microbiome is a functional organ, with complex overall effects on host health. Therefore, it is crucial to identify functions commonly associated with the microbiota in health and disease. Whole shotgun metagenomics is widely used for identifying functional potential of the intestinal microbiome. Initially, the metagenomic

core was analyzed based on the DNA isolated from fecal samples of 18 individuals with

European ancestry (Turnbaugh, *et al.* 2009).

Table 3. Selected core genus-like taxa in human large intestine. Prevalence, general metabolic trait and health associations of few key taxa in the human intestine. Microbiota profiling was done as described previously (Lahti, *et al.* 2014). The analysis is based on 130 genus-like groups as defined by (Rajilić-Stojanović, *et al.* 2009). For a more detailed information on all the common core genus-like taxa can be found in the supplementary Table S2.

Sr. No.	Genus-like taxa	Prevalence	General metabolic trait	Health association(s)
1	*Faecalibacterium prausnitzii*	100%	Produce butyrate, formate and lactate (Duncan, *et al.* 2002b)	Decreased abundance in Crohn's disease and colon cancer (Kang, *et al.* 2010, Chen, *et al.* 2012)
2	*Bifidobacterium*	99.50%	Produce lactate and acetate by utilizing various oligosaccharides (Turroni, *et al.* 2008)	Widely used in probiotic preparations for health benefits (Turroni, *et al.* 2008)
3	*Akkermansia*	96.26%	Dominant mucin degrader (Belzer and de Vos 2012)	Indicator of a healthy metabolic profile in humans (Dao, *et al.* 2015)
4	*Prevotella melaninogenica*	89.53%	Acetate producer (Wu, *et al.* 1992)	One of the tipping elements (Lahti, *et al.* 2014)
5	*Ruminococcus bromii*	88.53%	Amylolytic activity (Ze, *et al.* 2012)	Keystone species and healthy effects via breakdown of resistant starch (Amir, *et al.* 2017)
6	*Bacteroides fragilis*	86.28%	Few strains have ability to degrade mucin (Tailford, *et al.* 2015)	Anti-inflammatory effects and also opportunistic pathogen (Sansonetti 2011)

Despite the limited sample size, this analysis suggested the presence of a minimal intestinal metagenome that was shared by almost all subjects in this study. In a first comprehensive metagenome study, the functional similarities of the fecal metagenome of 124 European individuals with 3.3 million unique open reading frames were reported and indicated that on average 38% of an individual's total gene pool is shared with others (Qin, *et al.* 2010). In a recent study, the coding capacity of the large intestinal microbiome was addressed in further detail by studying the metagenome of 1070 individuals from three continents, i.e. Europe, America and Asia (Li, *et al.* 2014). This study, which included both healthy and compromised subjects, identified around

9.9 million unique microbial genes within the intestinal microbiome, and the analysis indicated that this estimate is close to the saturated coverage of core gene content and functions. In case of the intestinal ecosystem, the minimal functional core can be expected to also contain more ecosystem-specific pathways, such as those involved in resistance to bile and ability to grow in an anoxic environment using fermentation or anaerobic respiration with terminal electron acceptors other than oxygen. In the first large-scale metagenomic study, more than 1000 clusters of genes that encode vital functions necessary for survival in the large intestine were identified (Qin, *et al.* 2010). This would suggest that the genomes of core bacterial taxa would have evolved toward high functional redundancy. To assess whether the common core observed in this study also demonstrated functional redundancy, we analyzed the representative genomes of the top 50 bacterial taxa that are part of the phylogenetic core (Figure 1). Since our analysis was performed at genus level, we included known intestinal genomes from different species targeted by the HITChip which resulted in selection of 80 bacterial genomes representative of 50 genus-like taxa (Table S3, Supporting Information) (Rajilić-Stojanović, *et al.* 2009). We first identified the KEGG orthologs (KO) abundances for each of the genomes using the Integrated Microbial Genomes (IMG) system and calculated the Spearman correlation coefficient (Markowitz, *et al.* 2012) (Figure 2, see Supplementary data for details). From the inferred network, we observed that bacteria belonging to the phylum Bacteroidetes have high functional redundancy, whereas the phylum Firmicutes was comprised of a large number of more functionally diverse core bacteria (Figure 3). Recently, analysis of the gyraseB encoding gene (gyrB) in fecal samples from humans and related primates (wild chimpanzees, bonobos and gorillas) indicated that Bacteroides and *Bifidobacterium* spp. have cospeciated with their mammalian hosts over thousands of generations suggesting strong symbiotic association of these taxa (Moeller, *et al.* 2016). This would indicate that the host may also play a role in selecting functionally redundant species, which results in high competition for resources in the ecosystem but at the same time leads to higher ecosystem stability, which ultimately benefits the host (Coyte, *et al.* 2015). However, the common core also includes taxa with specialized functional roles.
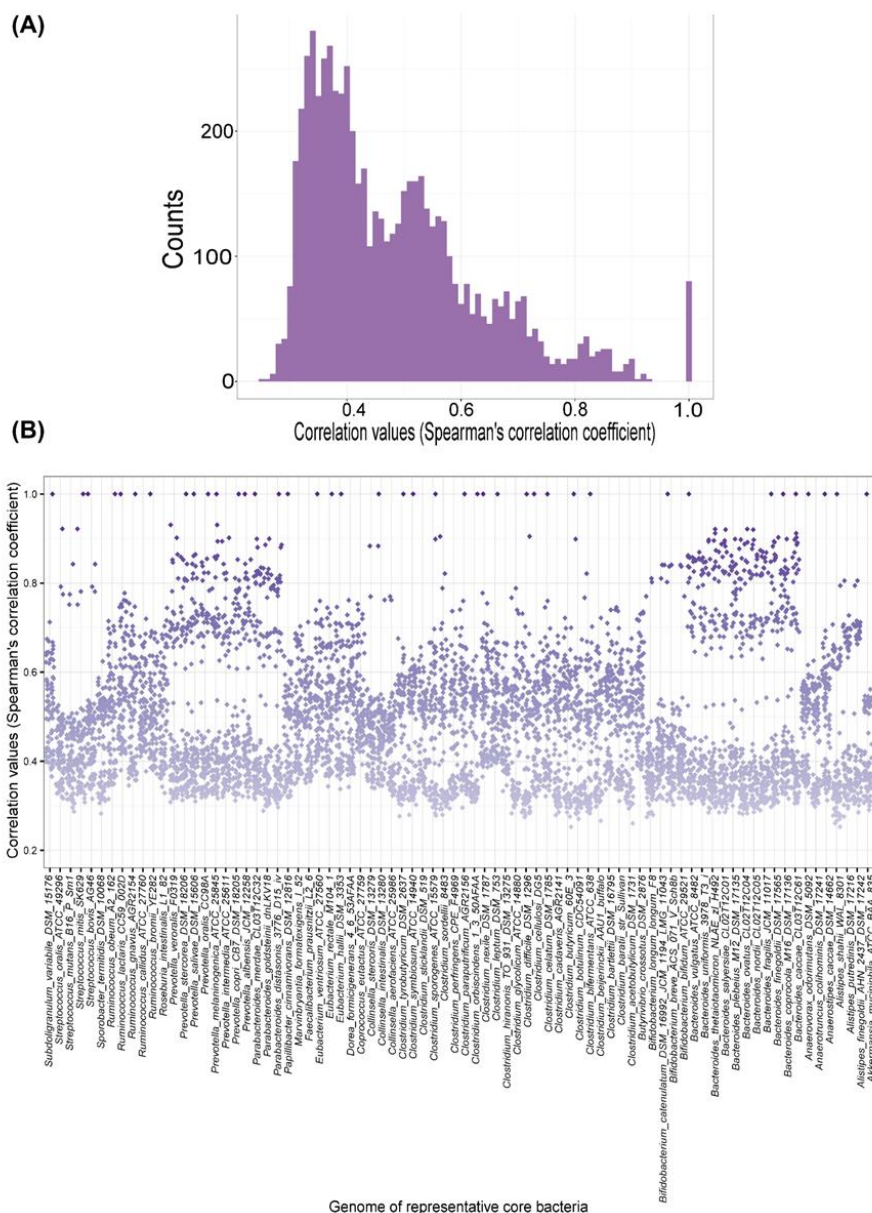
Figure 2. Spearman correlation coefficient for KO profiles for 80 genomes representative of the top 50 genus-level bacterial taxa including the phylogenetic core. (A) The distribution of Spearman correlation values in the core taxa. (B) The representative core bacterial genomes and their correlation to other genomes (each point depicts an independent genome correlated to genome on y-axis). The genomic data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community and were accessed using the IMG system.

A striking example is *Ruminococcus bromii* that was previously reported as a keystone species for its ability to degrade resistant starch and supporting growth of other bacteria capable of utilizing glucose, maltose, panose and isomaltose for growth (Ze, *et al.* 2012). A similar role of providing simple carbon sources and amino acids for supporting growth of microbes in the vicinity of the mucus layer can be played by

degraders of mucin, another important food source in the colon. The abundant representative of the Verrucomicrobia, *Akkermansia muciniphila*, another core member (Figure 1), is considered a keystone species in this process and is capable of using mucus as sole carbon and energy source, while signalling to the intestinal mucosa (Belzer and de Vos 2012, Everard*, et al.* 2013, Derrien*, et al.* 2017).
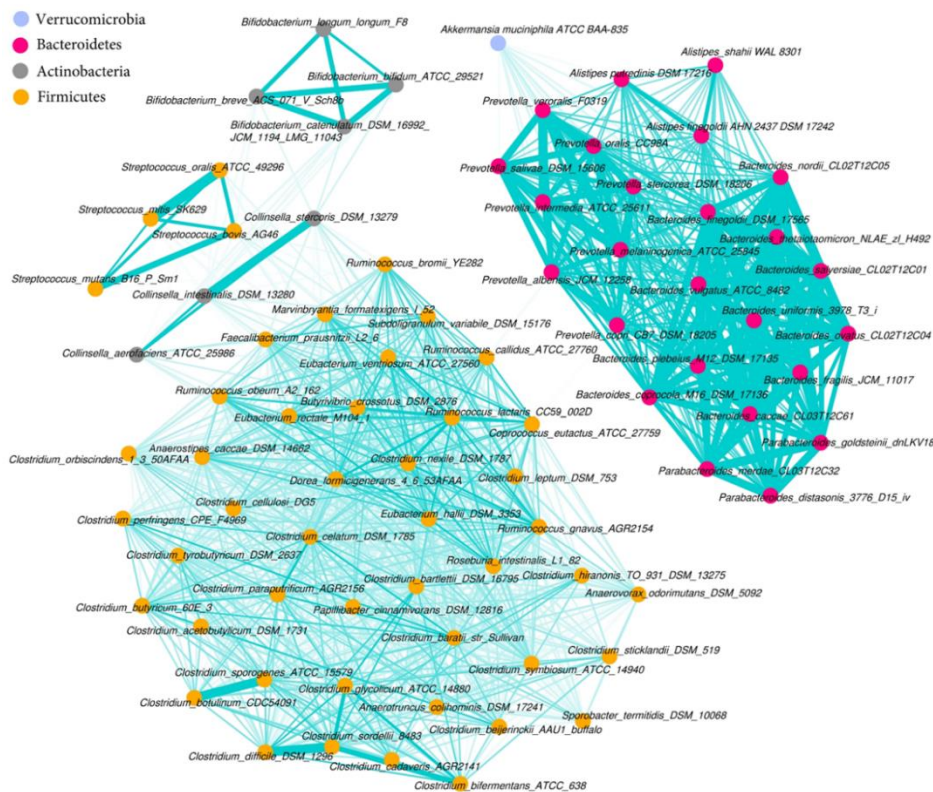

Figure 3. Functional correlation network of 80 bacterial genomes representative of the top 50 common core bacteria. The Spearman correlation matrix is represented as a network in which each genome is a node and each correlation an edge; the width of the edges is proportional to the magnitude of the correlation (the higher the correlation the thicker the edge line). The nodes with correlation coefficients below 0.5 are not connected, and genomes were placed by a graph 'spring' layout algorithm in qgraph (Epskamp*, et al.* 2012). The genomic data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community and were accessed using the IMG system.

Metagenomic sequencing data can provide insights into the function, taxonomic position and strain-level diversity of both known and not yet cultured species (Sunagawa*, et al.* 2013, Li*, et al.* 2014, Nielsen*, et al.* 2014, Jeraldo*, et al.* 2016). However, the active functions within the intestinal microbiome cannot be identified based on DNA sequencing data alone but require analysis of transcripts and proteins. To this end, several studies have addressed the active functional capacity of the

intestinal microbiota at the RNA and protein level, supporting the presence of an active functional core (Verberkmoes, *et al.* 2009, Erickson, *et al.* 2012, Kolmeder, *et al.* 2015, Kolmeder, *et al.* 2012). Metatranscriptomics analysis of fecal samples from 10 American adults revealed alterations in gene-expression profiles after changes in dietary intake (David, *et al.* 2014). Similarly, metaproteomic comparison of 29 non-obese and obese adults has shown larger differences in the active functional bacteria between the two groups than those observed at the compositional level (Kolmeder, *et al.* 2015). While these studies are a step toward functional understanding of the human microbiome, the field is still young. Improved statistical techniques and clinically relevant predictions will play a key role in a wider adoption to identify changes of functional features characteristic of different disease states.

Clusters, Continuous Gradients and Alternative Stable States

Several approaches have been employed to identify overarching community types, as well as alternative stable states in more specific taxonomic groups (Arumugam, *et al.* 2011, Holmes, *et al.* 2012, Ding and Schloss 2014). Moreover, longitudinal analysis of bimodal bacteria revealed temporal stability of the alternative states of low and high abundance, which are in some cases observable as ecosystem- level shifts (Lahti, *et al.* 2014). One potential explanation for the observed bistability could be that these states are resilient to perturbations and stable over ecologically relevant timescales (Connell and Sousa 1983). Additionally, the transition from one stable state to an alternative stable state as a result of a disturbance by e.g. dietary change, may comprise a period which can be stable for a long time relative to the extreme end state, called the alternative transient state (Fukami and Nakajima 2011). In alternative transient states, the communities are in a variable, non-resilient state and vary in structure and/or function as a result of processes such as priority effects (time and order of species arrival), external conditions and deterministic processes (Fukami and Nakajima 2011). Thus, the intestinal bacterial community can be viewed as a landscape representing varying global and/or local levels of stability, species richness and diversity (Arumugam, *et al.* 2011, Thaiss, *et al.* 2014, Koren, *et al.* 2013, Zoetendal and de Vos 2014). In this

context, the core bacteria and their abundances can be considered as potential biomarkers that may indicate changes in the intestinal ecosystem.

Alternative Community States

In an early approach to detect ecosystem-level structures in the intestinal microbiome, deep metagenome analysis suggested the presence of three clusters, termed enterotypes and characterized by co-occurring taxa named after their dominant bacterial groups i.e. *Bacteroides*, *Prevotella* and *Ruminococcus* (Arumugam*, et al.* 2011). One can argue whether it is possible to infer community assembly rules for the complex and highly dynamic intestinal ecosystem based on discrete alternative community types or whether the intestinal microbiome should rather be viewed as a continuous dynamic ecosystem, which is in constant flux but existing in a microbiota– host homeostatic state (Jeffery*, et al.* 2012, Knights*, et al.* 2014). Recently, it was reported that the enterotype-like clustering observed in ordination analysis could be influenced by the initial abundances of *Prevotella* and *Bacteroides* in the samples that were analyzed (Gorvitovskaia*, et al.* 2016). Moreover, it has been demonstrated that high levels of *Prevotella* and *Bacteroides* can co-exist in the host, suggesting that these types could be complementary rather than mutually exclusive (Lahti*, et al.* 2014). In addition, a separate study showed that removal of these two taxa from analysis did not result in differentiation of communities into clusters, suggesting that the dominance of *Prevotella* or *Bacteroides* did not represent underlying differences in community structure (Gorvitovskaia*, et al.* 2016). However, that study also found that the dominance of *Prevotella* or *Bacteroides* correlated well with dietary habits and lifestyle. Thus, it was suggested that the abundance of *Prevotella* and *Bacteroides* might be considered as a 'biomarker' of diet and lifestyle instead of using the term enterotypes (Gorvitovskaia*, et al.* 2016). A different approach to clustering and classification of samples based on microbial profiling data uses a probabilistic approach, called the Dirichlet multinomial mixture (DMM) models (Holmes*, et al.* 2012, Ding and Schloss 2014). This approach found evidence for the presence of alternative community configurations in the human intestinal microbiome regardless of the extensive intra- and inter- personal variations (Falony*, et al.* 2016, Holmes*, et al.* 2012).

Ecologists have also applied graph theory-based approaches to identify species interactions networks and modularity. In most studies, the co-occurrence and co-abundance network have been inferred from correlation-based and other similarity measures (Weiss*, et al.* 2016). One such attempt incorporated sparse regression with a number of similarity measures to predict microbial co-occurrence and co-exclusion relationships (Faust*, et al.* 2012). It was observed that the microbial community is organized into microcommunities which would indicate a trade-off between certain combinations of microbes that may represent alternative communities (Faust*, et al.* 2012). Recently, other tools have been developed to handle sparse compositional data (Friedman and Alm 2012, Gevers*, et al.* 2014, Kurtz*, et al.* 2015b). Compared to the standard correlation analyses, these new methods take better into account the sparsity, compositionality and high dimensionality of the taxonomic networks. One example is the use of CCREPE to identify Crohn's disease associated community network where it was observed that proinflammatory bacteria dominate and co-exclude other potentially beneficial taxa (Gevers*, et al.* 2014). This seems to be an example of a community shift. A similar observation of variable modularity in microbial interaction network was made in inflammatory bowel disease (IBD) patients and non-IBD controls (Baldassano and Bassett 2016). At the population level, the two states had a different modular structure and highlighted the different community memberships. However, it is vital to remember that the tools and models for identifying co-occurrences and modularity in microbiome data still need improvement (Weiss*, et al.* 2016). In all cases, these studies are observational and do not include interventions and hence inferences on causality cannot be made (de Vos and de Vos 2012). In general, more comprehensive longitudinal analyses and intervention studies will be needed to establish causal relationships.

Continuous Gradients and Bistability

The observed dynamics of the large intestinal ecosystem suggests the possibility for transitions between alternative and potentially stable states, with diet being one of the major driving factors (Falony*, et al.* 2016, Lahti*, et al.* 2014, Cotillard*, et al.* 2013, Wu*, et al.* 2011). For instance, it has been observed that a diet rich in proteins and animal fat is associated with the Bacteroides dominance, whereas carbohydrates have been

linked to high abundance of *Prevotella* (Wu*, et al.* 2011). The extremes of dietary habits point toward a *Prevotella*/*Bacteroides* trade-off driven by diet and suggesting a gradient in relative abundances between the two extreme states (high and low abundance) (Kelsen and Wu 2012). The joint population frequencies (represented as peaks and valleys) of the bistable taxa, *Prevotella* and *Bacteroides* exhibit a landscape distribution in the Western adult population (Figure 4A). This highlights the population-level variation in these key taxonomic groups. These two taxa exhibit high degree of functional redundancy as based on their genomic content when compared to functions shared between populations belonging to the Firmicutes or *Bifidobacterium* spp. (Figure 3). Previously, ecosystem-level models based on metagenomic and genomics data indicated that the habitat filtering (selection pressure inflicted by the habitat) may have greater impact than functional independence or cooperation on the community composition and thus, may result in selection of functionally redundant species (Levy and Borenstein 2014). It has also been proposed that the stability of intestinal microbiota is affected more by competing species than cooperating species (Coyte*, et al.* 2015). A new hypothesis emerged from the analysis of genome-based models that showed the module-derived functional redundancy to be reduced in the *Bacteroides* spp. as compared to other genera, potentially linking this to a decreased resilience to perturbations (Vieira-Silva*, et al.* 2016). Future studies aimed at mechanistic understanding of the interaction between these and other highly abundant or bistable taxa in the community and their potential role in ecosystemic transitions could provide the clues for microbiota modulation for instance based on dietary, antibiotic or other interventions.

To assess the dynamics of the large intestinal microbiota, it is necessary to study microbiota composition over different time scales and investigate the influence of perturbations. Compared to cross-sectional studies, analysis of time-series can provide information on intraindividual variations. Longitudinal analysis of larger populations gives insights into dynamics and stability of microbiota and could be informative in identifying key species that may play a role in ecosystem shifts (Faust*, et al.* 2015). By applying a discrete time Lotka-Volterra model to characterize bacterial interaction networks, it was observed that a high abundance does not necessarily imply a large

role in intestinal microbiota stability (Fisher and Mehta 2014). The identified key- stone species (*Bacteroides fragilis* and *B. stercosis*) in this study had moderate abundances but showed disproportionate influence on the interaction network. However, this study was limited by the number of individuals (two) included in the study. Variations of the Lotka-Volterra model have been used widely in ecology for the understanding of two species interactions. It is well known that the use of antibiotics may have a large impact on the microbial community and makes it prone to invasion by pathobionts (Bokulich*, et al.* 2016). A combination of Lotka-Volterra modeling and regression analysis in antibiotic-treated mice predicted that a microcommunity network was important in protection against *Clostridium difficile* invasion (Stein*, et al.* 2013). It was observed that a compromised community showed a catastrophic shift into an alternative state, which was stable and did not revert back to the initial community state after removal of the perturbations. Future studies applying these and other models on larger human cohorts may give better insights and aid in understanding ecology of the large intestinal microbiome.

In the intestinal tract microbiota, there exists a dynamic state and consists of species with varying fluctuations in abundances (Flores*, et al.* 2014, Caporaso*, et al.* 2011, Thaiss*, et al.* 2014). As there are thousands of species in the ecosystem, stability may arise in multiple states within a closed domain of possible alternative states also called 'basins of attraction' defined by a dynamic attractor (Scheffer*, et al.* 2001, Pepper and Rosenfeld 2012, Rosenfeld 2008). The microbial community in a given individual may move from one attractor to another over time and may include transiently occurring species from external sources (e.g. food) (Fukami and Nakajima 2011, Lang*, et al.* 2014). In many studies, single time point faecal samples are analysed and these may represent alternative transient states and consequently explain the high individuality observed in many studies (Flores et al. 2014). However, personalized patterns in the intestinal microbiota were detectable over 10 years of time, as shown by long-term monitoring of five healthy subjects (Rajilić-Stojanović*, et al.* 2013). Dense sampling of the human intestinal microbiota has revealed diurnal oscillations in relative abundance of almost 10% of all bacterial taxa and revealed a repetitive pattern of fluctuations in microbiota community configurations specific for a given time of the day (Thaiss*, et al.* 2014).

These fluctuations highlight the dynamic nature of the intestinal microbiota and indicate that there exists low and high abundance states for most bacteria. When put in perspective of the alternative transient states, these observations reveal an interesting feature of the intestinal microbiota. It is possible that the daily factors such as diet result in changes in the microbiota structure and composition (introduction of transient bacteria through food and water) and thus result in alternative transient states. This has been experimentally verified by small studies of extreme diets (plant-based diet or animal-based diet) and a diet swap study of African Americans and Native Africans (David*, et al.* 2014, O'Keefe*, et al.* 2015). There can be several possible paths of alternative transient states that may reach in one of multiple states within a limited set of attractors. However, if there is a major shift in dietary habit, this might push the community on an alternative transient path that ultimately results in a catastrophic shift in the community. Investigation of individual bacterial groups revealed bistability (i.e. alternative states of either high or low relative abundance) and identified six genus-level groups that were not only bimodal in their population frequencies but also exhibited bistability, *i.e.* being predominantly present at either high or low relative abundance with less frequent observations and a reduced temporal stability of the intermediate abundances (Lahti*, et al.* 2014). The intermediate abundance range (tipping point) exhibited reduced temporal stability within subjects (Lahti*, et al.* 2014). These bistable groups varied relatively independently and hence were termed `tipping elements`. At a community level, the corresponding compositional states could be described as specific combinations of such bistable tipping elements. Similar to *Prevotella melaninogenica* and *B. fragilis* distribution, we were interested in observing the population frequencies of highly abundant bistable bacteria *P. melaninogenica* and a lower abundance bistable taxon such as Dialister spp. in the cohort of Western adults. We used these examples to illustrate the alternative stable states mapping of the two-dimensional kernel density estimates for bacteria related to *P. melaninogenica* and *Dialister*. This analysis revealed three areas with high kernel density (Figure 4B). The less abundant Dialister group can exist in its high and low abundance states with *P. melaninogenica* low and high abundance states. Previously, the longitudinal assessment of bistability revealed that *Prevotella* groups exist in highly bistable states

and that their abundance levels are more stable in individuals over 3 months compared to other bistable taxa (Lahti*, et al.* 2014).
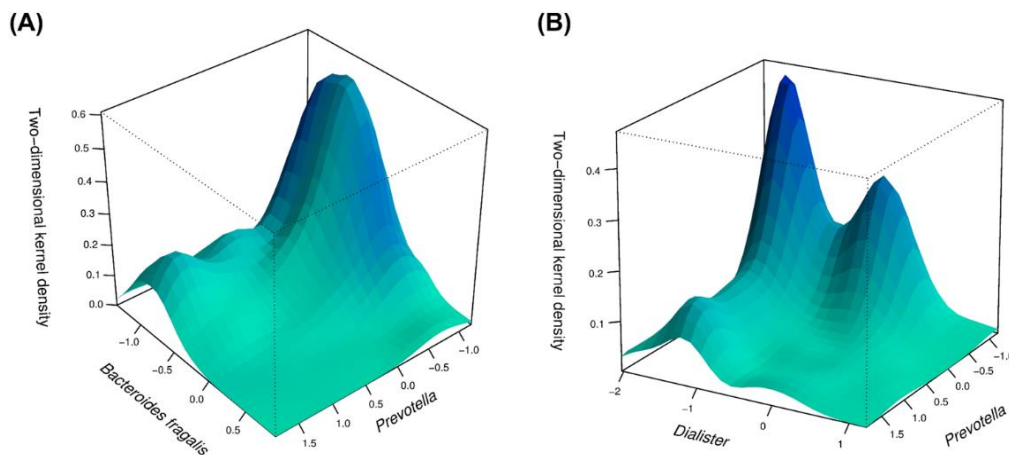

Figure 4: Two-dimensional kernel density (2D-kde) estimates. The abundance of bacteria (log10 transformed relative abundance data) mapped onto the x- axis and y-axis respectively, and the z-axis represents the kernel density estimate, in a 3D perspective plot showing the joint population frequencies of these two taxa demonstrating features similar to a landscape. (A) Mapping of *P. melaninogenica* and *B. fragilis* illustrates two distinct peaks that represent low *B. fragilis* and high *P. melaninogenica* population and low *B. fragilis* low *P. melaninogenica* population. (B) Mapping of *P. melaninogenica* and *Dialister* illustrates three distinct peaks that represent low *Dialister* and high *P. melaninogenica* population; low *Dialister* and low *P. melaninogenica* population; high *Dialister* and low *P. melaninogenica* population. The phylogenetic microarray data were obtained from previous study (Lahti*, et al.* 2014).

Furthermore, it was observed that the intestinal microbiota composition of individuals with bistable taxa closer to the tipping point is more variable in time (Lahti*, et al.* 2014). This indicates that the two extreme states are resilient, and changes in the ecosystem are more pronounced when bistable taxa have relative abundances closer to their tipping point.

Potential Implications of Community Characteristics for Microbiome Modulation Strategies

In the past few decades, research investigating the ecology of the intestinal microbiome has revealed various features, such as community assembly, temporal dynamics and stability and host covariates associated with intestinal microbiota (Table 4). The knowledge of positive, negative or neutral effects of bacteria on the ecosystem may help in identifying targets for therapeutic modulation. Additionally, features such as resilience of the ecosystem and effect of perturbations need to be understood in

detail. In the modification of an ecosystem, perturbation (random, episodic or programmed events) plays a major role (Paine*, et al.* 1998). The type of perturbation will affect the assembly of the community and also play a role in deter- mining the recovery state. For example, the low and high *Prevotella* states are highly stable and do not easily change over a short period of treatment (Lahti*, et al.* 2014). Furthermore, the initial state of a bacterial community has been reported to affect its recovery after antibiotic treatment (Raymond*, et al.* 2015). It has also been observed that a low-calorie dietary intervention could be helpful for obese individuals with lower bacterial richness (Tap*, et al.* 2015). Richness of the large intestinal microbiome is commonly associated with higher resilience (Lozupone*, et al.* 2012). Resilience and stability are the key features of the intestinal microbiome that need to be addressed while attempting modulation. Different alternative states may vary in both resilience (tendency of a system to return to a position of equilibrium when disturbed) and stability (tendency to maintain the state when subject to disturbance) (Walker*, et al.* 2004). Detailed understanding and identification of potential markers of these features will aid in designing intestinal microbiome modulation strategies.

In various diseased conditions, there is a shift in the composition and function of the intestinal microbiota driven by specific bacteria (possibly pathobionts). These bacteria are responsible for stabilizing an alternative community state, which is unfavourable for host health by producing specific metabolites that deter growth/survival of other bacteria. For example, in an inflamed intestine, increased levels have been observed of proinflammatory bacteria, such as members of the Enterobacteriaceae, associated with increased levels of the proinflammatory lipopolysaccharide produced by this group of Gram-negative bacteria (Wang*, et al.* 2007a). The resulting inflammatory cascade includes an increased concentration of host-produced reactive oxygen species that hamper growth of indigenous anaerobic bacteria (Lupp*, et al.* 2007). Thus, this aberration may represent an alternative stable state, the stability of which is maintained by an inflammatory milieu created by the host and influenced by members of the Enterobacteriaceae.

Table 4: Representative studies investigating the ecology of the intestinal microbiome employing high-throughput sequencing approach (most recent in each category have been listed).

Sr. No.	Names of the study	Conceptual insights
1	Temporal variability is a personalized feature of the human microbiome (Flores*, et al.* 2014)	Elucidates the temporal dynamics and stability of the microbiome including the human
2	Moving pictures of the human microbiome (Caporaso*, et al.* 2011)	First large-scale longitudinal sampling temporal dynamics and stability of the human microbiome
3	Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis (Thaiss*, et al.* 2014)	Address the diurnal changes in microbiota and identifies hourly fluctuations in some bacterial genera
5	Cospeciation of gut microbiota with hominids (Moeller*, et al.* 2016)	Provides support to the theory of cospeciation of human intestinal tract microbiota along with hominids
5	Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota (Stein*, et al.* 2013)	Temporal dynamics and stability and influence of perturbation on community dynamics and stability
6	The long-term stability of the human gut microbiota (Faith*, et al.* 2013)	Addresses the question of long-term stability of microbial community using novel high-throughput sequencing method
7	Antibiotics, birth mode and diet shape microbiome maturation during early life (Bokulich*, et al.* 2016)	Elucidated the community assembly in early life and effect of perturbation on shaping the microbiome
8	Population-level analysis of gut microbiome variation (Falony*, et al.* 2016)	Population level insights into intestinal microbial community structure and composition with analysis of large number of co-founding factors
9	An integrated catalog of reference genes in the human gut microbiome (Li*, et al.* 2014)	Population level insights into intestinal microbial community function
10	Enterotypes of the human gut microbiome (Arumugam*, et al.* 2011)	Differences in community types/clusters in human intestinal microbiome based on partitioning around the medoid method
11	Dynamics and associations of microbial community types across the human body (Ding and Schloss 2014)	Differences in community types/clusters in human intestinal microbiome on similar lines of Enterotypes paper but using DMM models
12	Linking long-term dietary patterns with gut microbial enterotypes (Wu*, et al.* 2011)	Stability of *Prevotella* enterotype from the *Bacteroides* enterotype was tested
13	Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle (Gorvitovskaia*, et al.* 2016)	Investigating data bias in detecting enterotypes and address the issues pertaining to it
14	Microbial co-occurrence relationships in the human microbiome (Faust*, et al.* 2012)	Community interactions networks, niche specialization and assembly of microbial community have been demonstrated
15	The treatment-naive microbiome in new-onset Crohn's disease (Gevers*, et al.* 2014)	Disease and healthy interaction network with respect to Crohn's disease are elucidated

16	Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and IBD (Greenblum, *et al.* 2012)	Metabolic network modularity of microbiome in health and disease state
17	Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression (Fisher and Mehta 2014)	A novel method for identifying the keystone species based on longitudinal studies is reported
18	Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa (Trosvik and Muinck 2015)	Using network statistics and reverse ecology the authors identify foundation and keystone taxa in human intestinal microbiome

Hence, the transition from higher densities of obligate anaerobes to proinflammatory facultative anaerobes would be critical for the establishment and/or maintenance of a state of increased inflammation. In such cases, a reduction in numbers of proinflammatory bacteria will be crucial for a shift toward an alternative state (may or may not be stable) no longer supportive of inflammation, and thus leading to re-establishment of anoxic conditions in the intestine. Along with the reduction in proinflammatory facultative anaerobes, it will be crucial to support the growth of bacteria capable of tolerating oxygen at low concentrations, such as lactobacilli and *Akkermansia muciniphila* that have anti-inflammatory effects for re- establishment of an anoxic milieu (Tien*, et al.* 2006, Ouwerkerk*, et al.* 2016).

The relative abundance of diagnostic indicator taxa, such as bacteria belonging to Enterobacteriaceae (increasing relative abundance) and *A. muciniphila* (decreasing relative abundance), was shown to indicate particular environmental conditions such as an inflamed intestinal milieu (Png*, et al.* 2010). In addition, indicator taxa can be sensitive to perturbations and therefore might serve as an early warning sign of unhealthy shifts in the community. Such indicator taxa can be monitored as their abundance may reflect the efficacy of intervention strategies. Using the knowledge of drivers and indicators, it would be possible to detect diagnostic biomarkers, identify specific therapeutic microbes (single keystone species or mixed consortia) for treatment and suggest lifestyle changes (such as dietary habits) that will result in a 'healthier' intestinal bacterial community.

Evidence for the predictive nature of intestinal bacteria was recently presented when investigating obese subjects undergoing dietary interventions (Korpela*, et al.* 2014). The responsiveness of the intestinal bacterial community to a specific dietary intervention was predicted by the baseline relative abundances of *Clostridium* clusters IV, IX and XIVa, and Bacilli. The members of several of the *Clostridium* clusters also include a number of the core phylotypes in the human intestine, most notably *Clostridium* cluster IV (*C. leptum* group, a major constituent of which is the Ruminococcaceae family) and *Clostridium* cluster XIVa (*C. coccoides* group, which resembles the Lachnospiraceae family) (Tap*, et al.* 2009, Rajilić-Stojanović and de Vos 2014). Based on these findings, it is tempting to speculate that some core phylotypes may also serve as indicator species for responsiveness, and their relative abundance may have profound influence on the efficacy of a given intervention strategy. Another example is *A. muciniphila*, which is a part of the intestinal core, and has been reported to be indicative of a healthy metabolic profile. Further- more, its abundance along with high gene richness was indicative of better clinical outcomes after calorie restriction diet in overweight/obese adults (Dao*, et al.* 2015). Evidence supporting the importance of bacterial community structure in determining individuals' responses to dietary modulation (in ecological terms; ecological disturbance) was provided by the findings of Salonen *et al.* (2014). It was observed that the dietary intervention had little effect on the intestinal bacterial composition of non-responders who were characterized by higher initial bacterial diversity. Thus, a higher baseline bacterial diversity represents a more resilient community, and increased responsiveness would be characteristic of an unstable community. This high- lights that the baseline bacterial diversity could be one of the potential markers in predicting responsiveness to dietary modulation. However, it is important to acknowledge the effect of other, more specific, factors such as relative abundance of certain bacteria, e.g. bistable bacteria, in determining an individuals' response to intervention. It will be interesting to investigate whether the bistable groups can serve as predictors for responsiveness or non-responsiveness to modulation strategies.

The functional core and the identification of bistability of specific metabolic processes could also provide clues for predicting efficacy of interventions. For instance, also the

overall bacterial gene richness has been demonstrated to exhibit bi- modal population frequencies (Le Chatelier*, et al.* 2013). Western individuals have been observed to be separated into two groups based on gene count, i.e. low gene count (LGC) and high gene count (HGC), with low gene richness being associated with pronounced dys-metabolism and low-grade inflammation. Thus, LGC and HGC could represent two alternative states in which one state (LGC) represents a state that is indicative of increased risk for obesity (Cotillard*, et al.* 2013). In this study, it was observed that the LGC group after dietary intervention had improved gene richness concomitant with improvement in the clinical phenotypes but showed little change in the inflammation status of these individuals. These functional alternative states, i.e. LGC and HGC, provide us with an opportunity to predict the efficacy of any approach used for modulation. This further highlights the importance of richness of the intestinal microbiome (both compositional and functional) as a factor in determining resilience to ecosystem changes. It is known that species richness determines the susceptibility to invasion by non-resident species and resilience to overall ecosystem disturbances (Levine and D'Antonio 1999). In case of the intestinal microbiome, resilience of a community with high functional richness adds an important dimension to our understanding of its complexity. Thus, there is a need for considering both species richness and functional gene richness during intervention studies aimed at microbiome modulation.

Future Prospects: Applying Community Level Understanding for Beneficial Landscaping of The Intestinal Microbiome

In the past decade, numerous studies targeting the human intestinal microbiome have led toward remarkable advances in our understanding of the crucial role these microbes play in human health. Recognizing the importance of this microbial community, the intestinal microbiome is often called a 'metabolic organ' signifying its crucial functional role which has an effect on our health status (Bocci 1992). Importantly, this metabolic organ represents a functioning ecosystem, which is governed by ecological mechanisms. At present, clinically meaningful evidence for the potential application of modulating the intestinal microbiota for therapeutic gain has

created considerable interest and enthusiasm (Li*, et al.* 2016, Smits*, et al.* 2013). Treatment of metabolic syndrome and *Clostridium difficile* infection by ecosystem-level cleansing restoration via fecal microbiota transplantation (FMT) has demonstrated the potential for translating microbiota research to clinical practice (Li*, et al.* 2016). While FMT is promising with proven efficacy in several cases, the underlying mechanisms that result in beneficial effects on host health are unclear and will need further refinement. One of the approaches is targeting specific group(s) of bacteria to bring about the desired changes in the intestinal microbiome. The state of a dynamic system may be manipulated by triggering a switch to an alternative state, for instance with a strong but temporally limited, transient perturbation ('pulse', antibiotics), or longer impact ('press', diet) (Costello*, et al.* 2012, Faust*, et al.* 2015). These transient perturbations can also be combined with introduction of single bacteria or groups of bacteria that support or inhibit growth of specific bacteria resulting in desired changes in composition and function of the intestinal microbiome.

Managing Invasion or Overgrowth Of Pathobionts

In order to manage invasions or overgrowth of pathobionts, we first need to identify phylotypes that are common residents and their relative proportions in the intestinal ecosystem. The study of common core phylotypes and core functions is important in this aspect as this would form the baseline knowledge of 'normal' inhabitants. Comparison of this core healthy state with diseased states will assist in identifying invasive species (potentially harmful species that are not part of the core) and/or pathobionts (resident microbes with pathogenic potential). For instance, during IBD Enterobacteriacaeae are known potential pathobiont species that affect the numbers of anaerobic intestinal bacteria as mentioned above. In such cases, the use of therapeutic microbes combined with specific dietary interventions can be used to support growth of core phylotypes capable of fermenting complex polysaccharides to produce short- chain fatty acids (SCFA) and gas (CO_2 and H_2), which results in decreased intraluminal pH to establish an anoxic environment (Tien*, et al.* 2006, Damaskos and Kolios 2008). Other examples of pathobionts in the human gastrointestinal intestinal tract are *Helicobacter pylori*, *Fusobacterium* spp., *Desulfovibrio* and *C. difficile* (Biagi*, et al.* 2010, Rowan*, et al.* 2010, Castellarin*, et al.*

2012, Kamada, *et al.* 2013, Martin and Solnick 2014). In most cases, the common observation in disease cases is the overgrowth of these pathobionts. In mouse models of antibiotic-mediated *C. difficile* infection, it was observed that a few bacteria confer protection and imbalance in their populations increases susceptibility to overgrowth by pathobionts (Stein, *et al.* 2013). These few bacteria included *Akkermansia*, *Coprobacillus* and *Blautia* spp. that are all part of the core and their inhibition by antibiotics was pro- posed to support overgrowth of *C. difficile* assisted by *Enterococcus* spp. in mice (Stein, *et al.* 2013). Thus, it could be possible that the knowledge of common core phylotypes combined with the low and high abundance states of pathobionts could help in designing specific approaches for managing invasion by pathogens or increased relative abundance of pathobionts. These approaches could include changes in the diet that support higher growth of core phylotypes and functions such as butyrate production which inhibits pathogen colonization (Scott, *et al.* 2014).

Adapting Intervention Strategies for Modulating the Microbiome

The outcome of a given intervention strategy will depend on the individual's response to certain modulation attempts. The response of the individual to a specific intervention can be predicted and monitored by observing overall community descriptors such as species and gene richness and diversity, as well as abundance/presence/absence of key indicator taxa that may reflect the efficacy of intervention strategies. Indicator taxa could include bistable bacteria, and their shift toward the tipping point could provide clues for transition from diseased to healthy alternate stable states. Studying the transition between low and high abundance states or vice versa with an intermediate tipping point could provide clues for the critical time point for ecosystem restoration by modulating the microbiome toward a different, possibly healthier state. It is also possible that the shift from a diseased state may not necessarily lead to a healthy state but rather to a state different from either of the states. These alternative scenarios may be dependent on the strength of perturbation that resulted in shift to a diseased state. In any case, monitoring of microbiome composition and its dynamics over time should be incorporated in intervention strategies.

Maintenance of A Healthy Intestinal Microbiome Landscape Throughout Life Span

Human health is interlinked with the human body and its associated microbiome, and thus, maintaining a healthy microbiome is essential. The characteristics of a healthy community com- position have been debated, and could be potentially identified by observing the common core and bacterial communities associated with alternate stable states. The core phylotypes have co-evolved to provide crucial ecosystem services such as production of butyrate (energy source for colonic epithelial cells) and protection against invasive pathogens and outgrowth of endogenous pathobionts. The process of modulating microbiome can be initiated in early life by managing certain host associated and environmental factors like diet. The mode of delivery has an effect on the pattern of development of the intestinal bacterial community (Grölund*, et al.* 1999, Cox*, et al.* 2014). Caesarean delivery is associated with a reduced number of *Clostridium* species most of which are part of the core microbiota in adult life (Figure 1) (Bokulich*, et al.* 2016). However, another recent report did not observe associations between the delivery mode and adult microbiota composition (Falony*, et al.* 2016). Maternal health status (obesity) and maternal stress have also been associated with an aberrant intestinal bacterial community in infants (Galley*, et al.* 2014, Zijlmans*, et al.* 2015). However, modifying the dietary habits, breast feeding and incorporating specific oligosaccharides have shown to aid in improving the abundance of anaerobic bacteria (Haarman and Knol 2006). Exercising regularly is known to have various health benefits. A recent study on professional athletes revealed that exercise could aid in increasing intestinal bacterial diversity (Clarke*, et al.* 2014). In elderly subjects, the intestinal microbiome is characterized by reduced bacterial diversity (Biagi*, et al.* 2010, Claesson*, et al.* 2012). Reduced abundance of beneficial core bacteria, increased inflammation and a decrease in the concentration of SCFA's have also been observed in elderly subjects (Salazar*, et al.* 2013). These deviations from a normal adult healthy state could be reduced by targeted modulation, which may include restoring the healthy core phylotypes (through probiotic and/or prebiotic interventions), reducing proinflammatory bacteria and increasing the SCFA production by incorporating dietary fibers in the daily diet.

Concluding Remarks

In the past few decades, we have gained significant insights into the composition, dynamics and function of our intestinal microbiome and its impact on our well-being. The Western adult population shares dozens of genus-level taxonomic groups that show high functional correlation. The intestinal microbiome should be viewed as a complex landscape, which is a mixture of common and individual characteristics. In addition to continuously changing gradients across bacterial abundance and function, this intestinal microbiome landscape also includes distinct features present as alternative stable states across varied compositional, structural and functional features. It is now clear that treatment for most of the disease conditions necessities careful consideration of the intestinal microbiome. A recent example is the application of a prediction algorithm based on dietary habits, physical activity and gut microbiota that allowed predicting blood glucose levels and designing personalized diets which were effective in controlling post-prandial blood glucose levels (Zeevi, *et al.* 2015). While the field of microbiome research rapidly moves forward, there has to be a careful consideration of the confounding variables in the studies, knowing that correlation does not imply causation, technical bias between studies, differences in statistical models, human error and most importantly reproducibility (Ioannidis 2005, Falony, *et al.* 2016). As the micro- biome research moves into the next phase of diagnostics and therapeutics, overcoming these issues will be a major focus of future research.

Acknowledgements

Supplementary data

Supplementary table S1: Genus-like taxa identified as member of core microbiota in Western Adults based on different microbial diversity profiling methods.

Sr. No	Genus-like groups	Rel. abundance (%)	Prevalence (%)
	HITChip based Core (This study)		
1	*Bryantella formatexigens (Marvinbryantia formatexigens)*	0.1	100
2	*Butyrivibrio crossotus*	0.1	100
3	*Clostridium cellulosi*	0.1	100
4	*Clostridium orbiscindens*	0.1	100
5	*Clostridium sphenoides*	0.1	100
6	*Clostridium symbiosum*	0.1	100
7	*Coprococcus eutactus*	0.1	100
8	*Dorea formicigenerans*	0.1	100
9	*Faecalibacterium prausnitzii*	0.1	100
10	*Lachnospira pectinoschiza*	0.1	100
11	*Oscillospira guillermondii*	0.1	100
12	*Outgrouping clostridium* cluster XIVa	0.1	100
13	*Ruminococcus callidus*	0.1	100
14	*Ruminoccocus obeum (Blautia obeum)*	0.1	100
15	*Sporobacter termitidis*	0.1	100
16	*Subdoligranulum variable* et rel.	0.1	100
17	Uncultured *Clostridiales* II	0.1	100
18	*Clostridium nexile*	0.1	99.75
19	*Lachnobacterium bovis*	0.1	99.75
20	*Anaerostipes caccae*	0.1	99.50
21	*Bifidobacterium*	0.1	99.50
22	*Bacteroides vulgatus*	0.1	99.25
23	*Clostridium leptum*	0.1	99.25
24	*Eubacterium rectale*	0.1	99.25
25	*Papillibacter cinnamivorans*	0.1	99.25
26	Uncultured *Clostridiales* I	0.1	99.00
27	*Eubacterium hallii*	0.1	98.75
28	*Ruminococcus gnavus*	0.1	98.50
29	*Streptococcus bovis*	0.1	98.25
30	*Allistipes*	0.1	97.50
31	*Eubacterium ventriosum*	0.1	97.50
32	*Clostridium (sensu stricto)*	0.1	97.00
33	*Akkermansia*	0.1	96.25
34	*Clostridium difficile*	0.1	95.76
35	*Prevotella melaninogenica*	0.1	89.53
36	*Ruminococcus bromii*	0.1	88.53
37	*Parabacteroides distasonis*	0.1	88.28
38	*Roseburia intestinalis*	0.1	88.28

39	*Anaerotruncus colihominis*	0.1	87.78
40	*Bacteroides fragilis*	0.1	86.28
41	*Bacteroides splachnicus*	0.1	85.04
42	*Anaerovorax odorimutans*	0.1	82.54
43	*Streptococcus mitis*	0.1	82.29
44	*Bacteroides ovatus*	0.1	81.55
45	*Tannerella*	0.1	70.57
46	*Clostridium stercorarium*	0.1	70.07
47	*Bacteroides plebeius*	0.1	69.33
47	*Collinsella*	0.1	69.33
47	*Eubacterium biforme*	0.1	66.33
50	*Ruminococcus lactaris*	0.1	66.33
51	*Lactobacillus plantarum*	0.1	65.59
52	*Clostridium colinum*	0.1	61.10
53	*Bacteroides uniformis*	0.1	59.60
54	Uncultured *Mollicutes*	0.1	53.86
55	*Sutterella wadsworthia*	0.1	51.37
16S rRNA gene sequencing based Core (Falony et al 2016)			
1	Unclassified *Veillonellaceae*	-	96.62
2	Unclassified *Ruminococcaceae*	-	99.86
3	Unclassified *Lachnospiraceae*	-	99.89
4	Unclassified *Hyphomicrobiaceae*	-	96.40
5	Unclassified *Erysipelotrichaceae*	-	99.20
6	Unclassified *Clostridiales*	-	97.80
7	Unclassified *Clostridiaceae*	-	98.41
8	*Roseburia*	-	99.12
9	*Faecalibacterium*	-	98.87
10	*Dorea*	-	98.02
11	*Coprococcus*	-	96.37
12	*Clostridium*_XlVa	-	97.39
13	*Blautia*	-	99.45
14	*Bacteroides*	-	99.64
15	*Parabacteroides*	-	95.30
16	*Clostridium*_IV	-	95.85
17	*Alistipes*	-	97.86
Metagenome based Core (Data re-analysed from Li et al 2014)			
1	*Bacteroides*	0.017061	99.75
2	*Eubacterium*	0.017061	86.67
3	*Faecalibacterium*	0.017061	83.02
4	*Ruminococcus*	0.017061	76.35
5	*Alistipes*	0.017061	73.46
6	*Clostridium*	0.017061	70.19
7	*Roseburia*	0.017061	67.04
8	*Blautia*	0.017061	53.21

Supplementary table S2: Top 50 genus-like taxa that form the common core in human large intestine. Genus-like taxa with sequenced genome representatives are in bold type face. Microbiota profiling was done as described previously (Lahti, *et al.* 2014). The analysis is based on 130 genus-like groups as defined by (Rajilić-Stojanović, *et al.* 2009). The taxa with representative genome sequences in databases are in bold.

Sr. No.	Genus-like taxa	General metabolic trait	Health/disease association(s)
1	***Faecalibacterium prausnitzii***	Produce butyrate, formate and lactate (Duncan, *et al.* 2002b)	Decreased abundance in Crohn's disease and colon cancer (Chen, *et al.* 2012, Kang, *et al.* 2010)
2	***Ruminococcus obeum*** [*]	Produces acetate from cellobiose, arabinose and other carbohydrate sources (Lawson and Finegold 2015)	Related to *Blautia*, these are depleted in Elderly and colorectal cancer patients (Chen, *et al.* 2012, Claesson, *et al.* 2012)
3	***Subdoligranulum variable***	Butyrate producers (Holmstrøm, *et al.* 2004)	Depleted in patients with *Clostridium difficle* infections (Pérez-Cobas, *et al.* 2014)
4	*Oscillospira guillermondii*	Unknown	Abundant in subjects with low Body mass index (Tims, *et al.* 2013)
5	***Coprococcus eutactus***	Butyrate producers together with formate or lactate (Holdeman and Moore 1975)	Butyrate is associated with health (Canani, *et al.* 2011)
6	***Sporobacter termitidis***	Homoacetogenic, grows on methylated aromatic compounds for example 3,4,5-trimethoxycinnamate (Grech-Mora, *et al.* 1996)	Abundant in subjects with low Body mass index (Tims, *et al.* 2013)
7	***Clostridium cellulosi***	Associated with degradation of plant material (Yanling, *et al.* 1991)	Abundant in subjects with low Body mass index (Tims, *et al.* 2013)
8	***Clostridium orbiscindens*** [¶]	Ability to utilize flavonoids (Carlier, *et al.* 2010)	High abundance in non-obese subjects (Kasai, *et al.* 2015)
9	***Dorea formicigenerans***	Produce hydrogen and carbon dioxide as fermentation end products (Taras, *et al.* 2002)	Increased abundance in adult irritable bowel syndrome patients (Rajilić–Stojanović, *et al.* 2011)
10	***Clostridium symbiosum***	Acetate, butyrate, and lactate producer (Vos, *et al.* 2011)	Single individual case report of infection in colon cancer (Elsayed and Zhang 2004)
11	***Butyrivibrio crossotus***	Ability to produce formate, lactate and butyrate (Moore, *et al.* 1976)	Butyrate producer observed lower abundance in individuals with Low gene richness (Le Chatelier, *et al.* 2013)
12	***Bifidobacterium***	Produce lactate and acetate by utilizing various oligosaccharides (Turroni, *et al.* 2008)	Widely used in probiotic preparations for health benefits (Turroni, *et al.* 2008)
13	***Bacteroides vulgatus***	Can utilize few polysaccharides such as	High abundance in biopsies from colon cancer subjects (Fujita, *et al.*

			amylose and amylopectin (Salyers, *et al.* 1977)	2002) on the contrary, *B. vulgatus* is abundant in healthy controls compared to patients with Crohn's disease (Kang, *et al.* 2010)
14	*Anaerostipes caccae*		Saccharolytic, Acetate/lactate-utilising, butyrate-producing bacterium (Schwiertz, *et al.* 2002)	Butyrate is associated with health (Canani, *et al.* 2011)
15	*Lachnospira pectinoschiza*		Pectin degrading ability (Cornick, *et al.* 1994)	*Lachnospira pectinoschiza* correlate with butyrate,acetate and formate in fecal samples (O'Keefe, *et al.* 2015)
16	*Clostridium leptum*		Produces acetate (Moore, *et al.* 1976)	*C. leptum* numbers and diversity significantly reduced in both Crohn's disease and ulcerative colitis (Kabeerdoss, *et al.* 2013)
17	*Clostridium sphenoides*		Citrate-fermenting and acetate producer (Walther, *et al.* 1977)	Abundance of *C. sphenoides* is lower in the individuals with increased cholesterol levels (Korpela, *et al.* 2014)
18	*Bryantella formatexigens* [β]		Utilizes glucose to produce acetate (Wolin, *et al.* 2003)	NA
19	Outgrouping *Clostridium* cluster XIVa		*Clostridium aminovalericum* is only known cultured representative) known to ferment aminovalerate to produce acetate, propionate, butyrate, and valerate (Buckel, *et al.* 1994)	Abundant in non-secretors compared to secretors (Wacklin, *et al.* 2014)
20	*Ruminococcus callidus*		Cellulolytic activity (Chassard, *et al.* 2010)	More abundant in healthy controls than in patients with Crohn's disease (Kang, *et al.* 2010)
21	Uncultured *Clostridiales* II		Know bacteria include *Christensenella minuta* (Morotomi, *et al.* 2012)	UCII as group is one of the tipping elements and potential signature of compromised gut in Primary sclerosing cholangitis (Lahti, *et al.* 2014, Rossen, *et al.* 2014). *C.minuta* is reported to be a heritable taxon with capacity to reduced weight gain (Goodrich, *et al.* 2014)
22	*Clostridium nexile*		Ability to produce a trypsin-dependent antimicrobial substance active against *Clostridium perfringens* (Marcille, *et al.* 2002)	Unknown

23	*Eubacterium hallii*	Lactate utilizing and butyrate producing bacteria (Duncan, *et al.* 2004)	Known for preventing and/or treating insulin resistance (Nieuwdorp and De Vos 2012)
24	*Ruminococcus bromii*	Amylolytic activity (Ze, *et al.* 2012)	Keystone species and healthy effects via breakdown of resistant starch (Ze, *et al.* 2012)
25	*Uncultured Clostridiales I*	Unknown	One of the tipping elements and decreased in the obese individuals (Lahti, *et al.* 2014, Verdam, *et al.* 2013)
26	*Prevotella melaninogenica*	Produces acetate (Wu, *et al.* 1992)	One of the tipping elements (Lahti, *et al.* 2014)
27	*Eubacterium ventriosum*	Butyrate producer (Van den Abbeele, *et al.* 2013)	Positively correlated to BMI differences (Tims, *et al.* 2013)
28	*Streptococcus bovis*	Acetate, ethanol and formate at pH above 6.5 and lactic acid at lower pH (Finlayson 1986)	*S. bovis* overgrowth associated with colorectal cancer (Abdulamir, *et al.* 2011)
29	*Lachnobacterium bovis*	Glucose fermentation yields majorly lactic acid with minor amounts of acetic and butyric acids. Reported ability temperature-sensitive bacteriocin-like inhibitory substance (Whitford, *et al.* 2001)	*Lachnobacillus* as mentioned in HITChip original paper, the name is not validly published and is a spelling error. This group of bacteria are *Lachnobacterium bovis* and related organisms which bolong to the clostridial XIVa cluster of the low-G+C content Gram-positive bacteria (Whitford, *et al.* 2001)
30	*Allistipes*	Predicted ability to utilize lysine and produce butyrate (Vital, *et al.* 2014a)	Reduced abundance in patients with non-alcoholic fatty liver disease (Jiang, *et al.* 2015)
31	*Ruminococcus gnavus*	Mucin utilizing ability (Crost, *et al.* 2013)	Lost in Ulcerative Colitis patients (Nishikawa, *et al.* 2009)
32	*Papillibacter cinnamivorans*	Ability to utilize aromatic compounds like cinnamate (Defnoun, *et al.* 2000)	Correlated with C-reactive protein (Verdam, *et al.* 2013)
33	*Eubacterium rectale*	Polysaccharide degrading butyrate producer (Ze, *et al.* 2012)	Positively correlated to BMI differences (Tims, *et al.* 2013)
34	*Akkermansia*	Dominant mucin degrader (Belzer and de Vos 2012)	Indicator of a healthy metabolic profile in humans (Dao, *et al.* 2015)
35	*Clostridium difficile*	Includes potential pathobionts and acetogens	Species such as *C. bifermentas* are part of the core (this study)
36	*Bacteroides fragilis*	Few strains have ability to degrade mucin (Tailford, *et al.* 2015)	Anti-inflammatory effects and also opportunistic pathogen (Sansonetti 2011)
37	*Roseburia intestinalis*	Saccharolytic, butyrate-producer (Duncan, *et al.* 2002a)	Positively correlated to BMI differences (Tims, *et al.* 2013)

38	*Ruminococcus lactaris*	Acetogen	NA
39	*Clostridium* (sensu stricto)	NA	Includes opportunistic pathogens
40	*Streptococcus mitis*	NA	Includes opportunistic pathogens
41	*Parabacteroides distasonis*	Acetate and succinic acid producer (Sakamoto and Benno 2006).	Reduced in inflamed intestine (Zitomersky*, et al.* 2013)
42	*Prevotella oralis*	Lactose and cellobiose utilizer (Hayashi*, et al.* 2007)	Reduced in inflammatory bowel syndrome (Rajilić–Stojanović*, et al.* 2011)
43	*Bacteroides ovatus*	Lactulose utilizer (Sahota*, et al.* 1982)	Opportunistic pathogen
44	*Collinsella*	Fermentation products of glucose are H_2, ethanol, formate and lactate (Kageyama and Benno 2000)	Reduced abundance in IBS (Kassinen*, et al.* 2007)
45	*Anaerotruncus colihominis*	Acetate and butyrate producer (Lawson and Finegold 2015)	10 fold higher during bloating compared to healthy controls (Jalanka-Tuovinen*, et al.* 2011)
46	*Dialister*	Metabolic end products are detected in trace amounts only (Downes*, et al.* 2003)	Reduction in plasma IL-6 concentration significantly correlated with higher abundance of *Dialister* (Martínez*, et al.* 2013)
47	*Bacteroides uniformis*	Has α-Fucosidase activity, ability to utilize cellobiose and xylose (Bakir*, et al.* 2006)	Decreased in Obese subjects (Verdam*, et al.* 2013)
48	*Anaerovorax odorimutans*	Ability to ferment putrescine to acetate, butyrate, H_2 and ammonia (Matthies*, et al.* 2000)	No conclusive evidence till date on association/correlation
49	*Clostridium stercorarium*	Cellulolytic bacteria (MADDEN 1983)	Overabundance in diabetic children (de Goffau*, et al.* 2014)
50	*Bacteroides splachnicus*[φ]	Known acetate, propionate, and butyrate and potential lysine utilizers producing butyrate (Vital*, et al.* 2014a)	Reduced in ulcerative colitis subjects (Morgan*, et al.* 2012)

Note: [*]*Ruminococcus obeum* is now reclassified and transferred to the genus *Blautia* as *Blautia obeum* (Lawson and Finegold 2015); [¶]*Clostridium orbiscindens* is now known as *Flavonifractor plautii* (Carlier*, et al.* 2010); [β]*Bryantella formatexigens* is renamed as *Marvinbryantia formatexigens* (Wolin*, et al.* 2003). [φ]*Bacteroides splachnicus* is renamed as *Odoribacter splanchnicus* (Hardham*, et al.* 2008). Legend: NA: No association reported in literature till date.

Supplementary table S3: List of publicly available genome used for calculating the KEGG KO functional correlation. These data were obtained from the Integrated Microbial Genomes (IMG) system and deposited by various research groups including some of the data produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community.

Genome Name	Status	Study Name	Genome Size (bp)	Gene Count
Clostridium sordellii 8483	Permanent Draft	*Clostridium sordellii* 8483	7808455	8736
Butyrivibrio crossotus DSM 2876	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2482791	2546
Bifidobacterium catenulatum DSM 16992, JCM 1194, LMG 11043	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2058429	2009
Clostridium hiranonis TO-931 DSM 13275	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2423348	2302
Prevotella copri CB7, DSM 18205	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3507873	3414
Ruminococcus obeum A2-162	Finished	Human Microbiome Project (HMP) Reference Genomes	3757491	3210
Prevotella veroralis F0319	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2986735	3110
Bacteroides uniformis 3978 T3 i	Permanent Draft	Comparative genomics of human mucosal isolates of the enterotoxigenic *Bacteroides fragilis* group, etiologic agents of inflammatory diarrheal disease	5047633	4497
Anaerovorax odorimutans DSM 5092	Permanent Draft	Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project	3265634	3015
Bacteroides finegoldii DSM 17565	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4881901	4557
Prevotella salivae DSM 15606	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3140543	2992
Bacteroides vulgatus ATCC 8482	Finished	*Bacteroides vulgatus* ATCC 8482	5163189	4195
Bacteroides nordii CL02T12C05	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	5707589	4506
Clostridium sticklandii DSM 519	Finished	*Clostridium sticklandii* DSM 519	2715461	2688
Bacteroides caccae CL03T12C61	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	5479119	4674

Eubacterium ventriosum ATCC 27560	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2869695	2892
Coprococcus eutactus ATCC 27759	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3102087	3065
Collinsella aerofaciens ATCC 25986	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2439869	2440
Bacteroides salyersiae CL02T12C01	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	5781840	4677
Parabacteroides merdae CL03T12C32	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4918045	4248
Clostridium paraputrificum AGR2156	Permanent Draft	The Hungate 1000. A catalogue of reference genomes from the rumen microbiome	3561289	3568
Anaerostipes caccae DSM 14662	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3605636	3845
Subdoligranulum variabile DSM 15176	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3237471	3450
Dorea formicigenerans 4_6_53AFAA	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3825700	3902
Alistipes putredinis DSM 17216	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2549878	2799
Clostridium bifermentans ATCC 638	Permanent Draft	*Clostridium bifermentans* ATCC 638	3609653	3652
Prevotella oralis CC98A	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2981962	2550
Prevotella melaninogenica ATCC 25845	Finished	Human Microbiome Project (HMP) Reference Genomes	3168282	2361
Clostridium leptum DSM 753	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3270109	3983
Clostridium tyrobutyricum DSM 2637	Permanent Draft	Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project	3028196	3061
Ruminococcus bromii YE282	Draft	The Hungate 1000. A catalogue of reference genomes from the rumen microbiome	2539482	2571
Clostridium orbiscindens 1_3_50AFAA	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4323157	4375

Bifidobacterium bifidum ATCC 29521	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2198551	2165
Clostridium glycolicum ATCC 14880	Permanent Draft	Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project	4003570	3918
Faecalibacterium prausnitzii L2-6	Finished	Human Microbiome Project (HMP) Reference Genomes	3321367	2820
Akkermansia muciniphila ATCC BAA-835	Finished	*Akkermansia muciniphila*	2664102	2238
Streptococcus mutans B16 PSm1	Permanent Draft	*Streptococcus mutans* str. B16 P Sm1	2329852	2290
Streptococcus bovis AG46	Permanent Draft	The Hungate 1000. A catalogue of reference genomes from the rumen microbiome.	1930741	1924
Clostridium nexile DSM 1787	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3861016	4297
Bacteroides plebeius M12, DSM 17135	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4421324	4022
Collinsella intestinalis DSM 13280	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	1804297	1846
Clostridium butyricum 60E.3	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4618205	4232
Marvinbryantia formatexigens I-52	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4548960	4960
Streptococcus oralis ATCC 49296	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2054852	2052
Ruminococcus callidus ATCC 27760	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3070687	2913
Prevotella intermedia ATCC 25611	Permanent Draft	Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project	2575089	2248
Eubacterium hallii DSM 3353	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3290996	3297
Eubacterium rectale M104/1	Finished	Human Microbiome Project (HMP) Reference Genomes	3698419	3267
Prevotella stercorea DSM 18206	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3092669	3066
Clostridium beijerinckii AAU1_buffalo	Permanent Draft	Clostridium beijerinckii AAU1_buffalo	6954347	10315
Bacteroides thetaiotaomicron NLAE-zl-H492	Draft	Genomic variation in Bacteroides isolates	6792460	5906

Clostridium celatum DSM 1785	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3553721	3484
Clostridium bartlettii DSM 16795	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	2971856	2875
Clostridium perfringens CPE F4969	Permanent Draft	Clostridium perfringens CPE F4969	3510272	3116
Alistipes shahii WAL 8301	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3763317	2616
Streptococcus mitis SK629	Permanent Draft	*Streptococcus mitis* genome project, IGS	2213700	2209
Prevotella albensis JCM 12258	Permanent Draft	*Prevotella* Genome sequencing - Univ of Tokyo	2674720	2602
Parabacteroides distasonis 3776 D15 iv	Permanent Draft	Comparative genomics of human mucosal isolates of the enterotoxigenic *Bacteroides fragilis* group, etiologic agents of inflammatory diarrheal disease	5422204	4803
Ruminococcus gnavus AGR2154	Permanent Draft	The Hungate 1000. A catalogue of reference genomes from the rumen microbiome	3724077	3784
Parabacteroides goldsteinii dnLKV18	Permanent Draft	*Parabacteroides goldsteinii* dnLKV18	7066206	5752
Bacteroides ovatus CL02T12C04	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	7880760	6465
Clostridium acetobutylicum DSM 1731	Finished	*Clostridium acetobutylicum* DSM 1731	4145581	4029
Papillibacter cinnamivorans DSM 12816	Draft	Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project	2890610	2751
Bacteroides coprocola M16, DSM 17136	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4295617	4394
Bacterioides fragilis JCM 11017	Permanent Draft	*Bacterioides* species genome sequencing - The University of Tokyo	5270831	5342
Roseburia intestinalis L1-82	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4380675	4776
Clostridium sporogenes ATCC 15579	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4102125	3869
Clostridium difficile DSM 1296	Permanent Draft	Genomic Encyclopedia of Type Strains, Phase I: the one	4215232	3935

|				
		thousand microbial genomes (KMG-I) project		
Sporobacter termitidis DSM 10068	Draft	Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project	4198214	3867
Bifidobacterium breve ACS-071-V-Sch8b	Finished	Human Microbiome Project (HMP) Reference Genomes	2327492	1890
Clostridium symbiosum ATCC 14940	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	4810675	4440
Clostridium botulinum CDC54091	Permanent Draft	Diversity of *Clostridium botulinum* type F strains CDC	4079045	5896
Alistipes finegoldii AHN 2437, DSM 17242	Finished	A Genomic Encyclopedia of Bacteria and Archaea (GEBA)	3734239	3302
Ruminococcus lactaris CC59_002D	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3051582	2832
Clostridium cadaveris AGR2141	Permanent Draft	The Hungate 1000. A catalogue of reference genomes from the rumen microbiome	3542160	3459
Bifidobacterium longum longum F8	Finished	Human Microbiome Project (HMP) Reference Genomes	2384987	1744
Anaerotruncus colihominis DSM 17241	Permanent Draft	Human Microbiome Project (HMP) Reference Genomes	3718888	4506
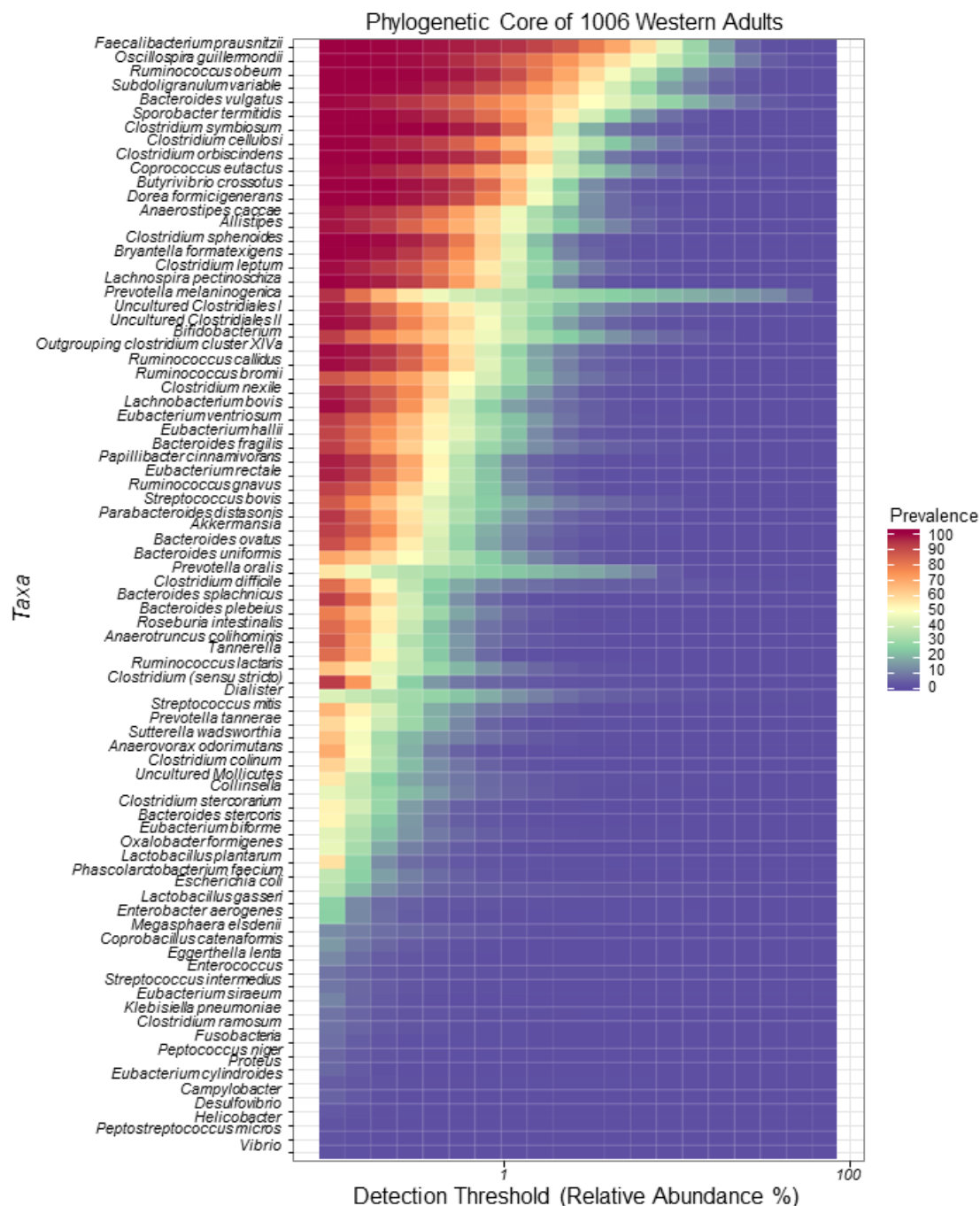
Figure S1: Bacterial core observed in the large intestine of 1006 Western adults. The data here includes core microbiota analysis of all the samples irrespective of the method used to extract faecal DNA.
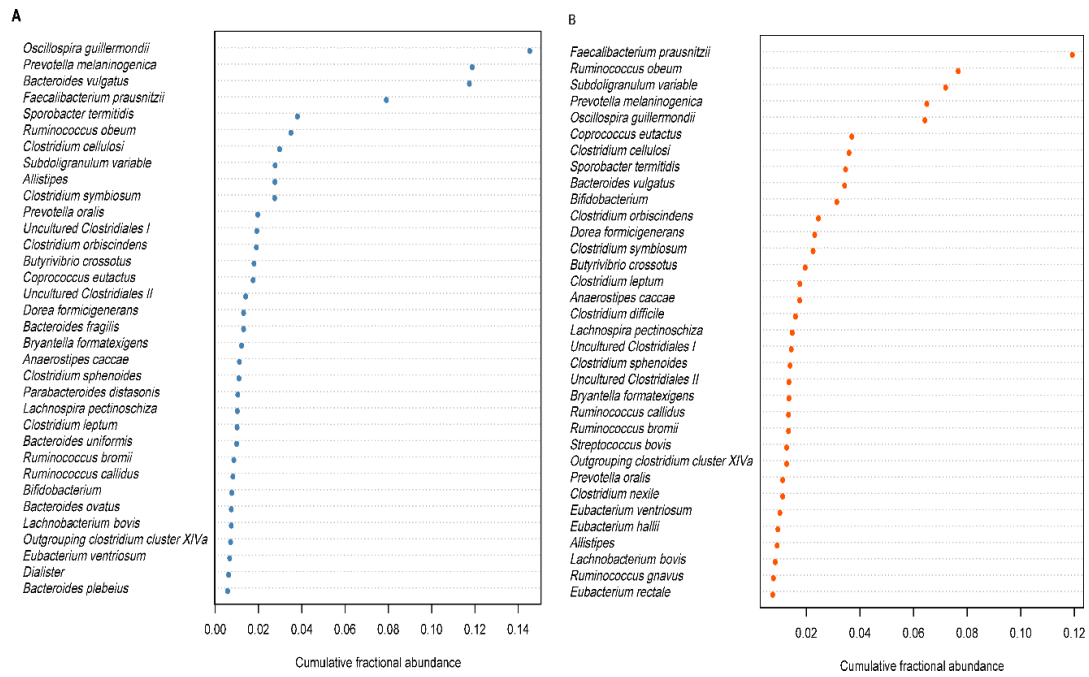
Figure S2: Ranked abundances based on cumulative fractional abundance of top 50 genus-like taxa using different DNA extraction methods. A) Ranked abundances based on cumulative fractional abundance for non-mechanical DNA extraction methods. B) Ranked abundances based on cumulative fractional abundance for mechanical (repeated bead beating) DNA extraction methods.

Chapter 3

Hyperdominance in microbial communities

Sudarshan A. Shetty, Leo Lahti, Egbert H. van Nes, Willem M. de Vos,

Marten Scheffer and Hauke Smidt
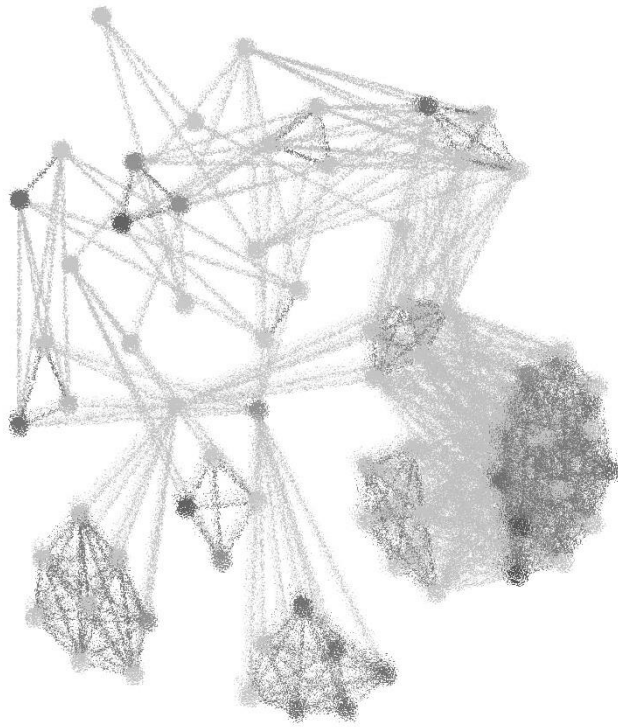
Manuscript in preparation

Abstract

More than an estimated 10^{12} microbial phylotypes inhabit planet earth, shaping a broad range of properties of both free-living and host-associated ecosystems. The major impact is through the collection of diverse microbes that form ecosystem-specific communities. Knowledge of the fundamental characteristics of microbial communities is an important step towards understanding the functioning and long-term stability of any ecosystem. Recently, an investigation of soil microbial communities across the globe revealed that around 2% of the observed microbial phylotypes accounted for 50% of the total community. This so called hyperdominance has been commonly observed also in macro-ecosystems. However, knowledge of the prevalence of such hyperdominance for other microbial communities is lacking. Here, using the Earth Microbiome Project (EMP) 16S ribosomal RNA (rRNA) gene-based communal catalogue covering a broad range of diverse ecosystems, we show that hyperdominance is a characteristic feature of both free-living and host-associated microbial ecosystems. Globally, on average less than 4.6 % of 16S rRNA gene amplicon sequence variants (ASVs) account for a cumulative relative abundance of 50% in diverse ecosystems. The human microbiome is no exception, and hyperdominance is highly prevalent in gut, oral and skin microbiomes. This was validated by analysis of public metagenomic datasets. Using temporal microbial profiling data, we furthermore show that several of the hyperdominant ASVs have a high and low abundance state in an individual's life. These observations provide an intriguing view of microbial communities, and especially the human gut microbiome, where hyperdominance is driven by immigration and perturbations, metabolic versatility of bacteria, habitat-filtering and host-associated selective forces.

Keywords: Hyperdominance, Conditionally rare taxa, Human gut microbiota

Introduction

Micro-organisms drive various biogeochemical processes and have major impact on fitness of other, multicellular, species. One of the primary goals of microbial ecologists is to map and understand microbial diversity, structure and function, and variability thereof, across a broad range of free-living and host-associated ecosystems (Thompson*, et al.* 2017). The distribution of microbial taxa in diverse environments is heavily skewed towards the rare fraction, whereas only few taxa seem to dominate (Lynch and Neufeld 2015). However, an important question is how many taxa have a substantial dominance? Are there specific traits that can be associated with dominance of a specific taxon or is dominance driven by random processes? Are the taxa in the rare biosphere less important or do they play a role in stabilizing the large metacommunity through aspects such as functional redundancy following an insurance hypothesis (McCann 2000)? The first reported documentation of hyperdominance (defined as the fraction of the species that account for 50% of the total biomass) was made for the Amazonian tree flora, where 1.4% of the tree species was found to account for half of the total biomass (ter Steege*, et al.* 2013). Although causes for the observed hyperdominance remain unknown, an interesting observation from the study of tree species was that most of the hyperdominant species were habitat specialists (ter Steege*, et al.* 2013). Considering that recent conceptual and technological advances now for the first time allow comprehensively studying microbial ecosystems, observations and causes of hyperdominance and potential impact of hyperdominant species on the structure, stability, resilience and functioning of microbial ecosystems became an intriguing aspect also in microbial ecology.

Here we first identified the fraction of phylotypes present in a given community that represents 50% of the cells or specific marker sequences such as the 16S rRNA gene, in that community. Since such nucleic acid based molecular profiling data cannot be directly correlated with the contribution of species to the total biomass, we define hyperdominance as given in equation below:

$$\mathrm{HD}_{\mathrm{microbial}} = log_2(1/coverage)$$

Here *HD$_{microbial}$* is the hyperdominance value, and *coverage* is the percentage of phylotypes accounting for a cumulative relative abundance of 50%, with these phylotypes being defined as hyperdominant phylotypes. We used the recently published Earth Microbiome Project (EMP) microbial survey data to quantify hyperdominance in a broad range of microbial ecosystems (Thompson*, et al.* 2017). This communal catalogue provides a unique resource to investigate global biodiversity and to test ecological concepts in the microbial world (Figure S1). We then more specifically explored the human microbiome to further investigate the nature and dynamics of hyperdominance and hyperdominant taxa in a host-associated community. The human microbiome is a multifaceted complex ecological "model" with intricate relationships between the microbial community and its host (Cho and Blaser 2012). Especially in the intestinal tract, the microbes compete for resources, including both host-derived carbon and energy sources such as mucus, and dietary components such as complex fibres that mostly end up undigested in the large intestine. Non-continuous food intake patterns lead to sudden influx of substrates, which can in turn result in transient shifts in community activity and even composition (Lang*, et al.* 2014). On the other hand, the host immune system plays a major role in keeping a check on its microbial partners (Foster*, et al.* 2017a). The host-associated microbiome is driven by both bottom-up and top-down selection pressures (Costello*, et al.* 2012, Foster*, et al.* 2017a, Koskella*, et al.* 2017). The host is under pressure to select for microbial communities that will aid in improving host fitness and survival. Thus, we were interested in the hyperdominant bacteria in the human microbiome (skin, oral and more specifically the intestine) to identify bacteria that have been successful in coping with the evolutionary pressure to dominate in the ecosystem. In addition, the microbes associated with the human body have received much attention resulting in large amounts of multi-omics derived data with respect to their ecophysiology (Klaassens*, et al.* 2007, Erickson*, et al.* 2012, Bui*, et al.* 2015, Jeraldo*, et al.* 2016). This is important for finding potential links between ecological observations and functioning of the ecosystem.

Methods

Datasets analysed in this study

Earth Microbiome Project

We used a subset of the 16S rRNA gene based survey made available as part of the EMP catalogue (https://zenodo.org/record/890000) (Thompson*, et al.* 2017). The biom file used was "emp_deblur_150bp.subset_2k.rare_5000.biom", 150bp and rarefied to 5000 reads per sample. This included 975 samples consisting of 91364 ASVs. We removed ASVs that were observed only once in the complete dataset to avoid inflating our estimation of hyperdominance. The phylogenetic tree was pruned accordingly to remove these singleton ASVs. The second subset of the EMP dataset used for the human-associated microbial survey was from "emp_deblur_100bp.subset_2k.rare_5000.biom", which was filtered to keep only human-associated data. This included 160 samples with a total of 4330 ASVs. After removing ASVs classified as mitochondria and chloroplasts a total of 3754 ASVs were retained for final analysis.

Moving pictures of the human microbiome

The data for human microbiome time series was obtained from "emp_deblur_150bp.release1.biom" by filtering to keep only samples from the Qiita study ID 550 (Caporaso*, et al.* 2011). This data was not rarefied to equal sampling depth, however, we removed samples with less than 1000 reads. We also removed ASVs observed only once in the entire dataset as well as ASVs classified as mitochondria and chloroplasts. Finally, we used 2510 ASVs in a total of 943 samples.

Metagenomic datasets of the human microbiome

The metagenomic data was obtained from the curatedMetagenomicData R/Bioconductor package (Pasolli*, et al.* 2017) (Pasolli *et al.*, 2017). We used data from three studies, including those of M. Schirmer and co-authors (Schirmer*, et al.* 2016); C. Huttenhower and colleagues (Human Microbiome Project phase 1) (Huttenhower*, et al.* 2012); and H. Nielsen and co-workers (Nielsen*, et al.* 2014). Only samples marked as healthy were used for our study. We used 866 samples having a total of 1077 bacterial metagenomic species.

Identification of hyperdominance

We used only those samples which had a richness of more than 30 ASVs. This was done to capture hyperdominance in complex and diverse microbial communities. We identified the number of ASVs that accounted for 50% of all sequences in a given sample by converting sequence counts to relative abundance and identifying N (number of taxa) that occupy 50% of the community. The function used for this purpose, *coverage*, is available in the microbiome R package (Lahti and Shetty 2018).

Identification of the temporal core microbiota and conditionally rare taxa

Core ASVs were identified separately for gut, oral and skin microbiota. These core ASVs were those that were detected at a relative abundance of at least 0.0002 (abundance of the least abundant ASV) in at least 95% of the samples of a given time series. The conditionally rare taxa were identified using the previously described methods, and the codes were used from the GitHub repository (https://github.com/ShadeLab/ConditionallyRareTaxa) (Shade*, et al.* 2014).

Simulations of the neutral community dynamics

The simulations of time-series dynamics for neutral communities were done using the seqtime R package, function *simHubbell* (Faust*, et al.* 2018). The perturbations were defined using the function *perturbation*, where, times=seq(1,10000), durations=rep(1,10000), numberchanges=c(10000,-10,rep(0,8)), capacityConstant = TRUE, deathrate = 10.

Data availability

All the data and codes used for this study are available from the GitHub repository https://github.com/microsud/ecologicalmicrobiomes.

Results and Discussion

Half the abundance in microbial communities is contributed by only a few

We used the EMP Ontology (EMPO) for a hierarchical view of diverse microbial ecosystems (http://press.igsb.anl.gov/earthmicrobiome/protocols-and-standards/empo/) (Thompson*, et al.* 2017). Briefly, the EMPO framework includes four levels, including empo_0: EMP sample; empo_1: free-living or host-associated; empo_2: saline, non-saline, animal, plant, or fungus; and empo_3: specific habitat e.g.

animal distal gut. We first assessed the frequency of hyperdominance in free-living and host-associated environments collected from across the globe as a part of the EMP catalogue (Figure 1) (Thompson, *et al.* 2017). For this, we calculated the percentage of 16S rRNA gene amplicon sequence variants (ASVs) that made up 50% of all sequences in a given sample. We removed samples with inadequate richness (less than 30 ASVs) as they may not be representative of the real-world complexity of microbial communities. We observed a left-skewed distribution of coverage (*i.e.* percentage of ASVs accounting for 50% of the total number of sequences) in free-living and host-associated microbial communities (Figure 1A).
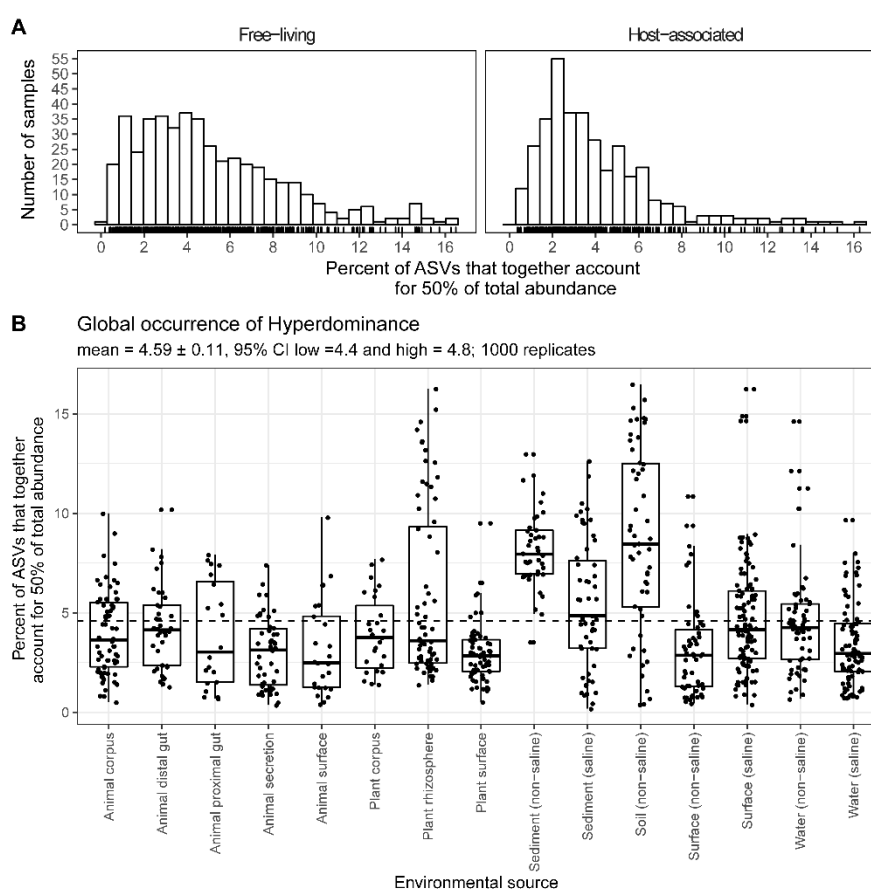
Figure 1: Fraction of ASVs that account for a cumulative relative abundance of 50% in microbial communities globally. **A]** Frequency distributions of percentages of ASVs accounting for 50% of the total abundance in free-living and host-associated communities. **B]** Comparison of the fraction of species accounting for 50% of the total abundance in samples from diverse environments classified according to EMP ontology level 3. The horizontal dotted line indicates the mean percentage of the observed ASVs that account for 50% of the total number of sequences.

Among the 813 samples used for this analysis, on average 4.6% ± 0.11 (mean ± SD, at 95% Confidence interval (CI), 1000 bootstrap replicates) of the observed ASVs accounted for 50% of the total number of sequences, with 509 (60.5%) of all samples with percentages below the mean (*i.e.* 4.6%) (Figure 1B). Hyperdominance in host-associated environments was prominent, where the majority of samples included in our analysis (69%) had a coverage below 4.6 %. The mean coverage for host-associated microbial communities was 4% ±0.1 (at 95% CI, 1000 bootstrap replicates). In free living communities, 53.9% of the samples had a coverage lower than the global mean (4.6%) of the ASVs contributing to a cumulative relative abundance of 50%. The mean coverage specifically for free-living communities was 5.06 ± 0.1 (at 95% CI, bootstrap 1000 replicates). A large variation in coverage in free living communities was mostly observed in soil and sediment samples, which had higher richness compared to host associated ecosystems (Figure 1B and Figure S2). These results strongly support the observation that hyperdominance is an intrinsic feature of microbial communities in natural ecosystems, irrespective of the type of environment.

Hyperdominance and its relation with richness and inequality

Next, we evaluated to what extent inequality and hyperdominance relate to richness in microbial communities. We used the Gini index as a measure of inequality, a commonly used measure for relative inequality in economical income (Gini 1936). We observed a negative correlation between richness and hyperdominance in both free-living and host-associated microbial communities (Figure 2A and 2C). This was in line with the observation of a negative correlation between richness (number of observed ASVs) and Gini index in both free-living and host-associated ecosystems (Figure 2B and 2D). In both types of ecosystems, it was observed that inequality is prominent in microbial communities with low richness. Similar observations were made when using Faith's phylogenetic diversity, which is a commonly used phylogenetically weighted richness measure (Figure S3). Further investigation at more specific habitat level (EMPO, level 3) indicated that hyperdominance in microbial communities associated with plants correlated negatively with richness, especially the plant rhizosphere communities ($R > 0.8$, $P < 0.05$) (Figure S4). In contrast, animal associated microbial communities only showed marginal association of richness with hyperdominance.
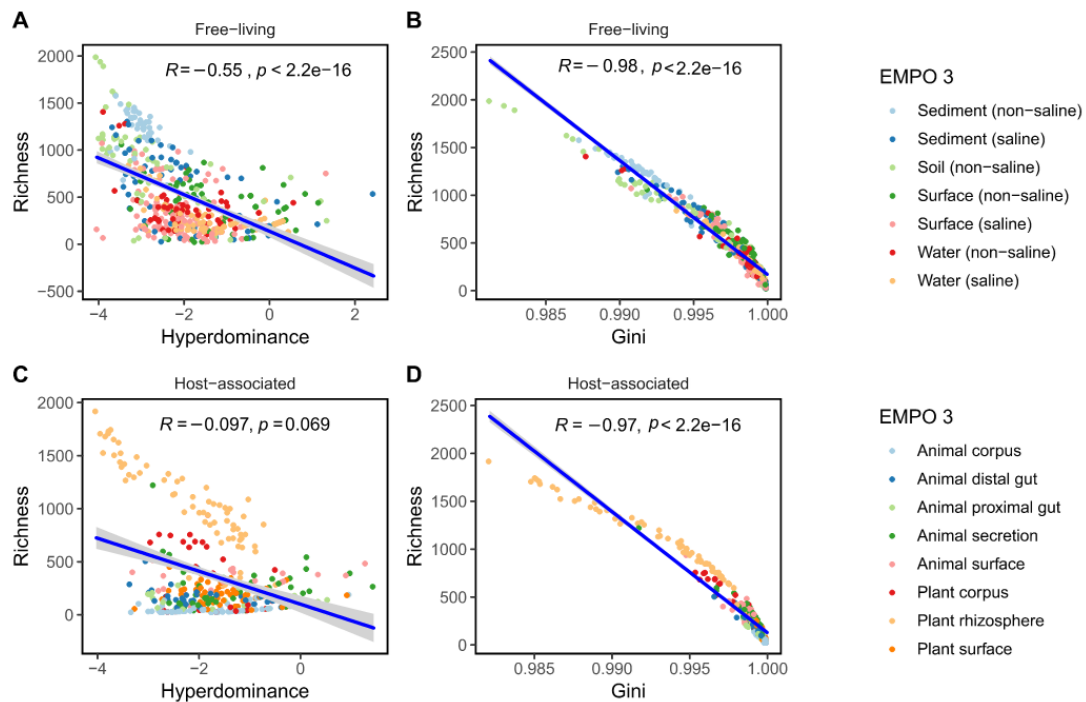
Figure 2: Relationship of richness (observed ASVs) with hyperdominance and inequality (Gini coefficient) in free-living and host-associated ecosystems. The correlations were calculated using Spearman's correlation. The blue lines depict a linear regression line with shaded regions showing the confidence interval at 95%. **A]** Richness versus hyperdominance for free-living microbial communities. **B]** Richness versus inequality as measured by the Gini coefficient for free-living microbial communities. **C]** Richness versus hyperdominance for host-associated microbial communities. **D]** Richness versus inequality as measured by the Gini coefficient for host-associated microbial communities.

Hyperdominance is a characteristic feature of human microbiomes

Using publicly available data, including both 16S rRNA gene amplicon sequencing and shotgun metagenomic data from large scale studies, we aimed to identify the hyperdominant bacteria in human microbiomes. For population level identification of bacteria that are generally hyperdominant in the human skin, oral and gut microbiome, we investigated the human microbiome dataset from the EMP 16S rRNA gene catalogue. Overall, 5.5%, 6.2% and 10.0% of the total ASVs present in the dataset were part of the hyperdominant taxa in at least one individual in oral, skin and gut microbiota, respectively. The top four hyperdominant ASVs identified in oral skin and gut microbiota are given in Table S1. Both oral and skin are habitats, which are reported to have higher richness compared to gut microbiota (Li, *et al.* 2012). The higher percentage of ASVs that can become part of the hyperdominant fraction within the gut

of individuals suggests that roles are more interchangeable within the gut microbiota (Turnbaugh*, et al.* 2009, Li*, et al.* 2014).

To cross-validate our observations based on amplicon sequencing data, we analysed publicly available shotgun metagenomic data sets (Huttenhower*, et al.* 2012, Schirmer*, et al.* 2016, Nielsen*, et al.* 2014). Here we made use of the fact that data have been processed and made readily available in standardised format as a part of the curatedmetagenomicDB package (Pasolli*, et al.* 2017). We observed that irrespective of the nationality, on average 5.5 ± 0.07% (at 95% CI, lower CI boundary = 5.4 and higher CI boundary = 5.7, 1000 replicates) of the metagenomic species (MGS) accounted for 50% cumulative relative abundance. This confirmed our findings from the ASV data about inequality and that hyperdominance is a characteristic of gut microbial communities (Figure 3A and 3B).
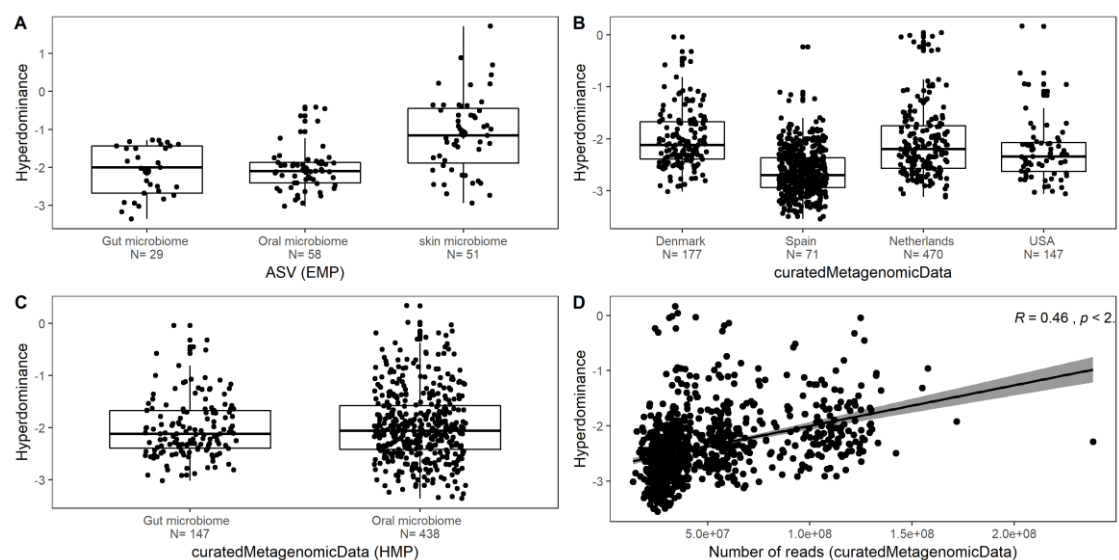

Figure 3: Hyperdominance observed in metagenomic datasets. **A]** Observed hyperdominance at different body sites using 16S rRNA gene amplicon sequencing data (n = 138)**. B]** Hyperdominance in gut metagenomes of healthy adults from Denmark, Spain, Netherlands and USA (n = 865). **C]** Hyperdominance in oral (n = 438) and gut (n = 147) samples based on Human Microbiome Project (HMP) metagenomic data. **D]** Relationship between hyperdominance and number of reads per samples in human gut metagenomic data.

In the oral microbiome HMP metagenomic data, we observed hyperdominance with a mean MGS coverage of 4.3 ±0.1% at 95% CI, bootstrap 1000 replicates) (Figure 3C). These values are similar to those observed in the ASV data for oral microbiome. In metagenomic datasets from the gut, we observed that at higher sequencing depth, hyperdominance becomes more prominent (Figure 3D). In line with our previous

observations in ASV data from the gut (Table S1), the hyperdominant fraction in metagenomic data was mostly comprised of bacteria from the phyla Bacteroidetes and Firmicutes (Table 1). These observations indicate a close association between the human host and bacteria from several families within the phyla Bacteroidetes and Firmicutes (viz. Bacteroidaceae, Lachnospiraceae and Ruminococcaceae). Bacteria from the genera *Bacteroides*, *Eubacterium*, *Alistipes*, *Faecalibacterium* and *Ruminococcus* are well known fermenters in the human gut microbiome. Some of these, such as *Bacteroides ovatus*, *B. uniformis*, *Eubacterium rectale* and *Ruminococcus bromii,* are known for their ability to degrade complex polysaccharides.

Some strains of *Faecalibacterium prausnitzii* have also been shown to utilize long chain inulin (Flint*, et al.* 2012a). Members of the genus *Bacteroides* have been shown to be characterized by high adaptability to the gut environment, including their ability to rapidly adapt within an individual's microbiome (Zhao*, et al.* 2017). Moreover, a recent study has identified that 40% of the species of the phylum Bacteroidetes (mostly species within the genera *Bacteroides* and *Parabacteroides*) and three *Alistipes* spp. are generalists (Vieira-Silva*, et al.* 2016). This would explain the observed hyperdominance of ASVs belonging to *Bacteroides, Faecalibacterium* and *Alistipes* in the human gut microbiome.

Hyperdominant taxa can also be conditionally rare

In the following, we assessed the stability of the dominance of the hyperdominant fraction over longer time scales. To this end, we investigated the dynamics of the hyperdominant fraction in publicly available dense time series data for human skin, oral and gut microbiomes (Caporaso*, et al.* 2011). We calculated the coefficient of variation for relative abundance for each of the ASVs across time to check for stability. We observed that most of the ASVs that are part of the hyperdominant fractions show low variation across time (Figure 4A-C). Interestingly, in gut and skin samples, we observed few ASVs with low mean abundance, but high coefficient of variation (Figures 4A and 4C). These ASVs are particularly interesting, as these are rare for the majority of the time and only increase in abundance at some time-points. In order to further explore these conditionally rare ASVs, we used the framework of conditionally rare taxa (CRT) as described previously (Shade*, et al.* 2014) . Using the same data from one individual

(M3), we observed that nine out of the 175 hyperdominant ASVs in skin are also part of the conditionally rare fraction. In the gut, we observed that six out of 79 hyperdominant ASVs are part of the conditionally rare fraction. This observation further supports the notion that certain ASVs are part of the hyperdominant fraction only occasionally (Table 2).

Table 1: Hyperdominant metagenomic species (MGS) in the human gut microbiome of different populations. The relative abundances were calculated separately for each population. Here**,** the top four MGS that had more than 25% of prevalence in hyperdominant fractions across samples are shown.

| Country	Hyperdominant MGS	Percentage of samples where MGS is part of hyperdominant fraction (%)	Relative abundance (proportions)			
			Maximum	Mean	Median	Std.dev
Spain	*Subdoligranulum* (unclassified)	57.8	0.50	0.08	0.07	0.08
	Faecalibacterium prausnitzii	53.5	0.14	0.05	0.05	0.03
	Ruminococcus bromii	42.3	0.31	0.05	0.03	0.06
	Eubacterium rectale	40.9	0.23	0.06	0.04	0.06
	Alistipes putredinis	32.4	0.10	0.03	0.03	0.03
Denmark	*Subdoligranulum* (unclassified)	67.8	0.41	0.11	0.09	0.09
	Eubacterium rectale	36.7	0.32	0.06	0.03	0.06
	Bacteroides uniformis	27.1	0.41	0.04	0.03	0.05
	Faecalibacterium prausnitzii	25.9	0.16	0.05	0.04	0.03
Netherlands	*Eubacterium rectale*	66.5	0.46	0.10	0.09	0.08
	Faecalibacterium prausnitzii	57.3	0.21	0.06	0.06	0.03
	Bifidobacterium adolescentis	56.7	0.71	0.08	0.06	0.09
	Subdoligranulum (unclassified)	42.7	0.32	0.05	0.04	0.05
USA	*Alistipes putredinis*	39.5	0.34	0.06	0.05	0.06
	Bacteroides ovatus	28.6	0.69	0.08	0.02	0.12
	Bacteroides uniformis	28.6	0.45	0.07	0.04	0.08
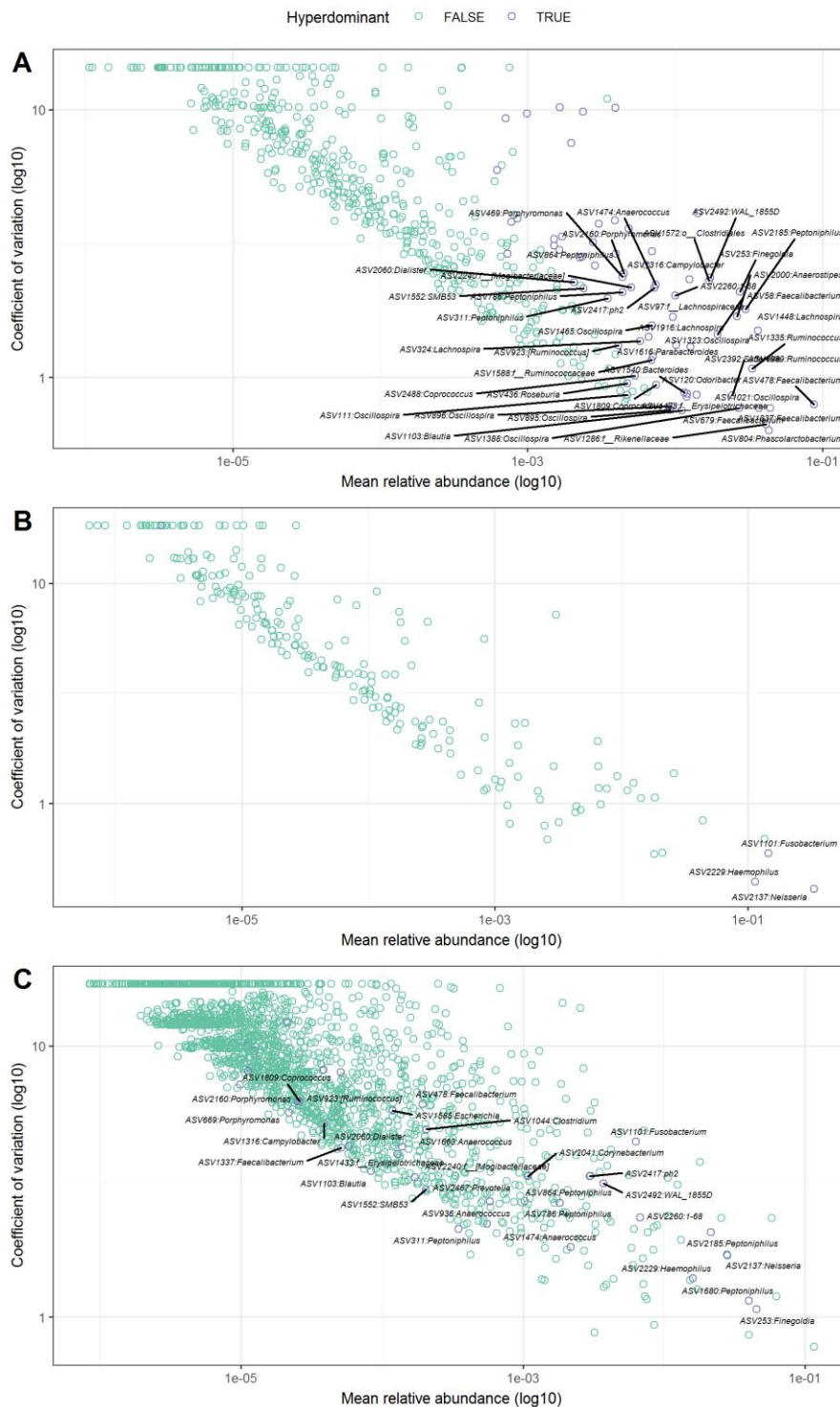
Figure 4: Stability of hyperdominant ASVs across time in M3 individual. **A]** Gut (N = 207)**. B]** Oral (N = 334). **C]** Skin (N = 287). The x and y axis values were log10 transformed for visualization.

Notably, we did not observe any conditionally rare ASVs in the oral samples, in line with the lack of hyperdominant ASVs with high coefficient of variation (Figure 4B). The

conditionally rare ASVs that are part of the hyperdominant fraction had a bimodality coefficient of more than 0.90, which can be suggestive of alternative states of these bacteria (Lahti, *et al.* 2014).

Table 2: List of conditionally rare ASVs identified in gut and skin samples of individual M3. The CRT were identified using a relative abundance threshold of 0.1. *The maximum relative abundance was calculated without taking into account the singletons. The number in parentheses in column 'Hyperdominant' is the number of times a given ASVs was part of the hyperdominant fraction.

Site	ASVs	Coefficient of bimodality	Maximum relative abundance*	Hyperdomin ant
Gut (N = 207)	ASV240:*Streptococcus*	0.99	0.16	No
	ASV606:o__Clostridiales	0.91	0.33	Yes (2)
	ASV1101:*Fusobacterium*	0.95	0.20	Yes (1)
	ASV1228:*Streptococcus*	0.96	0.53	No
	ASV1323:*Oscillospira*	0.91	0.17	Yes (5)
	ASV1837:*Coprococcus*	0.91	0.11	Yes (2)
	ASV2137:*Neisseria*	0.93	0.44	Yes (2)
	ASV2229:*Haemophilus*	0.94	0.17	Yes (1)
Skin (N = 287)	ASV157:*Lactococcus*	0.98	0.59	Yes (2)
	ASV584:*Diaphorobacter*	0.94	0.11	Yes (3)
	ASV867:*Flavobacterium*	0.91	0.19	Yes (2)
	ASV948:*Pseudomonas*	0.94	0.15	Yes (2)
	ASV956:*Pedobacter*	0.94	0.17	Yes (1)
	ASV980:*Janthinobacterium*	0.90	0.17	Yes (4)
	ASV1061:*Leuconostoc*	0.93	0.11	Yes (1)
	ASV1177:f__Leuconostocaceae	0.99	0.19	Yes (1)
	ASV1557:*Staphylococcus*	0.98	0.47	Yes (1)

The top hyperdominant ASVs fluctuated little in relative abundance and were dominant for the majority of the times sampled (Figure S5). It is suggested that such stability in abundance can have a stabilizing effect on the nutrient and energy flux (Henderson and Magurran 2014). In the particular case of the individual M3 gut time series data, we observed *Faecalibacterium* ASVs predominantly as abundant ASVs (Figure 5). Most of the hyperdominant bacteria such as those from genera *Bacteroides*, *Faecalibacterium*, *Eubacterium* and *Subdoligranulum* are known to have versatile metabolic features with the ability to grow on substrates commonly available in the colon (Holmstrøm, *et al.* 2004, Miquel, *et al.* 2013, Heinken, *et al.* 2014a, Cockburn, *et al.* 2015, Johnson, *et al.* 2017). Other bacteria have a more specific growth requirement and/-or occupy a small

niche. In the human gut, methanogens and sulfate reducing bacteria are good examples of bacteria with a defined niche.
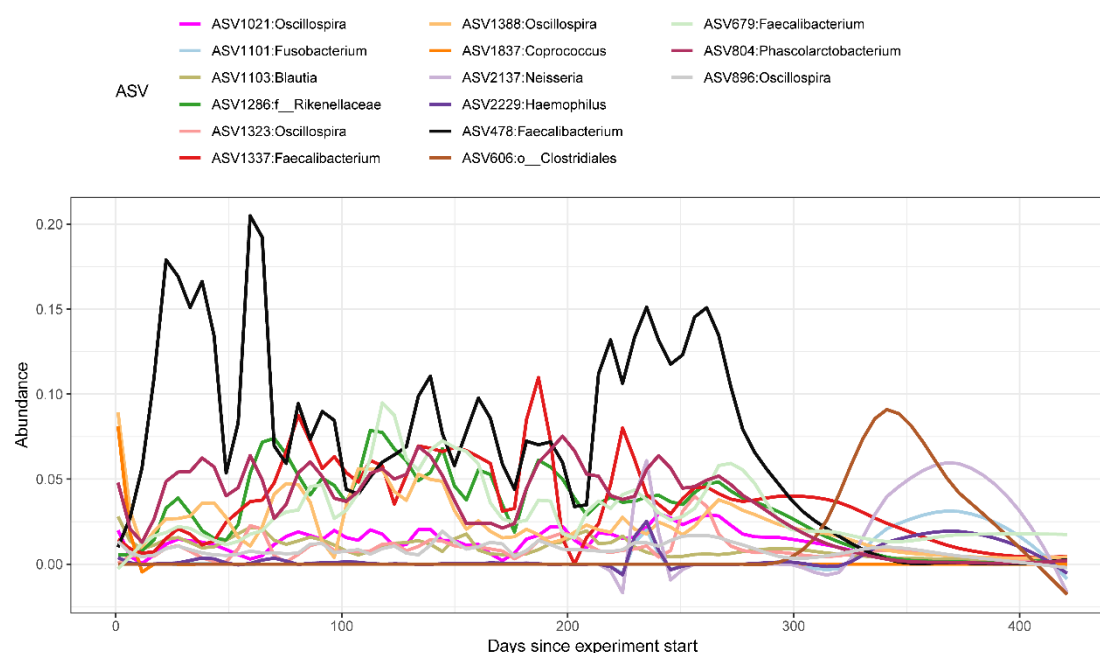

Figure 5: Dynamics of ASVs which are hyperdominant, part of the core and also including conditionally rare ASVs over more than 400 days of faecal sampling from a single individual (M3). The y-axis represents the relative abundance and x-axis represents the days of sampling. The line fits were obtained with loess smoothing ("geom_smooth" in R/ggplot2with a span of 0.1 (using 10% of the data at each step).

Although we did not detect methanogens in the human gut time series data, we observed that *Desulfovibrio* was consistently part of the rare rather than the dominant biosphere in both amplicon as well as cross-sectional metagenomic data (Figure S6). In addition, sulfate reducing bacteria are usually associated with inflammation and hence the host may have a role in "checking" for abundances of sulfate reducing bacteria (Rowan*, et al.* 2010). Therefore, having a limitation in niche size and selective pressure from the host may well be important deterministic processes shaping the microbial community structure in the human gut microbiome.

Immigration, perturbation events and patterns of hyperdominance

Analysis of hyperdominance in gut microbiota suggested that the hyperdominant species are functionally similar. It has been previously observed that neutral processes are likely to play an important role in dynamics of species sharing similar niches (see e.g. (Scheffer and van Nes 2006)). In addition, the assembly and dynamics of microbial

communities has been shown to be influenced by ecological processes such as immigration from the metacommunity and perturbation events (Mutshinda, *et al.* 2009, Hekstra and Leibler 2012).
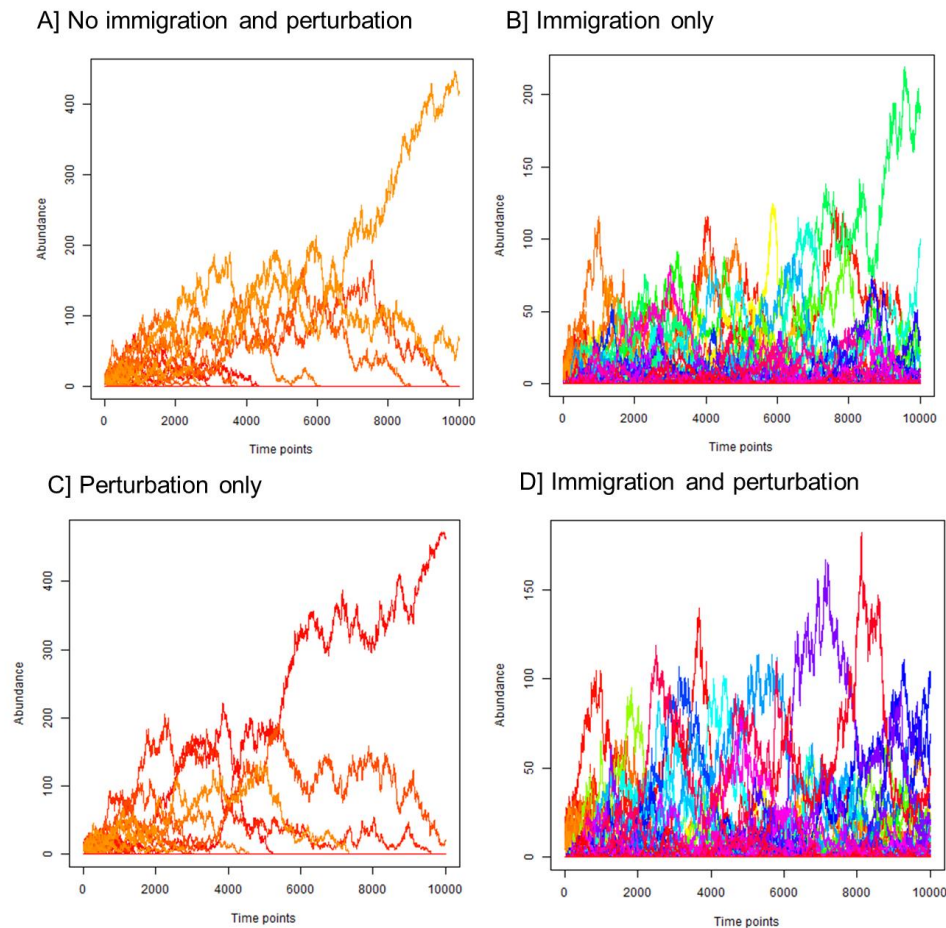

Figure 6: Simulation of community dynamics under a neutral model. Each line denotes abundance of different species in time; different colors correspond to different species in the community. **A]** Community dynamics in absence of immigration and perturbation. **B]** Community dynamics in presence of immigration and absence of perturbation. **C]** Community dynamics in presence of perturbation and immigration. **D]** Community dynamics in presence of both perturbation and immigration. The immigration rate was 0.02, and perturbations were modelled on daily time scales for 10,000 time points. The following parameters were set: number of species in local community, N=50; number of species in metacommunity, M = 500; number of deaths at each timepoint, d = 10.

Closed and unperturbed communities tend to result in dominance of single species, highlighting the important effects of immigration and environmental perturbations on community dynamics (Hekstra and Leibler 2012). Therefore, we were interested to know whether immigration and perturbation would result in specific patterns in changes in hyperdominant species and the observed rarity for the majority of the

species. We considered a neutral model of microbial community dynamics under four conditions a) no immigration and perturbation b) with immigration only; c) with perturbation only and d) with immigration and perturbation. To mimic the daily influence of diet on human gut microbiota, perturbations were modelled on a daily basis for 10,000 time points. In absence of both immigration and perturbation, the community reached to a state where a single species dominated and several species were lost (Figure 6A). In presence of only immigration, a pattern of changes in the dominant species emerged over time (Figure 6B). Interestingly, daily perturbations alone in absence of immigration had a negative impact on species diversity and lead to collapse of the community (Figure 6C). However, the most representative dynamics of frequent alternations in dominant species emerged when both immigration and perturbation events were acting on the community (Figure 6D). These events had a major role in maintaining species diversity and hyperdominance, and avoiding absolute dominance of a single species in natural environments.

Conclusion

The presence of a large rare biosphere in microbial communities has been generally accepted, however, to what extent dominance and hyperdominance are exhibited by specific taxa remains unknown. Here, using data from diverse microbial ecosystems made available through the EMP catalogue, we showed that on average 4.6% of all microbial ASVs dominate their respective communities. This observation is in line with the previous estimate that 2-6 % of species demonstrate hyperdominance in many ecosystems (manuscript in preparation (van Nes, *et al.*)). We also observed that ecological processes such as immigration and environmental perturbations give rise to fluctuations in the identity of the dominant species and natural patterns of oscillation in species abundances.

To avoid biases associated with PCR based approaches, we validated the amplicon-based results by analyzing metagenomic datasets and found close similarities in identity and fraction of hyperdominant taxa. This validation suggested that observed hyperdominance is not a result of variation in 16S rRNA gene copy number or biases associated with primers or other PCR-related technical biases. Furthermore,

metagenomic datasets also included populations from different European countries, and aided in identification of population-specific hyperdominant bacteria, suggesting the role of host-associated factors such as diet or lifestyle in selecting for the most dominant bacteria.

Certain traits can be associated with the hyperdominance potential of specific bacteria, such as, for example, the ability to take full advantage of the carrying capacity of the ecosystem by occupying the largest possible niche. Based on currently available evidence, it is clear that hyperdominant bacteria occupy the largest niche, i.e. degradation and utilization of complex polysaccharides that enter undigested in the colon. We also provided an example of *Desulfovibrio* species, the relative abundances of which in longitudinal and cross-sectional data suggest that rarity in the total community can be associated with a specialized and relatively small niche compared to hyperdominant taxa. However, these rare taxa can play an important role as *Desulfovibrio* species are known to influence the microbial community via reduction in H_2 concentrations (Rey*, et al.* 2013). *Desulfovibrio* spp. are also known to have metabolic cross-feeding with hyperdominant *Bacteroides*, as the latter encode sulfatases which release sulfate from host derived compounds (Rey*, et al.* 2013).

The data presented here are an important step towards a better understanding of dominance and rarity in microbial communities. Our results suggest that fluctuations in identity of hyperdominant taxa and avoidance of mono- dominance can be attributed to additive effects of immigration and perturbations.
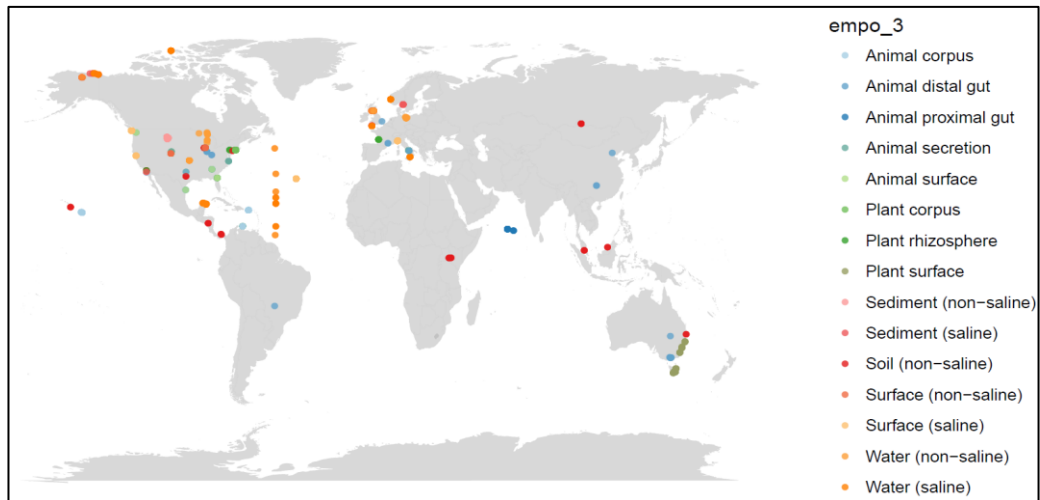
Acknowledgements

Supplementary information

Figure S1: Origin and type of samples analysed in the framework of the Earth Microbiome Project and used in this study, as based on the provided metadata (Thompson*, et al.* 2017).
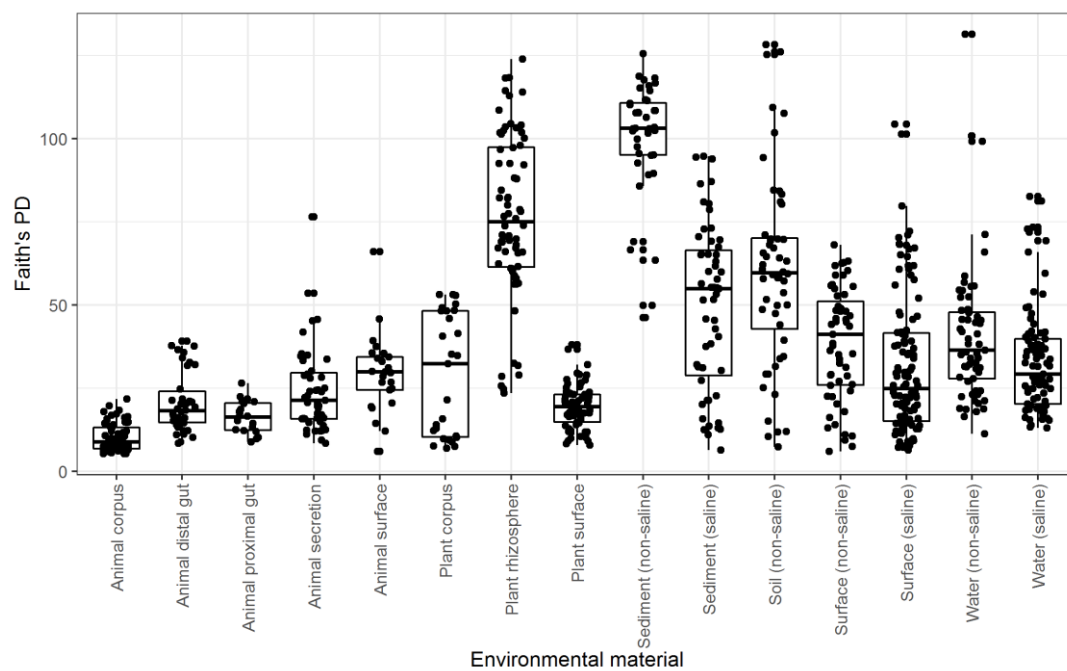

Figure S2: Comparison of Faith's phylogenetic diversity (PD) in samples from diverse origin classified according to the EMP ontology level 3 (Thompson*, et al.* 2017).
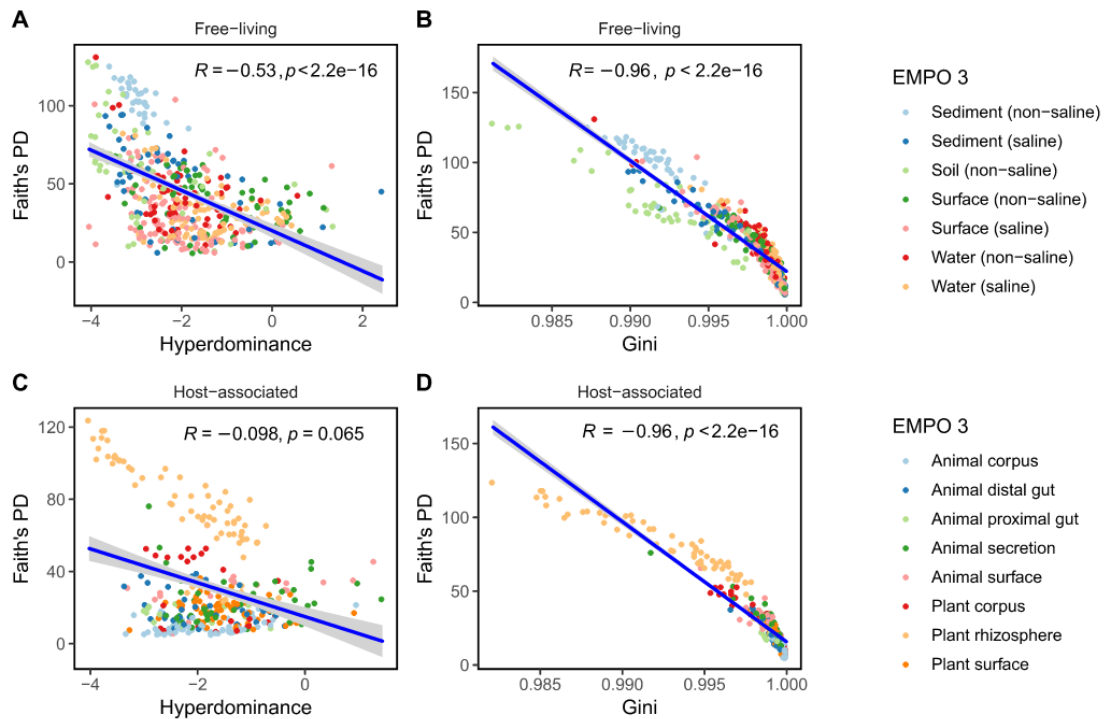
Figure S3: Relationship of Faith's phylogenetic diversity (PD) with hyperdominance and inequality in free-living and host-associated ecosystems. We calculated Spearman's correlation **A]** Faith's PD versus hyperdominance for free-living microbial communities. **B]** Faith's PD versus inequality as measured by the Gini coefficient for free-living microbial communities. **C]** Faith's PD versus hyperdominance for host-associated microbial communities. **D]** Faith's PD versus inequality as measured by the Gini coefficient for host-associated microbial communities.
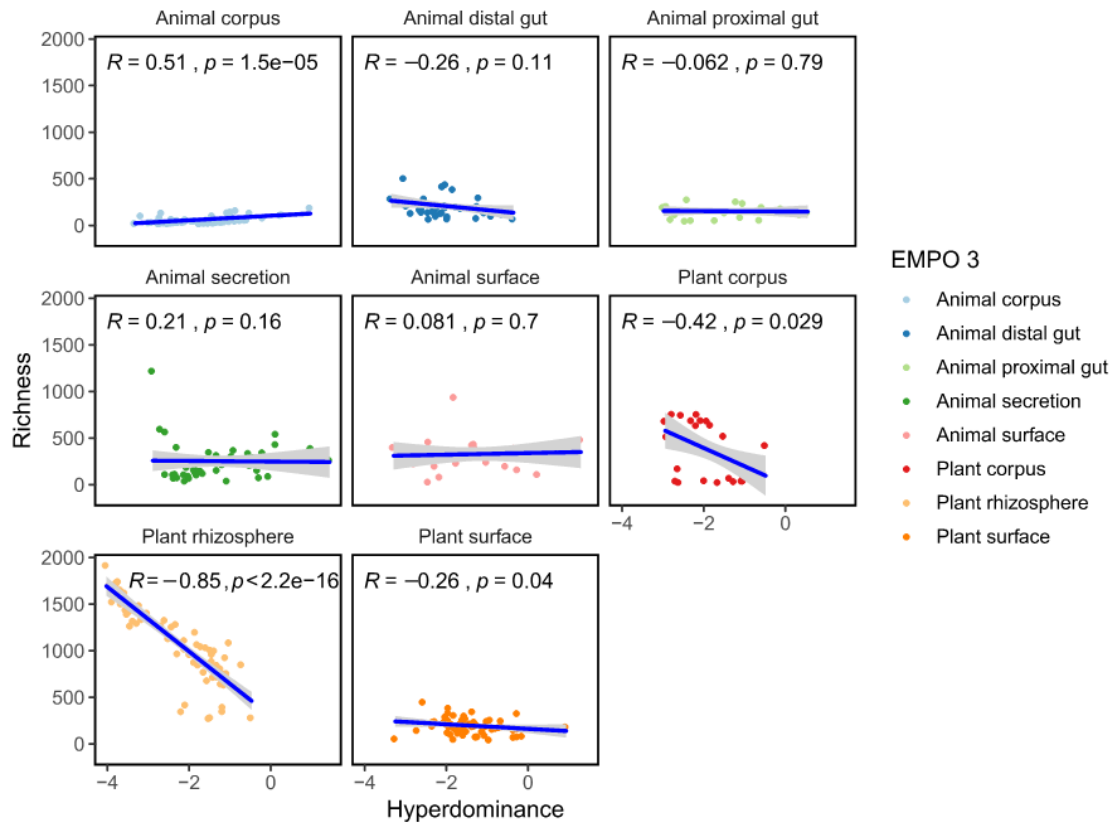
Figure S4: Relationship of richness with hyperdominance for specific habitats within host-associated ecosystems.
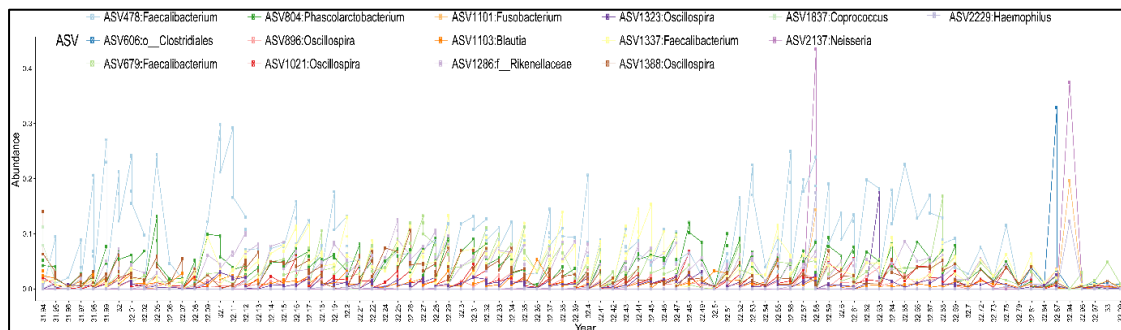

Figure S5: Dynamics of hyperdominant and conditionally rare ASVs in gut samples of subject M3. Here, we show changes in relative abundances. The x-axis represents the time-stamp of sampling of faeces and y-axis is the relative abundance in proportions.
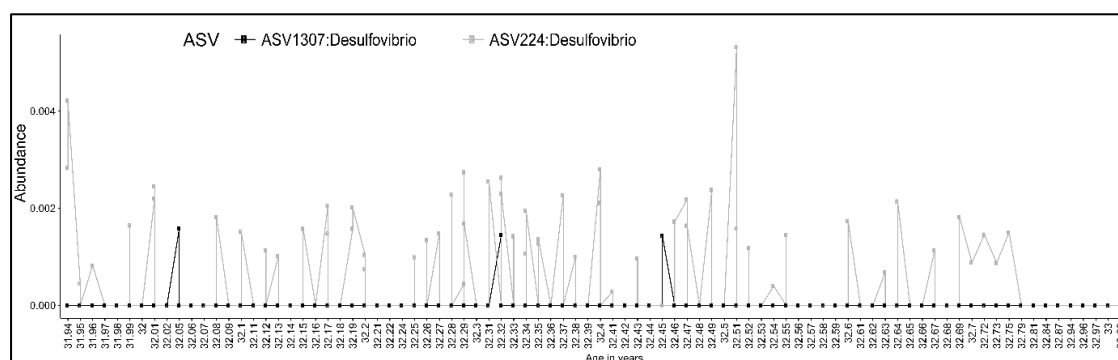
Figure S6: Relative abundance distribution of ASVs classified as members of genus *Desulfovibrio*. The data is for gut samples of individual M3. The x-axis represents the time-stamp of sampling of faeces and y-axis is the relative abundance in proportions. These ASVs were consistently observed to have a minimum relative abundance of 0.0044 (when detected) and 0.095 indicating their consistent rarity (in terms of relative abundance) in the gut ecosystem.

Table S1: Top four prevalent hyperdominant ASVs in the oral, skin and gut microbiota in humans.

Site	ASV	Percent of samples it is part of hyperdominant fraction (%)	Relative abundance (proportions)			
			Maximum	Mean	Median	Std. dev
Oral (N = 59)	ASV1751:Streptococcaceae	59.32	0.82	0.16	0.09	0.18
	ASV753:Pasteurellaceae	52.54	0.38	0.10	0.08	0.08
	ASV2071:Prevotellaceae	40.68	0.27	0.08	0.07	0.07
	ASV3053:Neisseriaceae	32.20	0.58	0.07	0.03	0.10
Skin (N = 51)	ASV1751:Streptococcaceae	41.18	0.61	0.07	0.03	0.12
	ASV706:Staphylococcaceae	35.29	0.71	0.07	0.03	0.13
	ASV1088:Corynebacteriaceae	27.45	0.17	0.03	0.02	0.03
	ASV753:Pasteurellaceae	19.61	0.06	0.01	0.00	0.02
Gut (N = 30)	ASV2820:Bacteroidaceae	53.33	0.46	0.13	0.05	0.16
	ASV1624:Lachnospiraceae	26.67	0.19	0.03	0.01	0.05
	ASV1813:Bacteroidaceae	26.67	0.22	0.04	0.01	0.06
	ASV3595:Bacteroidaceae	26.67	0.18	0.03	0.02	0.04

Chapter 4

microbiomeutilities: An R package for utilities to guide in-depth marker gene amplicon data analysis

Sudarshan A. Shetty, Leo Lahti, Willem M. de Vos and Hauke Smidt
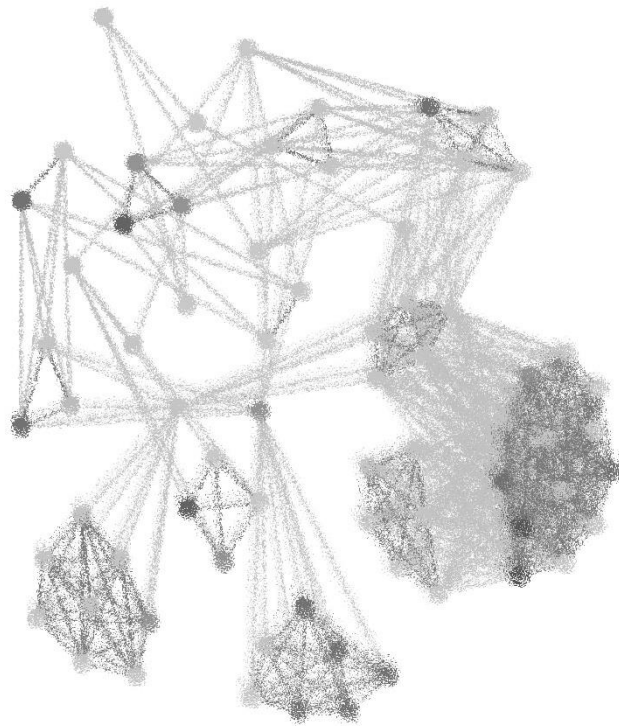
Manuscript under preparation

Abstract

Amplicon sequencing is widely used for microbial community profiling studies. Several tools have been developed, many of them in the R statistical programming environment, for analyzing amplicon sequencing data. Tools such as `phyloseq' can store data related to phylogenetic marker gene sequencing and provide a basic data structure for extending data analysis and visualization in this context via new tools. One such tool based on the phyloseq-data format is the recently developed 'microbiome' R package, which is part of the Bioconductor' repository. This R package provides further extension and, in some cases, alternative functions to analyze 16S rRNA gene count data. Here, we expand this suite of tools by the development of the R package 'microbiomeutilities that eases the initial steps in processing operational taxonomic unit (OTU) tables in R by combining the functionality of widely used R packages. The package provides user-friendly functions for formatting and manipulating the OTU and taxonomy tables. The automated HTML report generation provides a first-hand, quick visual glance into the quality of count data, taxonomic composition, alpha diversity measures as well as beta diversity analyses, including multivariate ordinations. The HTML report records all user inputs and graphical outputs for tracking, sharing and reproducing results of the analysis. In addition, we developed a web-based version of 'microbiomeutilities' for users who lack experience with R software. The microbiomeutilities R package is a free, open source software and a useful extension to the existing R packages for microbial community analysis.

Keywords: Microbiome, R tools, 16S rRNA gene sequencing, Data analytics

Introduction

High throughput sequencing of marker genes, especially the 16S ribosomal RNA (rRNA) gene, has now become a routine approach in profiling the composition of microbial communities. With large amounts of sequence data, challenges related to data handling and analysis have been at the forefront of software tool development. Numerous tools, from command line interfaces such as Mothur (Schloss*, et al.* 2009) and QIIME (based on Python) (Caporaso*, et al.* 2010), web-based tools such as Calypso (Zakrzewski*, et al.* 2016) and MicrobiomeAnalyst (Dhariwal*, et al.* 2017), and standalone GUIs (graphical user interfaces) using the Shiny package in R such as shiny-Phyloseq (McMurdie PJ and S 2014) and DAME (Piccolo*, et al.* 2017) have been developed. Packages created in the R statistical programming environment are of specific interest as these require less programming experience, and with Rstudio the interface to using R software packages is user friendly (Venables and Smith 2006). In addition, using `Rmarkdown` (https://rmarkdown.rstudio.com/), sharing of complete project data analysis procedures is convenient. There are numerous R statistical programming environment packages for microbial community analysis, and one of the most widely used is the 'phyloseq' package, which allows incorporation of phylogenetic data. A detailed list of relevant R packages can be found in Supplementary Table S1.

Here, we developed an R package called, 'microbiomeutilities' which provides new functionalities and convenient visualization tools for downstream analysis of high throughput 16S rRNA gene sequencing data. This package can be used seamlessly with existing R packages such as phyloseq and microbiome.

Functional description and advantages

The latest version of the `microbiomeutilities` R package is available from the GitHub repository. An overview of the `microbiomeutilities` functionality and workflow is shown in Figure 1.

Data Handling and Formatting

Here, we use human fecal microbiota profiling data from normal, colorectal cancer and non-colorectal cancer subjects (Zackular*, et al.* 2013). The processed data set used in the `microbiomeutilities` R package was obtained from a recent meta-analysis study

(Duvallet*, et al.* 2017). Often naming of taxonomic id's is inconsistent in different databases.
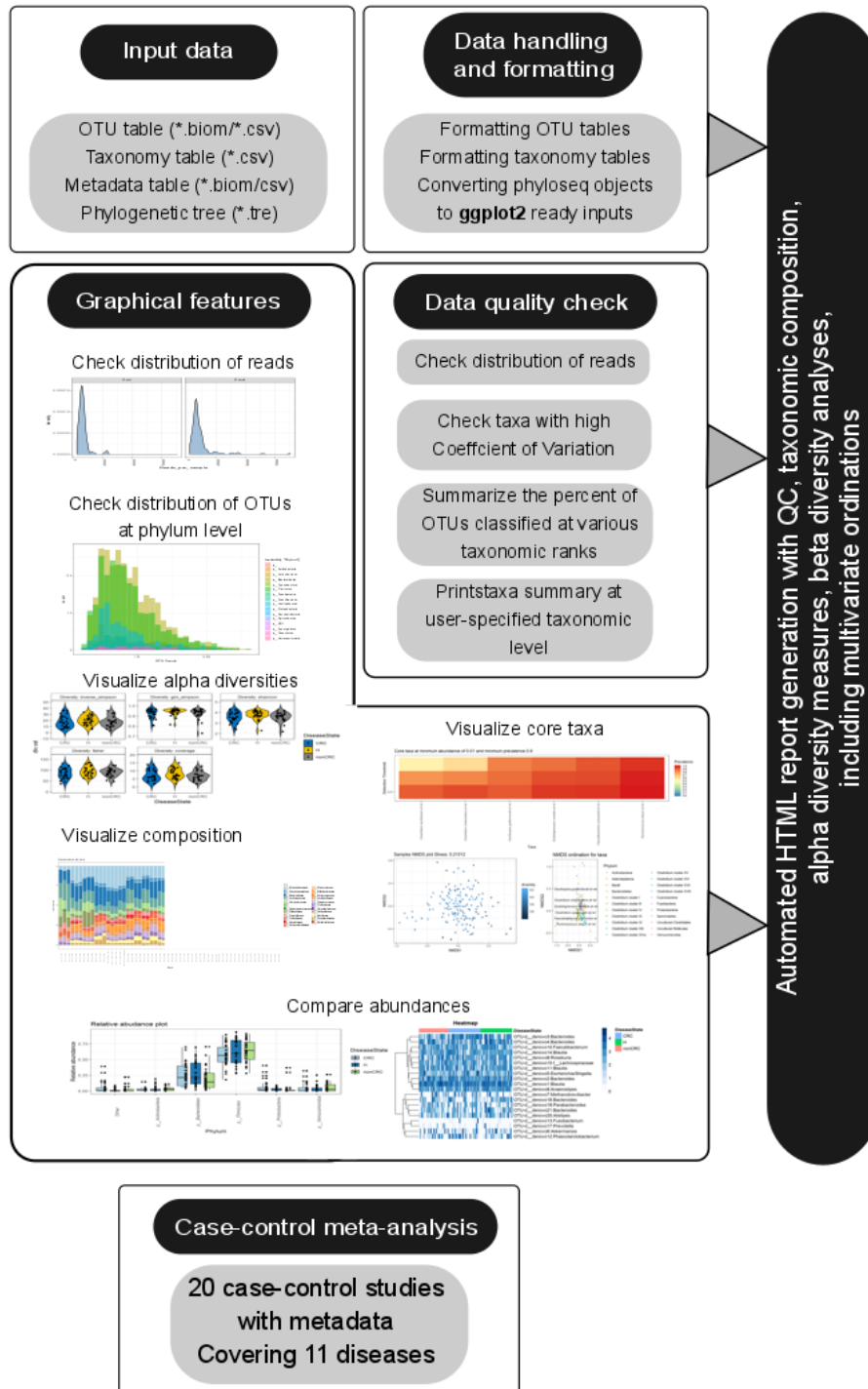

Figure 1: Workflow and overview of functionality of the `microbiomeutilities` R package. A list of functions for data handling, analysis and visualization is provided in Table S2.

The taxa IDs are usually numbers (e.g. OTU-99301) and difficult to interpret. Additionally, it is commonly observed that the taxonomy file has classification until a given taxonomic level. To make the interpretation easier, functions were implemented that can be used to clean the taxonomy tables to more readable outputs. For e.g. if a given Operational Taxonomic Unit (OTU) or Amplicon Sequence Variant (ASV) is classified up to family level then the name is given as OTU-99301:f_Lactobacillaceae-, alternatively, if the OTU is classified at genus level then the name is given as OTU13:*Fusobacterium*.

Next, knowing the overall percentage of OTUs/ASVs classified at different taxonomic levels in the dataset can be useful to check for resolution of the hypervariable region sequenced and/-or the quality of different taxonomic classifiers. The *percent_classified* function can provide a summary of classified and unclassified OTUs/ASVs at each of the taxonomic ranks (from Kingdom to species level). In addition, a summary of taxonomic data can be obtained using the *taxa_summary* function at any user-specified taxonomic level.

Phyloseq class objects cannot be used directly as input for widely used data handling and visualization packages such as `tidyverse` and `ggplot2` (Wickham 2017, Wickham 2011). Converting the information contained in `phyloseq` objects to a long data format can be useful for taking advantage of these data science R packages. For this and other purposes, we have created accelerated versions of existing methods.

Data Visualization

For quality check and pre-processing of the OTU/ASV data, the `microbiomeutilities` package consists of two functions, *plot_read_distribution* and *plot_taxa_cv*. The plot_read_distribution function gives the user first hand insights into library sizes within user defined sample groups, e.g. case and control group, whereas *plot_taxa_cv* can act as a visual guide to identify potentially spurious observations, i.e. those with a trivially high coefficient of variation.

As a utility function to support the `microbiome` R package diversity estimations, *plot_alpha_diversities* can be used to visualize differences in alpha diversity measures between groups of interest. In addition, we provide a user-friendly utility function to

plot commonly used visualizations, including taxa abundance plots, heatmaps and ordination with taxa loadings (Figure 1).

Preliminary Report Generation

The early steps in analyzing raw sequencing reads are commonly done using Mothur (Schloss*, et al.* 2009) QIIME and QIIME2 (Caporaso*, et al.* 2010), DADA2 (Callahan*, et al.* 2016), NG-Tax (Ramiro-Garcia*, et al.* 2016), or PANGEA (Giongo*, et al.* 2010). These tools use various algorithms for identifying i) OTUs when using clustering approaches or ii) ASVs or subOTUs when using error correcting models in DADA2 and Deblur (Callahan*, et al.* 2016, Amir*, et al.* 2017). Downstream analysis of OTU or ASV tables requires careful investigation of data properties such as sequencing depth, distribution of OTUs/ASVs in the data, identifying potential spurious OTUs/ASVs and variance in OTU/ASV distribution when identifying differentially abundant OTUs/ASVs. Knowledge of these properties is critical for choosing appropriate filtering and downstream statistical analyses. However, quality control steps can be time consuming especially for beginners. To shorten the time and manual effort required for such analysis, we introduce tools that generate an HTML report to have a first-hand, quick visual glance into the quality of count data, taxonomic composition, alpha diversity measures as well as beta diversity analyses, including multivariate ordinations. The microbiome_pipeline function saves all intermediate files, including figures and phyloseq objects in a user specified directory. Thus, the automated report generation helps in tracking back the analysis steps and data filtering parameters to make sure that the user is able to fine-tune and reproduce the analysis.

As an example we, we used data from mock communities (MCs) introduced to benchmark the NG-Tax pipeline (Ramiro-Garcia*, et al.* 2016), and routinely used in our laboratory for quality control of 16S rRNA gene sequencing runs. These MCs are synthetic bacterial communities of known composition. The dataset represents sequences from three different Illumina Hiseq runs spread over seven libraries generated with two different primer pairs covering the 16S rRNA gene variable regions V4 (F515-R806) and V5-V6 (F784-R1064) and different PCR settings (25,30 and 35 cycles and pooling of triplicate PCR reactions or a single reaction). More information with

regards to PCR settings and generation of the data can be found in the original publication that introduces NG-Tax (Ramiro-Garcia*, et al.* 2016). An example report generated by the *microbiome_pipeline* function using the mock communities (MCs) is provided as supplementary file. The online interactive version can be accessed at URL: https://goo.gl/1bkqWF.

MicrobiomeHD As User-Friendly Phyloseq Objects

Numerous case-control studies exist, however, accessing and analyzing the vast amount of data for statistical analysis is a hurdle. Convenient access to data for development and testing statistical approaches or for large-scale meta-analysis is crucial. For statisticians and epidemiologists with basic skills in R and no experience with processing sequencing data, availability of data from case-control studies in a format that is accessible in R can be highly convenient. Thus, we have converted the publicly available MicrobiomeHD (Duvallet*, et al.* 2017) database into a common phyloseq data class. We have assembled 20 case-control studies covering 11 diseases which can be conveniently accessed using the *get_microbiome_data* function (Table 2). While we provide most of the meta-data associated with the samples, users are advised to read the original publications listed in the table for more information, also to make sure that data is indeed comparable (e.g. with respect to methodological aspects of sample processing, DNA extraction, PCR and sequencing).

Web-Based Shiny Application

In many instances, the user may not be inclined towards learning a programming language as it may not outweigh the disadvantages of not knowing one. For example, undergraduate students as well as microbiologists focusing on wet lab experiments such as microbial community enrichments and defined mixed cultures or synthetic microbial communities often lack the necessary programming background. Analyzing microbial community composition is a relatively small part of the larger experiment, let alone undergraduate teaching modules, and investing time in learning a new programming language may not be a priority. In such instances, availability of a user-friendly online tool that does not require any software installation to carry out basic
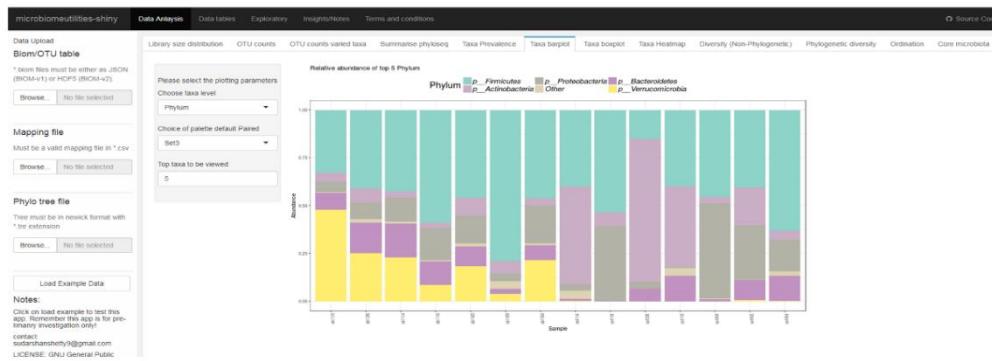
analysis such as taxonomic composition, identifying dominant taxa and differences in community structure under different conditions is of prime importance.

Table 2: List of case-control studies included as phyloseq objects in microbiomeutilities. The database is available at the following URL: https://github.com/microsud/microbiomedatarepo. ASD, autism spectrum disorder; CDI, Clostridium difficile infection; CRC, colorectal cancer; EDD, enteric diarrheal disease; HIV, human immunodeficiency virus; LIV, liver diseases; NASH, non-alcoholic steatohepatitis; OB, obesity; PAR, Parkinson's disease; T1D, type I diabetes; IBD, inflammatory bowel disease.
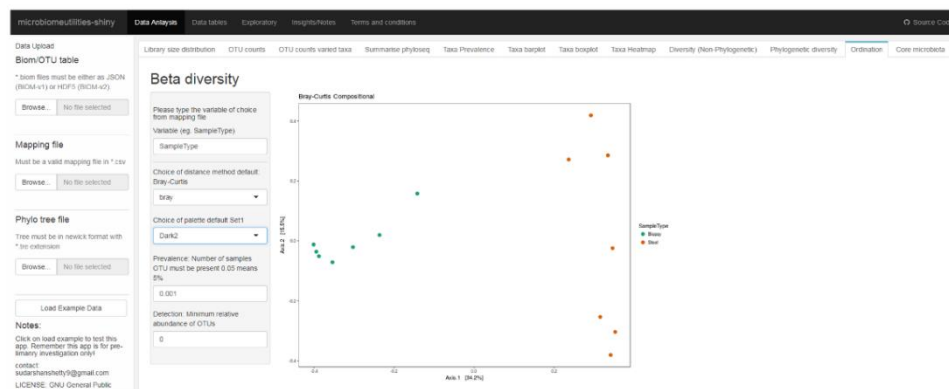
Study	Disease	Reference
Son2015_ASD	ASD	(Son, *et al.* 2015)
Kang2013_ASD	ASD	(Kang, *et al.* 2013)
Schubert2014_CDI	CDI	(Schubert, *et al.* 2014)
Youngster2014_CDI	CDI	(Youngster, *et al.* 2014)
Baxter2016_CRC	CRC	(Baxter, *et al.* 2016)
Zackular2014_CRC	CRC	(Zackular, *et al.* 2013)
Zeller2014_CRC	CRC	(Zeller, *et al.* 2014)
Singh2015_EDD	EDD	(Singh, *et al.* 2015)
NogueraJulian2016_HIV	HIV	(Noguera-Julian, *et al.* 2016)
Dinh2015_HIV	HIV	(Dinh, *et al.* 2014)
Lozupone2013_HIV	HIV	(Lozupone, *et al.* 2013)
Gevers2014_IBD	IBD	(Gevers, *et al.* 2014)
Zhang2013_LIV	LIV	(Zhang, *et al.* 2013)
Wong2013_NASH	NASH	(Wong, *et al.* 2013)
Ross2015_OB	OB	(Ross, *et al.* 2015)
Zupancic2012_OB	OB	(Zupancic, *et al.* 2012)
Scher2013_PAR	PAR	(Scher, *et al.* 2013)
Alkanani2015_T1D	T1D	(Alkanani, *et al.* 2015)
Scheperjans2015_PAR	PAR	(Scheperjans, *et al.* 2015)

For this purpose, we created a web application, which is based on the R `shiny` app (Cheng 2013) and the `flexboard` R package (https://goo.gl/B59uzf). The web application is hosted online on the freely available `shiny` app server (www.shinyapps.io). The `microbiomeutilities-shiny` web application can be accessed at the following URL: https://gutmicrobe.shinyapps.io/microbiomeutils-shiny/. The user-interface and some basic functionality are depicted in Figure 2.

Figure 2: The web interface for the microbiomeutilities-shiny app hosted on the shiny apps server. **A]** A relative abundance composition barplot. **B]** Ordination plot based on community dissimilarities. **C]** Page two depicting various data tables, with on the right-hand side the properties of data uploaded by the user.

Conclusion

The `microbiomeutilities` tool is a user-friendly R package that supplements existing R packages like `phyloseq` and `microbiome` (McMurdie and Holmes 2013, Lahti, *et al.* 2014, Lahti and Shetty 2018). The major functional extension provided by our package

to the existing R packages used for microbiome analysis is the generation of an HTML report with information on data properties such as library sizes, OTU count properties and taxonomic composition. It also includes tools for preliminary visualization for alpha and beta diversity. The entire report is generated in less than 1 minute (on average 56 seconds on a 64-bit Windows 7 operating system with 24 GB RAM, calculated using `microbenchmark` R package) for a dataset with 57 samples and 5,251,399 reads. This report is convenient for undergraduate students and biological researchers with little programming experience. The report also serves as a standardized template for common steps in microbial community analysis. The arguments provided in individual commands in the code section can be copy pasted into the terminal, and changes to arguments can easily done by the user. Furthermore, several utility functions for data visualization are provided, which are expected to help novice users.

The tools developed here are an important step towards transparency, standardization and reproducibility of all data analysis details. The interactive web interface can be integrated to other analysis pipelines such as NG-Tax to provide a complete workflow from raw reads to data analysis. The future development of the package will include addition of the case-control datasets for human intestinal microbiota surveys to cover diverse diseases associated with intestinal microbiota. These datasets will be useful for researchers interested in developing methods for identifying microbiota based biomarkers. Convenient access to such large collections of case-control datasets will aid in testing and validating novel methods. The tools developed here are a community resource and will be constantly updated to add additional functionalities as and when required.

Availability

The 'microbiomeutilities' package is implemented in R (as of version 3.4.4) and is available from GitHub (as of version 0.99.0) at URL: https://github.com/microsud/microbiomeutilities. A step-wise installation guide and usage instructions are available online at URL: https://goo.gl/9akCJT. The 'microbiomeutilities' R package is openly available under the 2-Clause BSD License wherein redistribution and modifications are permitted under minimal restrictions

(Morin*, et al.* 2012). The package is independent of the operating system and requires only R and Bioconductor software environments.

Acknowledgments

Supporting information

Supplementary Table S1: List of relevant tools available for use in the R/BioC software environment.

Sr.no	R/BioC package	Feature	References
1	ampvis2	Tools for visualising amplicon sequencing data. Powerful multivariate ordinations and heatmap visualisations	(Andersen KSS, *et al.* 2018)
2	deseq2	Differential expression analysis for sequence count data. Commonly used for RNA-seq data and now applied to marker gene count data	(Love, *et al.* 2014)
3	mare	*Microbiota Analysis in R Easily*. Handles raw reads, creates taxonomies, OTU tables and statistical analysis	(Korpela 2016)
4	metacoder	A novel way for data handling and visualizing the hierarchical structure of taxonomic classifications in community profiling data	(Foster, *et al.* 2017b)
5	metagenomeseq	Differential abundance analysis for microbial marker-gene surveys. Addresses zero-inflation in sparse count data	(Paulson, *et al.* 2013)
6	microbiome	Tools for microbiome analysis in R, extends and provides novel data analysis, manipulation and visualisations	(Lahti and Shetty 2018)
7	mixMC	Multivariate statistical analysis for compositional and sparse data. Includes function to carry out sparse Partial Least Squares Discriminant Analysis (sPLSDA)	(Le Cao, *et al.* 2016)
8	pathostat	R shiny package for statistical analysis of metagenomics sequecing data	(Manimaran, *et al.* 2018)
9	phylofactor	Phylogenetic factorization of compositional data which incorporates phylogeny in dimension reduction analysis	(Washburne, *et al.* 2017)
10	phylogeo	Integrated geographic analysis and visualization of microbiome profilling data	(Charlop-Powers and Brady 2015)
11	phyloseq	The most widely used and cited package for analysing microbiome census data in R. Several recent packages use the phyloseq-class objects as a base for extending and providing new functions	(McMurdie and Holmes 2013)
12	Shiny-phyloseq	Web-tool with user interface for Phyloseq	(McMurdie PJ and S 2014)
13	SigTree	Identifies and visualizes significantly responsive branches in a phylogenetic tree	(Stevens, *et al.* 2017)
14	spiec-easi	Includes methods for handling sparse and compositional microbiome data to infer ecological interactions	(Kurtz, *et al.* 2015a)
15	structSSI	A novel approach for simultaneous and selective inference for grouped or hierarchically structured data	(Sankaran and Holmes 2014)
16	labdsv	Ordination and multivariate analysis for ecology. Provides numerous functions for multivariate analysis and modeling species distributions	(Roberts and Roberts 2016)
17	vegan	R package for community ecologists, widely used for microbial community analysis	(Oksanen, *et al.* 2013)
18	Rhea	A pipeline with modular scripts. A useful set of scripts that can be modified by the user for personalising the analysis	(Lagkouvardos, *et al.* 2017)

| 19 | theseus | Extends functionality of phyloseq and microbiome R package for visualisating and analysing microbial community data | (Price, *et al.* 2018) |

Supplementary Table S2: Functions provided within microbiomeutilities.

Function	Description
format_to_besthit()	Formats the taxonomy table within a phyloseq-class object
get_microbiome_data()	Downloads test microbiome profiling datasets
list_microbiome_data()	Provides a list of available microbiome profiling datasets
microbiome_pipeline()	A wrapper function for generating reports of microbiome analysis
percent_classified()	Summarize the percent taxa classification for phyloseq-class
phy_to_ldf()	Converts phyloseq-class object to long data format
plot_alpha_diversities()	Creates a plot for alpha diversity indices calculated using the microbiome R package
plot_ordiplot_core()	Plots core microbiota on ordinations
plot_read_distribution()	Plots distribution of sequencing reads in user specified groups
plot_select_taxa()	Generates a boxplot for user specified list of taxa
plot_taxa_boxplot()	Taxonomic composition comparisons between user specified groups
plot_taxa_composition()	Taxonomic composition barplot or lineplot, legacy function from original microbiome R package
plot_taxa_cv()	Calculates and plots coefficient of variation for each OTU/ASV
plot_taxa_heatmap()	Plots a heatmap using phyloseq-class and pheatmap
print_ps()	Prints an overview of the content of the phyloseq-class object
taxa_summary()	Prints taxa summary at specified taxonomic level
zackular2014	Test data used for illustrating the microbiomeutilities package functionality

Chapter 5

Comparative genomics and physiology of the butyrate-producing bacterium *Intestinimonas butyriciproducens*

Thi Phuong Nam Bui, Sudarshan A. Shetty, Ilias Lagkouvardos, Jarmo Ritari, Bhawani Chamlagain, Francois Douillard, Lars Paulin, Vieno Piironen, Thomas Clavel, Caroline M. Plugge, Willem M. de Vos
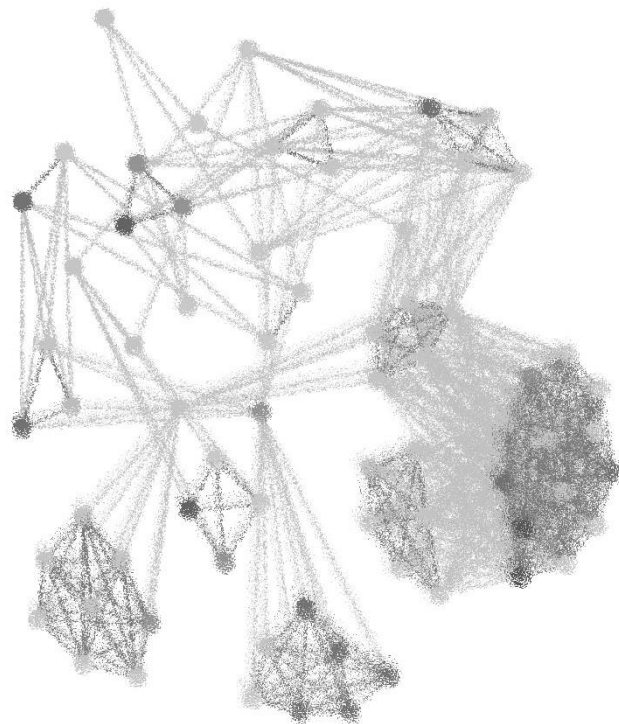
Abstract

Intestinimonas is a newly described bacterial genus with representative strains present in the intestinal tract of human and other animals. Despite unique metabolic features including the production of butyrate from both sugars and amino acids, there is to date no data on their diversity, ecology, and physiology. Using a comprehensive phylogenetic approach, *Intestinimomas* was found to include at least 3 species that colonize primarily the human and mouse intestine. We focused on the most common and cultivable species of the genus, *Intestinimonas butyriciproducens,* and performed detailed genomic and physiological comparison of strains SRB521^T and AF211, isolated from the mouse and human gut, respectively. The complete 3.3-Mb genomic sequences of both strains were highly similar with 98.8 % average nucleotide identity, testifying to their assignment to one single species. However, thorough analysis revealed significant genomic rearrangements, variations in phage-derived sequences, and the presence of new CRISPR sequences in both strains. Moreover, strain AF211 appeared to be more efficient than strain SRB521^T in the conversion of the sugars arabinose and galactose. In conclusion, this study provides genomic and physiological insight into *Intestinimonas butyriciproducens,* a prevalent butyrate-producing species, differentiating strains that originate from the mouse and human gut.

Keywords: *Intestinimonas butyriciproducens*, Butyrogenesis, Genomics

Introduction

The gut microbiome is a living organ within the human body that contains highly complex and diverse microbial communities contributing substantially to human health (Zoetendal*, et al.* 2006). While the majority of these communities consists of strictly anaerobic bacteria that are difficult to culture (Flint*, et al.* 2006), an increasing number has successfully been isolated, and presently approximately 1000 cultured species, including eukaryotes and archaea have been reported (Rajilić-Stojanović and de Vos 2014, Browne*, et al.* 2016). In spite of this, only a small number of these intestinal anaerobes have been characterized at the genomic level and often only draft genomes have been assembled. Complete genomes have the advantage that full genetic repertoires can be studied, and repetitive sequences can be revealed.

The chemical compound butyrate and butyrate-producing bacteria have gained increasing attention as they both play an important role in maintaining gut homeostasis and intestinal epithelial integrity (Hamer*, et al.* 2008). Butyrate is a major substrate for colonocytes and is reported to have anti-carcinogenic, anti-inflammatory, and anti-oxidative effects (Hinnebusch*, et al.* 2002, Klampfer*, et al.* 2003). Interestingly, decreased amounts of butyrogenic bacteria have been observed in inflammatory bowel diseases, colon cancer, and type-2 diabetes (de Vos and Nieuwdorp 2013, Qin*, et al.* 2012, Flint*, et al.* 2012b). Till date several butyrate-producing bacteria have been isolated that mainly belong to *Clostridium* clusters XIVa and IV (gathering many members of the family *Lachnospiraceae* and *Ruminococcaceae*, respectively). These bacteria mainly produce butyrate via the acetyl-CoA pathway (Louis*, et al.* 2010). A recent study based on comparative analysis of metagenomes reported three other pathways leading to butyrate production where amino acids serve as substrates, including lysine and glutamate (Vital*, et al.* 2014b). All these butyrogenic pathways merged at an energy-generating step where crotonyl-CoA is converted to butyryl-CoA by means of a butyryl-CoA dehydrogenase complex (Bcd/etf). The coupled oxidation of reduced ferredoxin and reduction of NADH from this conversion by the membrane-integrated Rnf complex creates a proton motive force (Li*, et al.* 2008). The lysine pathway that includes a specific acetoacetyl-CoA transferase involved in the butyrate

formation step was estimated to be the second dominant butyrogenic pathway in the gut. However, no cultured isolates capable of performing this relevant reaction had been described (Vital*, et al.* 2014b). Using a classical cultivation approach with human stool as inoculant, we isolated strain *Intestinimonas* AF211, which is capable of converting lysine completely and efficiently into butyrate and acetate (Bui*, et al.* 2015). This feature is unique for the gut ecosystem and links two important metabolic features, butyrogenesis and amino acid fermentation in the intestinal tract. The complete 16S rRNA gene sequence of strain AF211 revealed more than 99 % similarity to *Intestinimonas butyriciproducens* SRB521[T] isolated from mouse caecum (Kläring*, et al.* 2013), indicating that these two strains belong to the same species within *Clostridium* cluster IV. Of note, *Faecalibacterium prausnitzii*, another butyrogenic species within *Clostridium* cluster IV, is known to be prevalent and abundant in human (Arumugam*, et al.* 2011) but not in mice (Xiao*, et al.* 2015).

Predictions based on complete genome sequences combined with physiological and other functional studies are powerful tools for improving our understanding of the lifestyle and role of bacteria. A genomic approach can also be used to gain insight into the genetic potential of an organism and assists in predicting its ecological role. Presently, only a few closed genomes of intestinal butyrate-producing bacteria have been characterised, including that of *Faecalibacterium prausnitzii* (Heinken*, et al.* 2014b) and *Roseburia inulinivorans* (Scott*, et al.* 2006). More recently, a partial genome sequence of *Anaerostipes hadrus* PEL 85 was reported (Kant*, et al.* 2015) and a number of draft genomes of butyrate-producing bacteria from the *Roseburia/E.rectale* group have been characterised (Sheridan*, et al.* 2016). In the present study, we focused on the recently described genus *Intestinimonas* and describe its phylogenetic and host distribution. Moreover, we compare the complete genome sequences of the human isolate *I. butyriciproducens* AF211 and the mouse isolate *I. butyriciproducens* SRB521[T]. Finally, we describe the physiological characteristics of both strains to confirm their predicted metabolic features and provide explanations of their adaptation to different hosts.

Materials and Methods

Strains and Growth Condition

I. butyriciproducens strain AF211 was obtained from a stool sample of a healthy individual (Bui et al., 2015) and *I. butyriciproducens* strain SRB521[T] was isolated from the caecum of a mouse (Kläring*, et al.* 2013). Bacteria were routinely grown in anaerobic Reinforced Clostridium Medium (RCM, Difco) in 120-ml serum bottles sealed with butyl-rubber stoppers at 37 °C under a gas phase of 1.7 atm of N_2/CO_2 (80 : 20, v/v).

Metadata Search Of *Intestinimonas* spp. Using 16S Rrna Gene Amplicons

To explore the ecological distribution and sequence supported diversity of the genus *Intestinimonas* we followed a modified version of a method previously described (Lagkouvardos*, et al.* 2014, Kläring*, et al.* 2015) and now available through the IMNGS web platform at www.imngs.org. In short, the 16S rRNA gene of the type strain of *I. butyriciproducens* strain SRB521[T] (NR_118554) was used for a similarity query (95 %) against all 16S rRNA sequences available from the SRA in IMNGS (build 1508). The obtained sequences were filtered, aligned and used for calculation of species-like clusters (approx. 3% similarity) using CROP (Hao*, et al.* 2011). The detailed method is available as supplementary materials.

Genome Sequencing and Assembly

Cells of strain AF211 and SRB521[T] grown in RCM (overnight cultures) were used for genomic DNA extraction performed using ZR Fungal/Bacteria DNA MiniPrep kit (ZYMO, USA) according to manufacturer's instructions. Sequencing effort included the use of two complementary platforms based on the Illumina and PacBio technologies.

Genome sequencing of two 15kb libraries using with PacBio RS II instrument using P4/C2 chemistry (Pacific Biosciences, Menlo Park CA, USA) at the DNA sequencing and genomics laboratory, Institute of Biotechnology, University of Helsinki, Finland. Data processing and filtering was done with PacBio SMRT analysis pipeline v2.2 and the Hierarchical Genome Assembly Process (HGAP) protocol (http://www.pacb.com/devnet/). Filtering was run with minimum subread length 500 and polymerase read length quality 0.80. For pre-assembly, settings were: minimum

seed read length 7000, split target into chunks 1 and alignment candidate per chunk 24. Assembly was performed by the Celera V1 assembler with parameters genome size 3,000,000, target coverage 30, overlapper error rate 0.06, overlapper mini length 40 and overlapper K-mer 14. Assembly polishing was done with Quiver.

In addition, genome sequences were also collected using a HiSeq2000. This resulted in 3,202,992 and 2,759,530 paired reads for AF211 and SRB521[T], respectively, which were used to correct for possible PacBio sequencing errors. For the HiSeq paired-end reads, all rRNA reads were removed with SortMeRNA v1.9 using default settings except for an increased error value of 20% for the adaptors, and using the reverse complement of the adaptors as well (Kopylova*, et al.* 2012). Quality trimming was performed with PRINSEQ Lite v0.20.0 with a minimum sequence length of 40 and a minimum quality of 30 on both ends and as mean quality of read (Schmieder and Edwards 2011). For assembly, Ray v2.3 was used (Boisvert*, et al.* 2012a). Duplicate contigs in both the genome assemblies of AF211 and SRB521[T] were discarded if they had a hit with at least 99% sequence identity within a bigger contig, which spanned at least 98% of the contig query length. Furthermore, contigs with length <500bp were discarded. Next, Illumina contigs were aligned to the PacBio assembly using promer algorithm from the MUMmer 3.0 software package (Kurtz*, et al.* 2004) with the parameters show-tiling -c -R -i 50. The alignment was processed in R (http://www.r-project.org/) using in-house scripts to extract the Illumina sequence and combined with PacBio assembly at gap positions to produce the final closed genome sequence. A circular element was detected within this assembly, based on the BLASTP results of the predicted proteins (e-value 0.0001), and this added to the final assembly result. Scaffolding of the contigs was done with SSPACE-LongRead and the PacBio CCS reads using default settings (Boetzer and Pirovano 2014). The annotation was done by RAST server (Aziz*, et al.* 2008). Functional prediction of proteins was verified manually by BLASTing the amino acid sequences in Pfam (Finn*, et al.* 2014), Brenda, Interpro (Hunter*, et al.* 2012) and Uniprot databases. For a global overview of the genome features, the circular genome map was created using the virtual machine CGView Comparison Tool (CCT) (Grant*, et al.* 2012).

Specific Annotations

The core genome of *I. butyriciproducens* was calculated using Spine and the pan-genome was identified using AGEnt (Ozer*, et al.* 2014). For genomes of two strains, specific annotations for CRISPR genes was done using CRISPRfinder (Grissa*, et al.* 2007). Target phages for CRISPR were identifies using CRISPRtarget using Genbank Phage, and ACLAME genes databases with an E-value cut-off of 0.1 (Biswas*, et al.* 2013). The CRISPRTarget tool identifies target phages using different databases and the search algorithm has its own scoring system with a score of 20 which was kept as default minimum as suggested by the authors (Biswas*, et al.* 2013). All the spacers targeting phages in both SRB521^T and AF211 had an acceptable score of 21. Prophage genes in the genomes were annotated using PHAST (PHAge Search Tool) (Zhou*, et al.* 2011). The PHAST uses a scoring system based on three methods (i) Presence of known phage genes/proteins; (ii) >50% genes/proteins in the region are related to a known phage; and (iii) <50% genes/proteins in the region are related to a known phage (Zhou*, et al.* 2011). In general, if a given phage region's total score is < 70, it is marked as incomplete; between 70 to 90, it is marked as questionable; if >90, it is marked as intact. The carbohydrate active enzymes (CAZymes) were annotated using the amino acid sequences with default search parameters (Park*, et al.* 2010). Genomic islands were identified using online IslandViewer and a standalone program *i.e.* SeqWord Gene Island Sniffer that identifies genomic islands based on the analysis of oligonucleotide usage variations in DNA sequences (Langille and Brinkman 2009, Dhillon*, et al.* 2013, Bezuidt*, et al.* 2009).

Phenotypic Characterization

To investigate carbohydrate assimilation and detection of enzyme activities, API 20NE, API rapid 32, API ZYM were used with overnight-grown cells of both strains. The Gram reaction was determined using standard methods (Plugge*, et al.* 2000). In-vivo DNA-DNA hybridization was done at the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ, Braunschweig, Germany), as described previously (Ley*, et al.* 1970) with some modifications (Huss*, et al.* 1983).

For cellular fatty acid analysis, cells of both strains were grown in RCM medium for 5 days before being centrifuged (10,000 rpm) for 10 min at 4 °C. Pellets were stored at -

80 °C and later used to extract cellular fatty acids as described previously (Kämpfer and Kroppenstedt 1996). Products were analysed by Agilent model 6890N gas 199 chromatography (MIDI Sherlock, Newark, N.J.) as described previously (Kämpfer and Kroppenstedt 1996, Miller 1982).

Bile tolerance was tested in RCM supplemented with OX-Bile (0.1; 0.2; 0.3; 0.4; 0.5% (w/v); SIGMA ALDRICH). Growth was determined via OD measurement and product formation using HPLC as described previously (van Gelder*, et al.* 2012).

Growth On Carbohydrates And Amino Acids

Intestinimonas butyriciproducens strains were grown in 50 ml anaerobic bicarbonate buffered medium (Stams*, et al.* 1993) with 2 g/l yeast extract. Carbohydrates were added from 1M sterile stock solutions (10 mM final concentration). Tested carbohydrates were glucose, fructose, maltose, mannitol, cellobiose, lactose, raffinose, xylose, D-mannose, saccharose, galactose, D,L- arabinose and sorbitol. The polysaccharides starch, inulin, pectin, and mucin were added before autoclaving and used at a final concentration of 0.5% (w/v). Valine, serine, asparagine, threonine, isoleucine, aspartate, glutamate, methionine, citrulline, glycine were added to make 10mM at final concentration from 1M stock filter-sterile solutions. The growth in the mixture of amino acid was performed in the bicarbonate buffered medium containing 0.1 g/l yeast extract. Samples were collected at different time points and fermentation profiles were analysed using GC and HPLC analysis as described previously (van Gelder*, et al.* 2012). A Thermo Scientific Spectra HPLC system was used to detect organic acids including lactate, formate, acetate, propionate, isobutyrate, butyrate and alcohols including methanol, ethanol, 1,2 propanediol, 2-propanol and butanol. The HPLC system was equipped with a Agilent Metacarb 67H 300 x 6.5 mm column kept at 45 °C and running with 5mM H_2SO_4 as eluent and a refractive index detector.

Antibiotic Susceptibility Tests

The E-test was done to identify minimal inhibitory concentrations (MICs) according to the manufacturer's protocol (bioMérieux, France). Both strains were pre-grown in RCM broth (overnight cultures) and 50µl was spread on RCM agar plates (1.5 % w/v agar) until the agar surface was dry and the liquid was absorbed by the agar. Two E-test strips were placed on one plate to test each antibiotic. Results were noted according to

manufacturer's protocol (bioMérieux, France). Antibiotics tested included ciprofloxacin, cefotaxime, erythromycin, oxacillin, teicoplanin, tetracycline, tobramycin, vancomycin and sulfamethoxazole. The concentration range was 0.016 – 256 µg/ml for chloramphenicol, oxacillin, tetracycline, tobramycin and vancomycin, and 0.016 – 32 µg/ml for ciprofloxacin, cefotaxime, erythromycin, teicoplanin and sulfamethoxazole. MIC values were recorded directly from the strips after 48 hours and rechecked after 4 days. Duplicate experiments were performed for each antibiotic.

Vitamin B12 Analysis

Cell pellet (0.5 g) was suspended in 10 ml of extraction buffer (pH 4.5; 8.3 mM sodium hydroxide and 20.7 mM acetic acid) containing one mg of sodium cyanide. Following heat extraction (100 °C; 30 min) and cooling on ice, the extract was recovered by centrifugation (6,900 *g*; 10 min). Five ml of the extract was purified using an immunoaffinity column (Easy-Extract; R-Biopharma, Glasgow, UK) according to the manufacturer's instructions, and finally recovered in 500 µl of water. To identify the vitamin form, the extract was analysed using UHPLC (Waters Inc.; USA) and quadrupole time-of-flight mass spectrometer (Q-TOF; Synapt, G2-Si, Waters) as described previously (Deptula*, et al.* 2015).

Results and Discussion

Diversity and Ecology Of *Intestinimonas*

Intestinimonas is a newly described genus with peculiar metabolic functions, but its phylogeny and ecological distribution remain unknown. By utilizing the accumulated sequences available in the Sequence Read Archive (SRA) (https://www.imngs.org/), 9,887 *Intestinimonas*-like 16S rRNA gene amplicons were obtained and their pairwise distances projected in two dimensions using MDS (Figure 1A). Using full-length substitutes for the 4 bins containing the most amplicon sequences, enabled further filtering of the sequences that could be confidently classified as members of the genus *Intestinimonas* to 5,732 reads (see Supplementary methods). Clustering those using CROP resulted in the formation of 8 none-singleton clusters. The vast majority of OTU sequences (99 %) fell into only 3 major molecular "species" clusters. The results mirrored the diversity supported by phylogenetic analysis of near full length 16S rRNA

genes (Figure 1B). Ecological assessment revealed that 90 % of *I. butyriciproducens*-like

sequences were predicted to originate from human samples (Figure 1B).

In contrast, the species represented by clone FJ676152 clustered with reads that

originate from various animal samples. Few amplicons, primarily originating from

murine gut, mapped to the species represented by clone FJ374227. *I. butyriciproducens*

strains have also been isolated from swine faeces, signifying its widespread presence

in mammals (Levine*, et al.* 2013, Rettedal*, et al.* 2014). As *I. butyriciproducens*-like

sequences represented a major clade within the genus and were most prevalent in

humans, but also present in the murine host, we performed in-depth genomic and

physiological comparison of strains AF211 and SRB521^T, isolated from the human and

mouse gut, respectively.
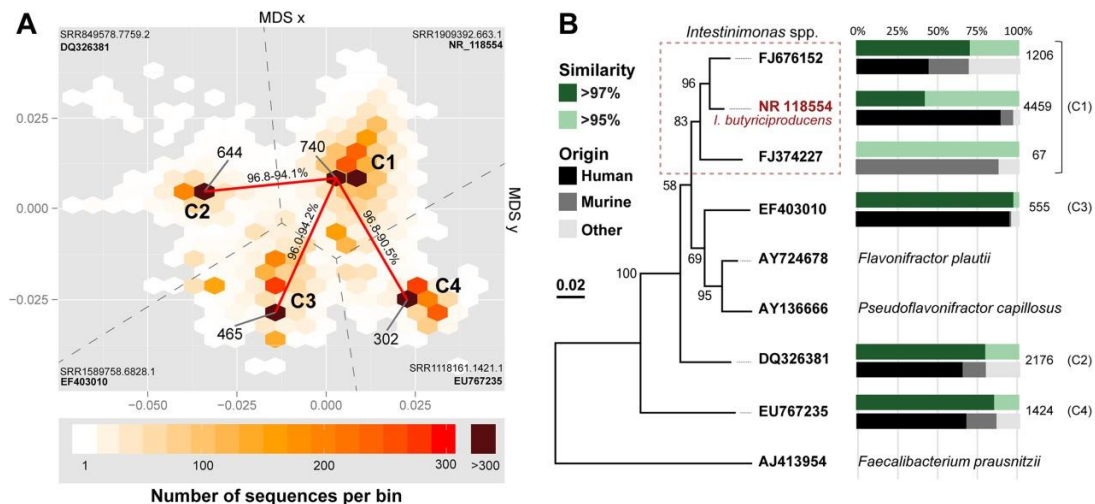
Figure 1: 16S rRNA-based diversity and host distribution of *Intestinimonas*. **A]** MDS plot of pairwise distances between *Intestinimonas*-like sequences extracted from SRA. Sequences were binned in hexagons color-coded according to the number of sequences they contained. The reads formed separated clusters (C1-C4) around bins containing over 300 sequences each. Similarity distances between representative short amplicon reads and clones from the three neighbouring cores and the one containing the reference 16S rRNA sequence of *I. butyriciproducens* (cluster C1) are shown next to the red lines connecting them, respectively; the corresponding sequence accession for each cluster core are given in the corresponding corner. The grey dash lines indicate midpoints between cores after assignment of each point (OTU) to its closest center. **B]** Maximum likelihood tree of *Intestinimonas*-related 16S rRNA gene sequences and representatives of most abundant amplicon reads (Fig 1A) together with selected neighbouring genera and outgroup. Bootstrap support is shown next to branches, with branch lengths measured in the number of substitutions per site. The box with red dashed-line indicates members of the genus *Intestinimonas*. The number of unique amplicon sequences mapped to each clone is shown on the right-hand side.

Phenotypic Characterization of *I. Butyriciproducens* Strains

Physiological and biochemical characterization of *I. butyriciproducens* strains AF211 and SRB521^T using API Rapid 32A, API 20NE, and API ZYM kits revealed that strain SRB521^T but not AF211 possessed N-acetyl-β-glucosaminidase activity (Table S1). Interestingly, analysis of cellular fatty acid composition, an important chemotaxonomic characteristic for microbial classification, revealed differences between the two strains (Table S2). More than 20 % content difference was observed in major fatty acids, such as $C_{14:00}$; iso-$C_{19:1}$ I and $C_{18:00}$. The cellular fatty acids have a major role in maintaining the integrity and viability of microbial cells, and may change in composition dependent on the environmental and growth conditions. Recently, differences in intra-species cellular fatty acid composition in *Lactobacillus plantarum* strains were found to be associated with the origin of isolation (Garmasheva*, et al.* 2015). Strains AF211 and SRB521^T were isolated from different hosts; the observed variations in cellular fatty acids may reflect the different environmental conditions in their natural habitats.

In the gut environment, bacteria need to resist several stress conditions. One of these is bile, which can act as a surfactant and has the ability to affect the phospholipids and proteins of cell membranes (Begley*, et al.* 2006). Bacteria are known to tolerate bile via two mechanisms: one is by using bile salt hydrolases (BSH) to hydrolyse bile and the other is by using efflux pumps to remove bile that passes through the outer membrane (Begley*, et al.* 2006, Jones*, et al.* 2008). The genomes of both AF211 and SRB521^T lacked the coding sequence for canonical BSHs but indicated the presence of genes for several MDR efflux pumps that could aid in resistance to bile. *In vitro* investigation of bile tolerance showed growth of strains AF211 and SRB521^T after 3 days of incubation in 0.1% bile (w/v) (Figure S1) but no growth was observed in the presence of higher concentrations. This may be due to the absence of a bile salt hydrolase gene in both strains. While strain AF211 had a shorter lag time and grew to a higher OD than strain SRB521^T without bile the differences between doubling times with and without bile were most pronounced (an increase of 17 and 20.4 hours for strain AF211 and SRB521^T, respectively) (Figure S1). The reduced bile inhibition of strain AF211 suggests that it to be better adapted to bile than the mouse strain although it may also reflect different

adaptation to bile components since these are known to differ between mouse and man (Sayin*, et al.* 2013).

Comparative Genomics - General Features of The *I. Butyriciproducens* Genomes

I. butyriciproducens strains AF211 and SRB521[T] both have a single circular chromosome with similar genome features (Figure 2A). No plasmid sequences were detected in either of the genomes. The genome size of AF211 is 112,795 bp larger than that of SRB521[T], indicating larger genetic potential in the human isolate that is also reflected in a higher number of coding sequences *i.e.* 3,363 vs. 3,268. Furthermore, the average nucleotide identity (ANI) between the two isolates confirmed that they belong to the same species.

A

	AF211	SRB521[T]
Host	Human	Mouse
Accession number	CP011307	CP011524
Genome Size (Base pairs)	3,376,476	3,263,681
GC content (%)	59.1	59.4
RNA genes	57	58
Number of Coding Sequences	3363	3268
Number of Subsystems	348	338
Average nucleotide Identity (ANI)	98.80 %	98.80 %

B
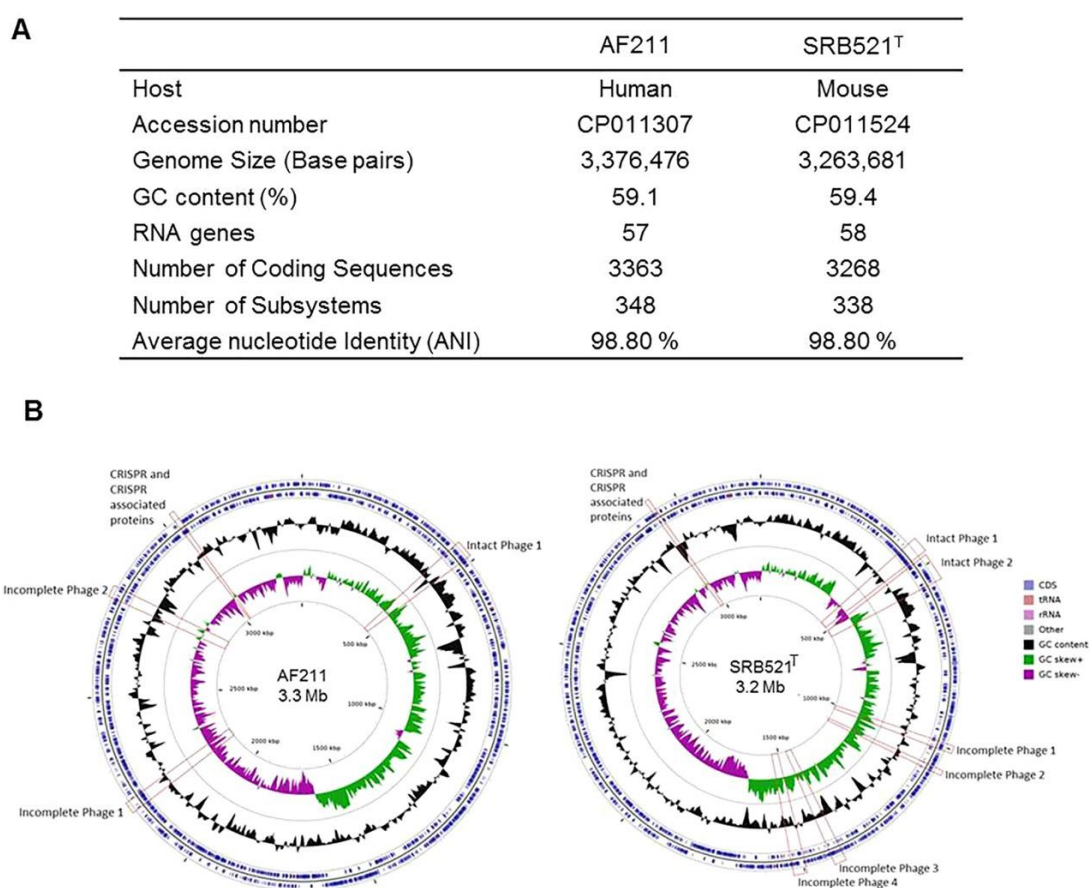
Figure 2: General genome features. **A]** and genome maps **B]** of *Intestinimonas butyriciproducens* isolates. From the outer circle inward, coding regions are marked on the first two rings: genes encoded on the positive strand (outside) and if encoded on the negative strand (inside). The third ring (black) shows local GC content. The innermost graph shows the CG skew. Phage regions and CRISPR associated genes are marked within red lines.

Additional genomic features such as the distribution of GC content, GC-skew, and location of genes including tRNA, rRNA, prophage genes are indicated in Figure 2B.

Pan- And Core Genome

The pan-genome of *I. butyriciproducens* was calculated to contain 4,762 coding DNA sequences (CDS), of which 2,612 CDS (2,603,042bp) were considered as core. The average number of CDS predicted per genome was approximately 3,427, indicating that an average of 76.21 % of the genome of each isolate was part of the core. The overall subsystem level core and accessory genome features were further analysed. This revealed that approximately 22.8 % of the genome sequences were unique to one of the isolates signifying the strain level differences in the genomic content of AF211 and SRB521[T]. The additional coding capacity in the AF211 was mostly assigned to genes coding for hypothetical proteins, mobile element proteins, and restriction enzymes that were located on genomic islands (Figure S2).

Flexible Genomic Regions in *I. Butyriciproducens*

To identify whether genomic features of *I. butyriciproducens* AF211 and SRB521[T] could be linked to the different origin of the strains, a detailed analysis of their genomes was made. As already evident from differences in GC skew (Figure 2), the two genomes carry a variety of genomic rearrangements including large inversions and integration of foreign regions of DNA (Figure 3). Alignment of the genomes of AF211 and SRB521[T] revealed the presence of 16 locally collinear blocks (LCBs) with several regions of inversion and rearrangement (Figure 3A). Moreover, an obvious X-pattern was observed when the genomes of AF211 and SRB521[T] were compared in a dot blot, indicating a symmetric chromosomal inversion, involving approximately one third of the genome around the origin of replication (Figure 3B). Not surprisingly, genes for a site-specific recombinase (SRB521_00768) and a probable integrase/recombinase (SRB521_002435c) flanking the inverted region in genome of SRB521[T] were detected. One of the LCBs, which includes a site-specific recombinase gene (SRB521_00768), is located at 738168-739784 bp in the mouse isolate; the corresponding LCB in the human isolate with the same gene (AF_02777) is located at around 2770592-2772208 bp, signifying the role of this gene in genomic rearrangement. In addition, we searched the

genomes for the presence of genomic islands to identify non-self-mobilising elements that code for proteins with diverse functions. The AF211 genome contained 24 genomic islands (GIs) while SRB521[T] had 18 (Figure S2A). These include the already noted (incomplete) prophages (see Figure 2 and Figure. S3) and series of other predicted gene clusters that are part of the GIs (Supplementary data 1). Remarkably, the majority of the genes located on the GIs encode hypothetical proteins in both AF211 (56%) and SRB521[T] (63%).
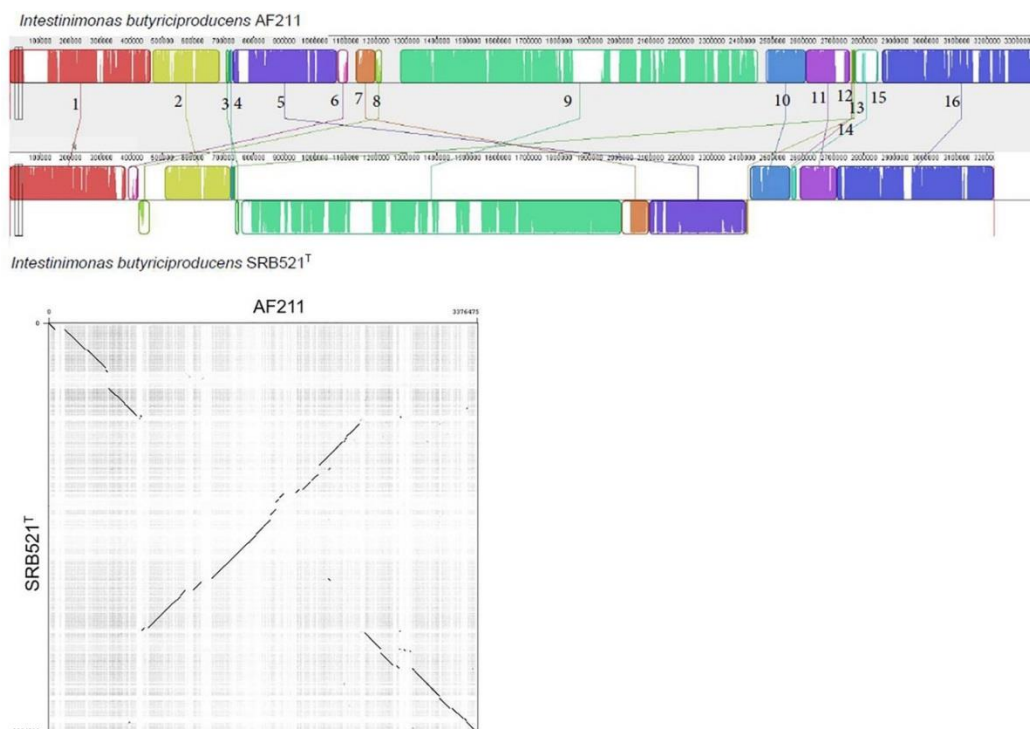
Figure 3: Alignment of the genome sequences of *I. butyriciproducens* strain AF211 and SRB521[T]. **A]** The genome of strain AF211 was used as reference for global alignment using progressive MAUVE. The numbers indicate the locally collinear blocks (LCBs) that were identified in both the genomes. Conserved and highly related regions are coloured, and low-identity unique regions are in white (colourless). LCBs below the mid-line in strain SRB521[T] are inverted relative to the strain AF211 sequence. **B]** Dot plot alignments of AF211 and SRB521[T] using Gepard. SRB521[T] is on the y-axis and AF211 is on the x-axis and the genomes are starting with the origin of replication at position 1. The black lines represent regions of similarity while the breaks in this line (syntenic line) represent regions of genomic variations at a given locus between the two genomes.

Prophages and CRISPR Systems In *I. Butyriciproducens*

There has been considerable attention for the presence and function of bacteriophages in the human intestine. Specifically, the intestinal mucosa has been identified as a niche for bacteriophages, potentially indicating their role in maintaining microbiota dynamics

(Barr*, et al.* 2013, Wootton 2013). As bacteriophage diversity and composition may vary between mouse and man, it was of interest to analyse the natural defence systems of the two *I. butyriciproducens* strains, including the analysis of prophages known to provide superinfection immunity. Comparative analysis of the genomes of AF211 and SRB521^T demonstrated major differences in the prophage elements. The genome of AF211 revealed one intact and 2 incomplete prophages while that of SRB521^T was predicted to have two intact and 4 incomplete prophages (see Figure. S3). Prophage integration sites in the genomes of both strains were found to have significantly different G + C content (Figure 2A). In AF211, the prophages had varying G+C content, intact prophage 1 and incomplete prophage 1 had a higher G+C content of 61.77% and 62.42% respectively vs. 54.95% for incomplete prophage 2 when compared to the overall G+C content of 59.4% for the entire genome of AF211. In case of SRB521^T, intact prophage 1 had a G+C content of 59.29%, while that of intact prophage 2 was 64.15%. Additionally, in the SRB521^T genome, two of the incomplete prophages had higher G+C content (63.84% and 64.10%), while for the other two it was lower, 55.66% and 54.90% respectively. These observed differences suggest that the genes were acquired from both high and low G+C content bacterial taxa. The comparison of prophage sequences between the two genomes revealed that the two strains do not share any prophage sequences, indicating exposure to different prophages in human and mouse gut. This observation was also supported by the CRISPR/Cas system search in the two genomes. Both strains possessed clustered regularly interspaced short palindromic repeats (CRISPR/Cas system) in their genomes. The CRISPR array involves associated proteins to confer resistance to bacteriophages. The CRIPSR arrays contained the same 33 bp long direct repeat (5- ATTTCAATCCACGCCCCCCGTGTGGGGGGCGAC- 3). The number of spacers was 58 and 21 for strain AF211 and SRB521^T, respectively. Interestingly, only two spacers were shared between the two strains (Figure S2B). Blast analysis of the spacer regions revealed no sequence with complete identity to phage database entries, which can be attributed to the lack of phage sequences in Genbank Phage, and ACLAME genes databases. However, spacer 11 of the CRISPR system in strain SRB521^T targeted the phage minor tail protein in prophage reported in *Nocardia farcinica* IFM 10152. Interestingly, spacer 46 of the CRISPR system in strain AF211 appeared to target phages

that made up 18% of total virus like particle (VLP) contig sequences in metagenomic DNA from human fecal samples (Minot*, et al.* 2011). Harbouring an immune system against this highly abundant viral family could be beneficial for survival of AF211 in the human gut.

Butyrogenesis From Sugars and Lysine

As the intestine of men and mice differ considerably in the composition, load, and kinetics of dietary components, it was of importance to address the metabolic capacity of the two *I. butyriciproducens* strains. Genomic analysis revealed the presence of a complete glycolytic pathway while the pentose phosphate and the tricarboxylic acid pathways were incomplete in both strains. *In vitro* tests confirmed a mixed acid fermentation using variety of carbohydrate substrates, which resulted in production of butyrate as a major product and ethanol, isobutyrate or lactate as minor metabolites. The two strains grew poorly in hexose sugars but growth was much improved by the addition of acetate (Figure 4A). A similar feature was previously described in *Faecalibacterium prausnitzii;* the presence of acetate increased energy harvest from glucose by this species (Heinken*, et al.* 2014b) via stimulation of the acetyl-CoA pathway.

This pathway has a butyryl-CoA dehydrogenase (Bcd) electron-transferring flavoprotein (Etf) complex that generates a proton motive force via a membrane-integrated Rnf complex (Li*, et al.* 2008). Comparison of the genes invovled in the acetyl-CoA pathway between the two genomes revealed high similarity in butyryl-CoA:acetate CoA transferases which are the key enzymes in butyrate formation (Figure 4B). Neither of two strains was capable of utilizing polysaccharides (inulin, starch, mucin and pectin) in vitro although the genomes contained a number of glycoside hydrolases. Strain AF211 was characterized by a higher number of genes related to auxiliary activites, carbohydrate esterases, glycoside hydrolases, and glycoside transferases (Figure S4). Genes coding for sugar phosphorylation such as hexokinase, galactosekinase, and xylulosekinase were detected, demonstrating potential to grow on hexose. Indeed, laboratory tests showed good growth on galactose, arabinose, glucose and some growth on fructose, maltose and lactose in presence of 10mM acetate and 2g/l yeast extract but mannitol, cellobiose, raffinose, xylose, D-mannose, saccharose or sorbitol

did not support growth (Figure 4A). Butyrate production was higher in cultures of strain AF211 compared to SRB521^T when grown on galactose and arabinose, suggesting the higher effeciency in sugar conversions of the human isolate over the mouse isolate.
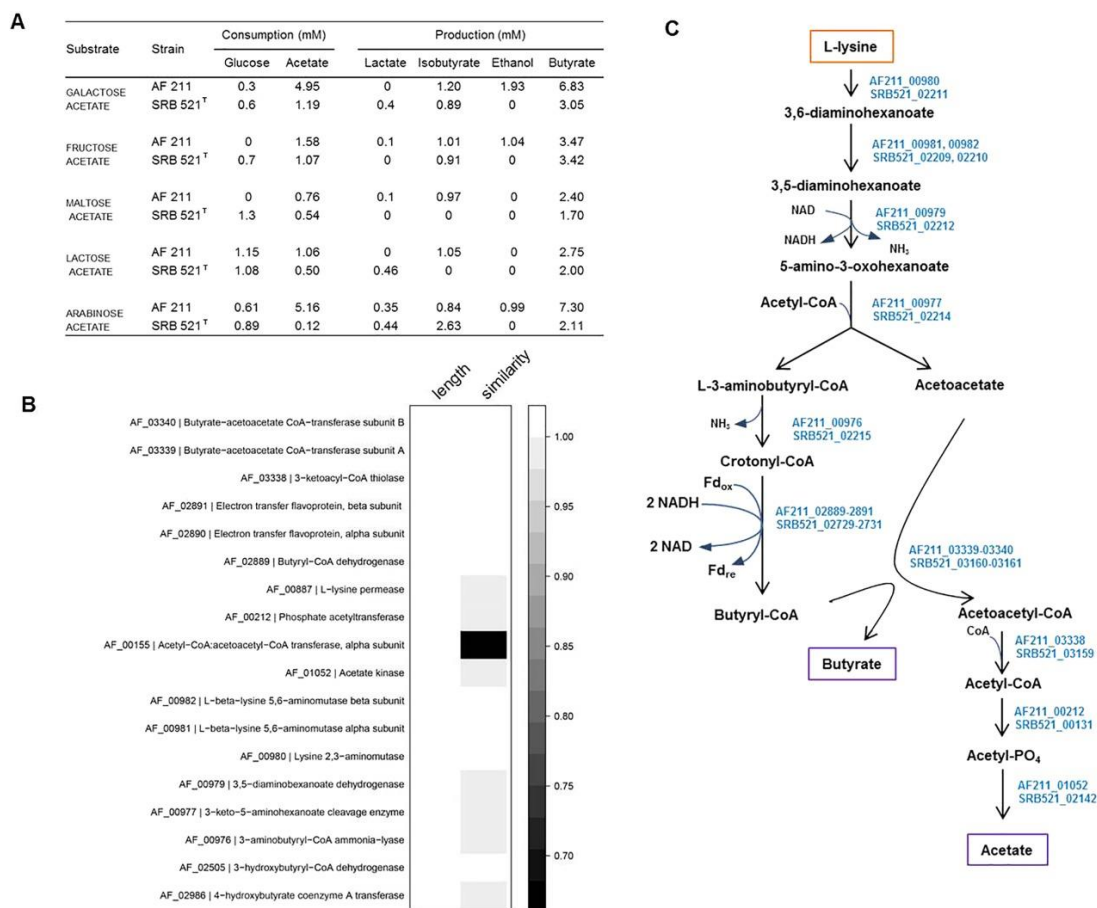
Figure 4: Butyrogenesis in strain AF211 and SRB521^T. **A]** Fermentation profiles of AF211 and SRB521^T in 10 mM sugars and 10 mM acetate for 2 weeks incubation. Product formation is calculated in mM. The highlighted values indicated discrepancy in butyrate formation and acetate consumption. Values were mean of duplicates and standard deviation was below 10%. **B]** Genes involved in butyrate formation in strain AF211 and strain SRB521^T. The dark colour indicates low similarity of sequences of strain SRB521^T as compared to that of strain AF211. **C]** Lysine conversion pathway in both strains: locus tags are indicated in blue.

The amino acid lysine has been found as substrate for butyrogenesis in *I. butyriciproducens* AF211 (Bui*, et al.* 2015). Strain SRB521^T showed similar capability of producing butyrate and acetate from lysine (Figure S5) and the genome of this strain also coded for the entire lysine pathway (Figure 4C). Genes involved in the lysine pathway in strain SRB521^T and AF211 were compared (Figure 4B) and found to have the same length and high nucleotide identities (above 95%), except for the acetyl-CoA: acetoacetate CoA transferase gene (about 70 %) (Figure 4B-C). Thus, the lysine

pathways could be a common feature of *I. butyriciproducens* species, irrespective of the host environment. This unique capacity of *I. butyriciproducens* to grow on lysine positions this strain at the cross-road between protein metabolism and butyrogenesis in the gut ecosystem. Of note, a small amount of iso-butyrate was formed during sugar fermentation, which was associated with the supplementation of yeast extract. We considered the possibility of the Stickland reaction (Stickland 1934). However, we found that growth on yeast extract alone did not give any iso-butyrate production and neither of the two strains grew in a mixture of amino acids (valine, serine, asparagine, threonine, isoleucine, aspartate, glutamate, methionine, citrulline, glycine). Iso-butyrate can also be formed via the branched chain amino acid pathway, as has been observed in a marine anaerobe although in this case no energy is produced and this pathway is assumed to prolong starvation survival (Harwood and Canale-Parola 1981). A similar pathway may explain the observed iso-butyrate production in *I. butyriciproducens,* since both strains contain the genes for this pathway, including a branched chain amino acid aminotransferase, a phosphotransbutyrylase and several copies of oxidoreductases and kinases.

Antibiotic Resistance of *I. Butyriciproducens*

The ability of antibiotics to modulate the gut ecosystem is well known. Strain AF211 and SRB521[T] showed a similar sensitivity profile to the tested antibiotics (Figure 5A). Several resistance mechanisms appear to be encoded by the genomes of the two strains (Figure 5B). Multidrug resistance efflux pumps were abundantly present with >20 copies in both strains. These are non-specific transport systems used to export (or import) toxic compounds in many bacteria and involved in housekeeping functions (Poelarends*, et al.* 2000). Antibiotic resistance analysis using the E-test showed highest minimum inhibitory concentration (MIC) on tetracycline (>256 µg/ml) for both strains and lowest MIC on teicoplanin (0.032 µg/ml for AF211 and 0.047 µg/ml for SRB521[T]). Results for tetracycline in the present study were higher from that reported earlier for strain SRB521[T] (Kläring*, et al.* 2013) as the E-test was done in different media and incubation condition that might result in different MIC values (Huys*, et al.* 2002). Tetracycline is a protein synthesis inhibitor that binds to the 30S subunit of ribosomes. The observed tetracycline resistance may be due to the potential production of *a TetW*

protein   that   shares   99% identity   with   TetW-like proteins in *Roseburia, Bifidobacterium* and some other intestinal strains. In addition, the presence of multiple efflux pumps may also contribute to tetracycline resistance (Fig 5), as demonstrated in *Pseudomonas aeruginosa* (Li*, et al.* 1994) and *E. coli* (Okusu*, et al.* 1996) and subject to activation by a transcriptional regulator TetR, which was also found to be predicted from the genomes of both *I.butyriciproducens* strains.

A

Antibiotics	*I. butyriciproducens* SRB521[T] (ug/ml)	*I. butyriciproducens* AF211 (ug/ml)
Ciprofloxacin	>32	>32
Cefotaxime	0.064	0.064-0.05
Erythromycin	1	0.75-1
Oxacillin	0.38-0.5	0.38
Teicoplanin	0.032	0.047
Tetracycline	>256	>256
Tobramycin	1.5-2	2
Vancomycin	0.75	0.75
Chloramphenicol	0.75	0.75
Sulfamethoxazole	>32	> 32
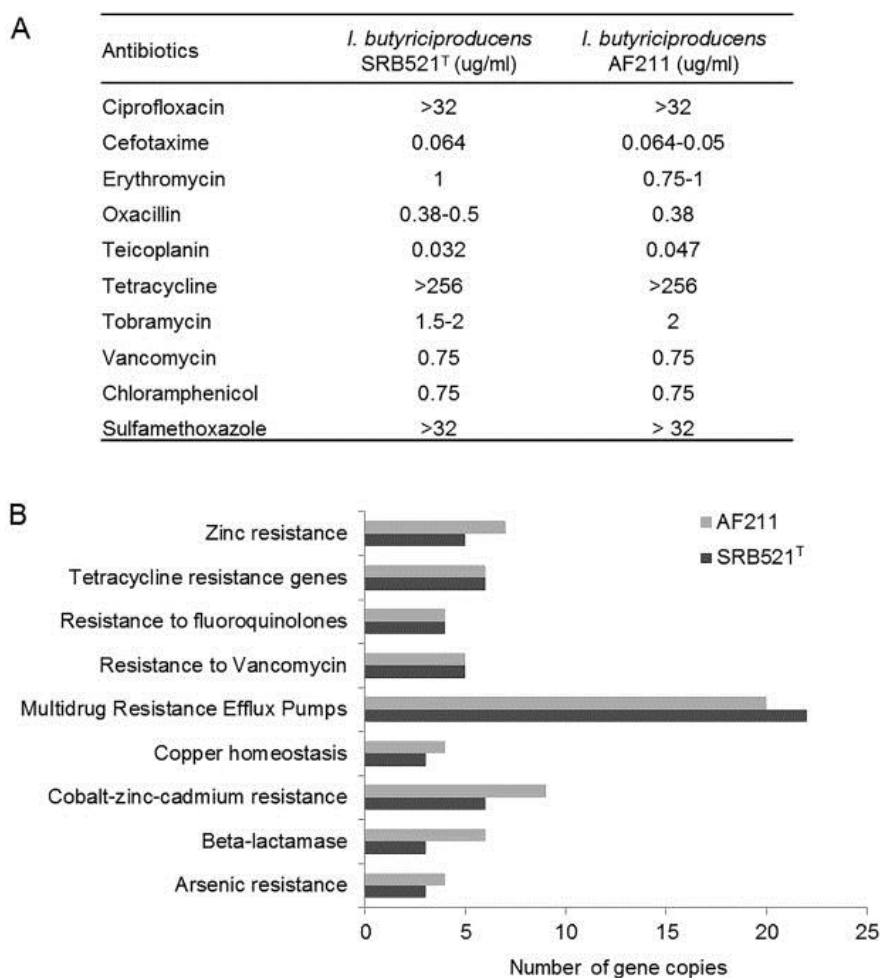
Figure 5: Resistance genes in strain AF211 and SRB521[T]. **A]** Numbers of resistance genes in the genomes of *I. butyriciproducens* AF211 and SRB521[T]. **B]** Minimum inhibitory concentration of the 2 strains (μg/ml).

The two strains were sensitive to teicoplanin, vancomycin, cefotaxime, and oxacillin, which are inhibitors of cell wall synthesis in Gram-positive bacteria, but they showed relatively high MIC values with ciprofloxacin and sulfamethoxazole (Figure 5B). Although (remnants of) some vancomycin resistance genes (*vanB, vanW, vanS* and

vanR) were detected in the two genomes, these were not sufficient to provide vancomycin resistance, which requires a complex operon with at least 5 genes (Kruse*, et al.* 2014). A set of genes predicted to code for resistance to inorganic compounds including arsenic resistance protein, cobalt-zinc-cadmium resistance protein, and copper resistance protein, were also found as part of the defence system of both strains.

Oxidative Stress, Transport Systems, And Vitamin Production In *I. Butyriciproducens*

The human intestine is considered as an anaerobic ecosystem, especially the large intestine, even though there is a gradient of oxygen from the epithelial layer outward (Albenberg*, et al.* 2014). The reduction of oxygen results in production of superoxide radicals, hydrogen peroxide and hydroxyl radicals that are extremely toxic for cells (Fridovich 1995). Hence, it is important for intestinal microbes that possess defence systems to prevent accumulation of these reactive oxygen species. Moreover, the oxidation of substrates and various xenobiotics might generate toxic oxidizing compounds in the environment, which intestinal anaerobes need to deal with for survival (Brioukhanov and Netrusov 2004). Several genes predicted to have protective role under oxidative stress were detected in both strains (Table S3). Among them, CoA-disulfide reductase, rubrerythrin, superoxide reductase, and alkyl hydroperoxide reductase subunit C-like protein were detected in the proteome of strain AF211 (Bui et al 2015), indicative of active functions of these proteins. Superoxide reductase (EC 1.15.1.2) and manganese superoxide dismutase (EC 1.15.1.1) aid in lowering superoxide concentration and this process is carried out in the cytosol (Sheng*, et al.* 2014) while catalase (EC 1.11.1.6) prevents accumulation of hydrogen peroxide. These enzymatic activities have been tested in strict anaerobes including Clostridia*, Bacteroides* spp., acetogens, sulfate reducers, and methanogens, indicating a wide distribution among anaerobes (Brioukhanov and Netrusov 2004). It has been reported that cell starvation also results in the synthesis of antioxidative defence systems apart from O_2 (Rocha and Smith 1997). Genes encoding the protein rubrerythrin were most abundantly present in the genomes of either strain (6-8 copies). Rubrerythrin has a protective effect under oxidative stress, although its mechanism of action still remains unknown (Mukhopadhyay*, et al.* 2007). In addition, a gene for the peroxide stress

regulator PerR was found in both genomes. Interestingly, the alkyl hydroperoxide reductase subunit C (AhpC) was found to be encoded by strain AF211 but not SRB521^T. AhpC is involved in organic peroxide detoxification (Rocha and Smith 1999). The presence of different oxidative stress defence systems in the two anaerobic strains enhances their survival in the gut environment where there is always a gradient of oxygen toward lumen.

Membrane transporters that facilitate the exchange of ions, nutrients and metabolites with the environment are essential for bacterial metabolism. A total of 102 genes were annotated as membrane transporters in the genome of strain AF211, and 86 in strain SRB521^T for the same mechanisms of transport system (Figure S6). This suggests that strain AF211 has an improved overall capability of exchanging molecules with its environment. For both strains, ECF class transporters and ABC transporters were the most prevalent groups (10 to 30 copies) while symporter and antiporter were the least abundant (<2 copies) (Figure S6). The analysis showed that the genome of strain AF211 harboured more copies of genes for ECF class transporters (30 versus 24), ABC transporters (47 versus 43), Nickel/Cobalt transporters (9 versus 4) and TRAP transporters (8 versus 6) than strain SRB521^T, and this might result in better capability of transporting nutrients. The ECF transporters are present in various microbial lineages and responsible for transporting vitamins (Rodionov*, et al.* 2009). ABC transporters for oligopeptides, dipeptides and branched-chain amino acids were found, indicating the capability of taking up these compounds, which corresponds with our previous report of growth on different protein-derived sources (Bui*, et al.* 2015). Galactose and multiple sugar ABC transporters were predicted from the genomes of both strains. Although sugar phosphotransferase systems (PTS) are often present in Gram-positive bacteria, no PTS system was found in the genomes.

Vitamin B12 is an essential cofactor for bacterial and host metabolism and has been suggested to modulate the intestinal microbe-host symbiosis (Degnan*, et al.* 2014b). This vitamin is produced by a selected group of intestinal microbes and hence there is microbial competition for uptake of this metabolite (Allen and Stabler 2008). We showed that strain AF211 is capable of producing pseudo-vitamin B12 (Figure S7). Although we did not test for vitamin B12 synthesis in strain SRB521^T, genomic analysis

of strain SRB521^T showed an identical vitamin B12 synthesis pathway to that of strain AF211, indicating the capability of strain SRB521^T to produce a pseudo-vitamin B12 (data not shown). Some genes in the corresponding metabolic pathway appeared to be missing, but this can be attributed to the shortcomings of annotations currently available in present genome databases. It has been demonstrated in *Clostridium sticklandii* that lysine 5,6-aminomutase is a B12-dependent enzyme (Chang and Frey 2000). This enzyme also appeared as a key player in lysine fermentation by *I. butyriciproducens* (Bui*, et al.* 2015)*,* indicating that this strain produced vitamin B12 to support its metabolic conversion of lysine into butyrate. In addition, several other conversions also require vitamin B12 as cofactor, including the conversion of 1,2-propanediol to propionate, ethanolamine to acetaldehyde and ammonia, and glycerol to 3-hydroxypropionaldehyde (the antimicrobial reuterin) (Garsin 2010, Engels*, et al.* 2016a). This could also apply to *I. butyriciproducens* although initial experiments showed no growth in the bicarbonate buffered medium on glycerol, 1,2-propanediol or ethanolamine. We did observe that both *I. butyriciproducens* strains encode a vitamin B12 ABC transporter comprising the membrane proteins BtuF (SRB521_01945, SRB521_01948 and AF_01368c, AF_01372c) and BtuC (AF_01370c, SRB521_01947), as well as the iron B12 siderophore hemin periplasmic substrate-binding component. This suggests that the *I. butyriciproducens* strains have the ability to exchange B12 with the surrounding environment and hence are contributing to the intestinal microbe-host symbiosis.

Conclusions

The 16S rRNA amplicon metadata search revealed that *I. butyriciproducens* is the most common known species within the *Intestinimonas* genus and is most prevalent in humans. Comparative genomic and physiological analysis of *I. butyriciproducens* strains AF211 and SRB521^T isolated from the human and mouse intestine, respectively, confirmed the so far unique capability of using lysine for growth as a major feature of the species. Lysine is abundant in the gut as it is part of human diet and is released via microbial activity. In healthy subjects lysine is usually excreted in the faeces and hence its conversion by *I. butyriciproducens* prevents this loss. Moreover, *I. butyriciproducens*

converts lysine into butyrate, which is an energy source for colonocytes, and releases ammonium, which can be used as nitrogen source by bacteria and neutralises acidity created by SCFA production. Both strains of *I. butyriciproducens* contained 2 butyrate synthesis pathways, where both lysine and simple sugars can serve as energy source. However, physiological data suggested that AF211 is more efficient in utilizing human-specific sugars such as arabinose and galactose in presence of acetate. In conclusion, the present study provides detailed genomic and physiological insight into newly discovered *I. butyriciproducens* strains with unique metabolic capabilities and differentiates two strains of this species with different host origins.

Acknowledgements

Supplementary information

Table S1: Physiological and biochemical comparison of *I. butyriproducens* strain AF211 and SRB521[T]. The experiments were performed by using API Rapid 32 (A), API 20NE (B), API ZYM (C). +: positive response; -: negative response; w: weak response; vw: very weak response.

A

API rapid 32	AF211	SRB521[T]
Urease	-	-
arginine dihydrolase	-	-
α-galactosidase	-	-
β-galactosidase	-	-
β-galactosidase 6 phosphate	-	-
α-glucosidase	-	-
β-glucosidase	-	-
α-arabinosidase	-	-
β-glucuronidase	-	-
N-acetyl-β-glucosaminidase	-	+
mannose fermentation	-	-
rafinose fermentation	-	-
glutamic acid decarboxylase	-	-
α-fucosidase	-	-
reduction of nitrates	-	-
idole production	-	-
alkaline phosphatase	-	-
arginine arylamidase	-	-
proline arylamidase	-	-
leucyl glycine arylamidase	+	w
phenylalanine arylamidase	-	-
leucine arylamidase	-	-
pyroglutamic acid arylamidase	-	-
tyrosine arylamidase	-	-
alamine arylamidase	-	-
glycine arylamidase	-	-
histidine arylamidase	w	w
glutamyl glutamic acid arylamidase	-	-
serine arylamidase	w	w

B

API 20NE	AF211	SRB521[T]
indole	-	-
urease	-	-
glucose	w	w
manitol	vw	-
lactose	vw	-
saccharose	vw	-
maltose	vw	-
salicin	vw	-
xylose	vw	-
arabinose	vw	-
gelatin	-	-
esculin	-	-
glycerol	-	-
cellobiose	-	-
mannose	-	-
melezitose	vw	-
raffinose	vw	-
sorbitol	vw	-
rhamnose	vw	-
trehalose	vw	-
catalase	vw	-

C

API ZYM	AF211	SRB521[T]
Alkaline phosphatase	-	-
esterase E4	w	w
Esterase lipase C8	w	w
Lipase C14	-	-
Leucine arylamidase	-	-
Valine arylamidase	-	-
Cystine arylamidase	-	-
Trypsin	-	-
α-chymotrypsin	-	-
Acid phosphatase	w	+
Naphthol-AS-BI-phosphohydrolase	+	+
α-galactosidase	-	-
β-galactosidase	-	-
β-glucuronidase	-	-
α-glucosidase	-	-
β-glucosidase	-	-
α-mannosidase	-	-
α-fucosidase	-	-

Table S2: Cellular fatty acid composition of the strain AF211 and SRB521[T]. Data were obtained in the present study. All strains were grown in RCM for 5 days at 37°C. Values are percentages of total cellular fatty acids.

Cellular fatty acids	*I. butyriciproducens* AF211	*I. butyriciproducens* SRB521[T]
12 : 00	4.33	3.15
13 : 1	0.19	0.55
14 : 00	32.08	39.83
16 : 00	3.12	2.3
18 : 00	12.79	4.06
19 : 00	ND	4.95
19 : 1 iso I	31.69	37.9
20 : 00	2.91	1.53
Summed feature 1 (15:1 iso H/13:0 3OH)	1.94	0.91
Summed feature 4 (17:1 iso I/antaiso B)	9.05	4.83

Table S3: Number of genes associated to oxidative stress in AF211 and SRB521^T.

Functions	SRB521^T	AF211
CoA-disulfide reductase	1	1
Peroxide stress regulator PerR, FUR family	1	1
Catalase	2	2
Manganese superoxide dismutase	2	2
Rubredoxin	2	3
Rubrerythrin	6	6
Superoxide reductase	2	2
Alkyl hydroperoxide reductase subunit C-like protein	0	1

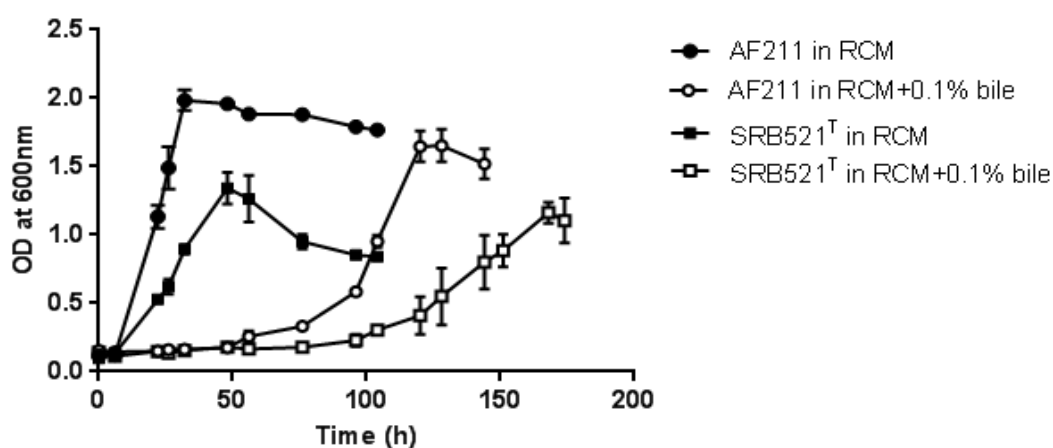**Figure S1:** Growth of strains AF211 and SRB521^T in the presence and absence of 0.1 % bile. Strains were grown in RCM and RCM plus 0.1 % bile (Oxgall). The doubling times of strain AF211 with and without bile are 6.6 hours and 23.6 hours (17 hours difference) while those of strain SRB521^T are 11.7 hours and 32.1 hours (20.4 hours difference), respectively. Values are mean of duplicates. Error bars are standard deviations.

A

B
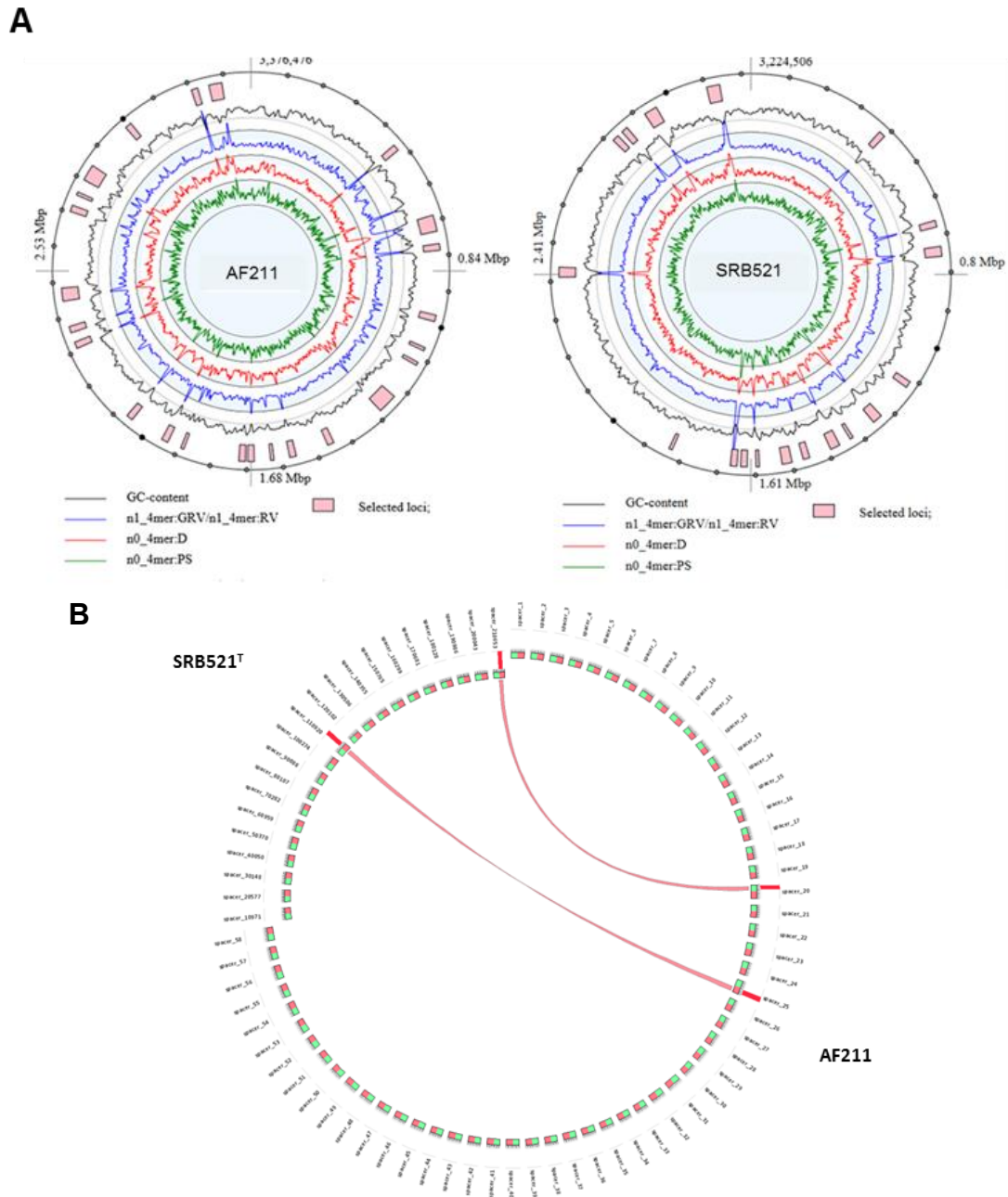
Figure S2: Genomic islands and spacer comparison. **A]** Positions of Genomic islands in both AF211 and SRB521[T] as predicted by SeqWord Sniffer. **B]** Spacer comparison.

Prophage genes detected in AF211

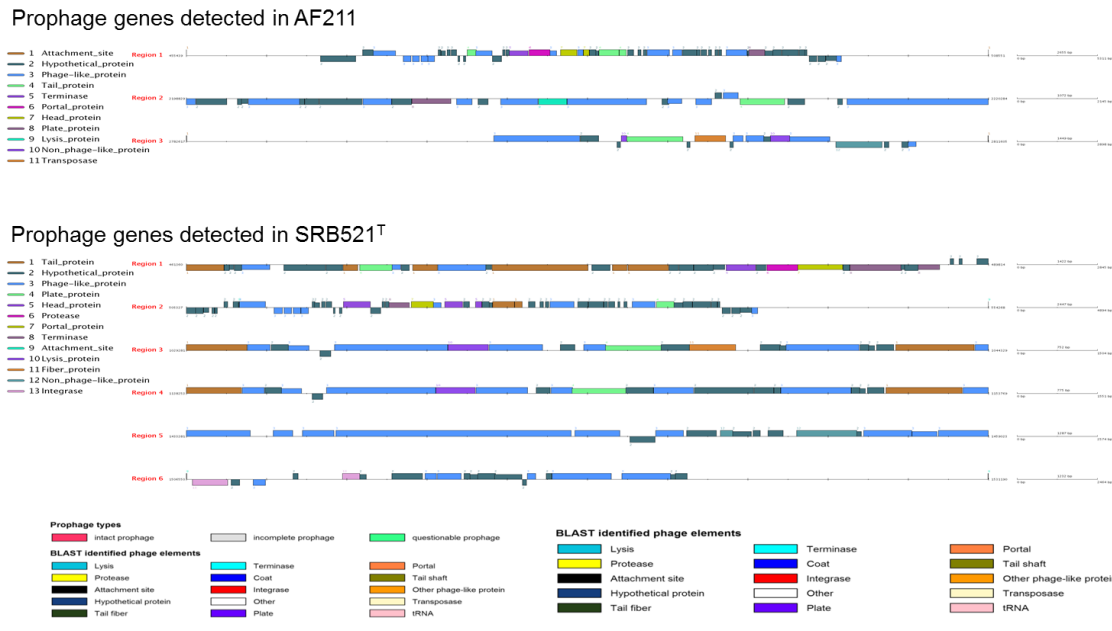
Prophage genes detected in SRB521^{T}

Figure S3: Prophage regions detected in *I. butyriciproducens* strain AF211 and SRB521^{T} genome sequences.
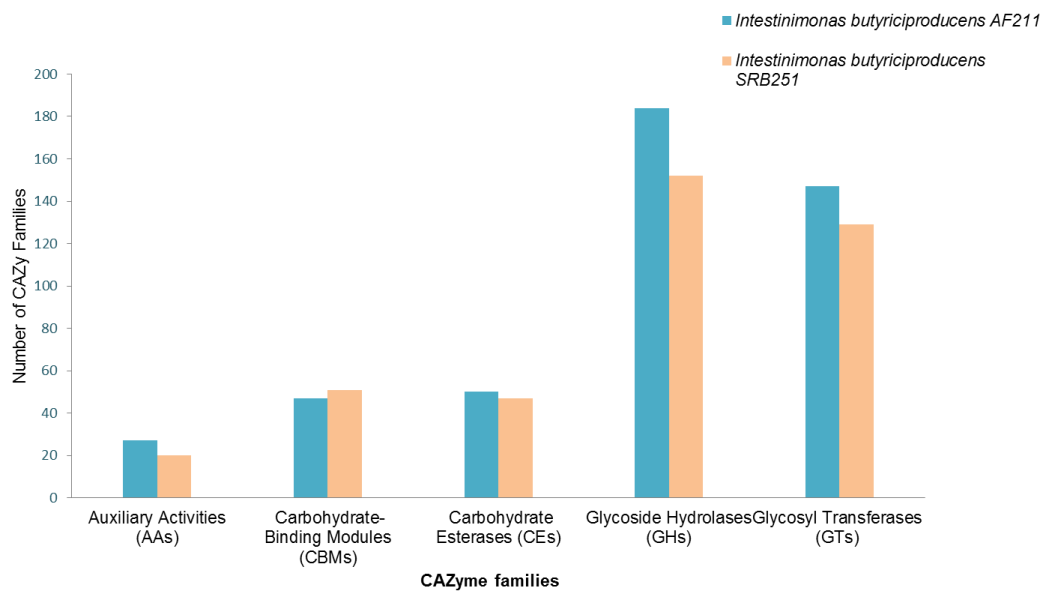
Figure S4: CAZyme families detected in AF211 and SRB521^{T}

Concentration (mM)
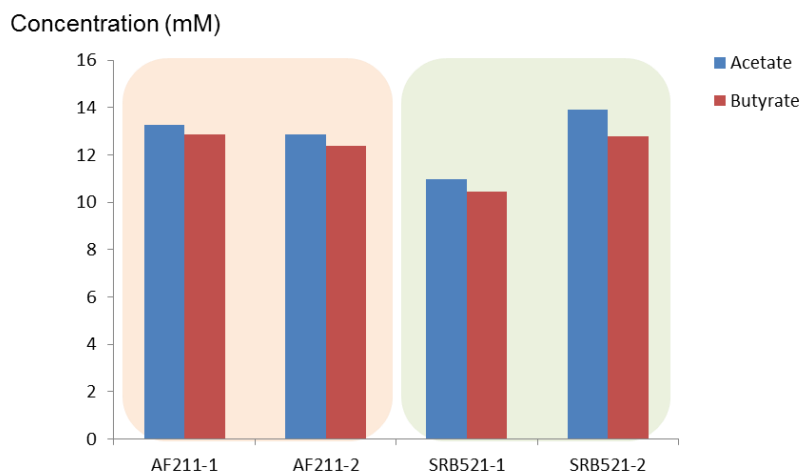
Figure S5: Lysine fermentation by strain AF211 and SRB521[T]. Product formation was determined after 2 days of growth in bicarbonate buffered medium (mM). Orange and green boxes show the results from duplicate experiments of strain AF211 and SRB521[T], respectively. 15 mM L-lysine was added as substrate.
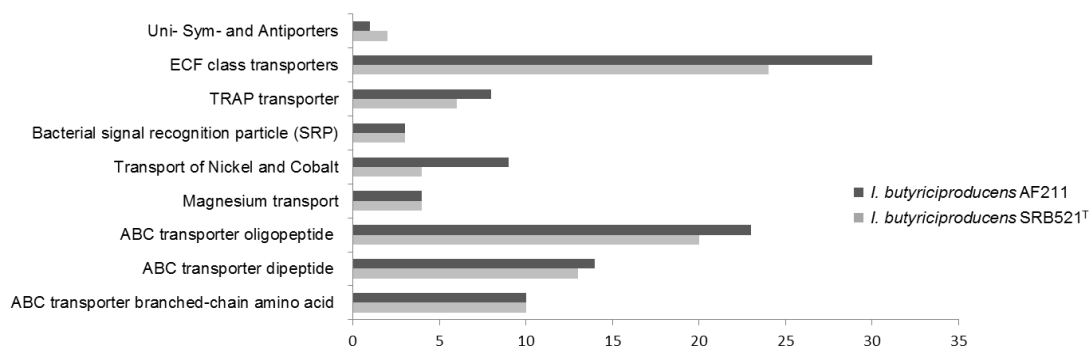
Figure S6: Annotated transport systems in the genomes of AF211 and SRB521[T]

FigureS7 Pseudovitamin-B12 production by *Intestinimonas butyriciproducens* AF211. **A]** UHPLC-UV (361 nm) chromatogram of the immunoaffinity purified cell extract of AF211 and **B]** the mass spectrum of the pseudovitamin B12 peak eluting at 3.16 min on performing MS/MS of the protonated molecular ions in a quadrupole time-of-flight (Q-TOF) mass spectrometer.

Supplement methods

Metadata search of *Intestinimonas* spp. using 16S rRNA gene amplicons

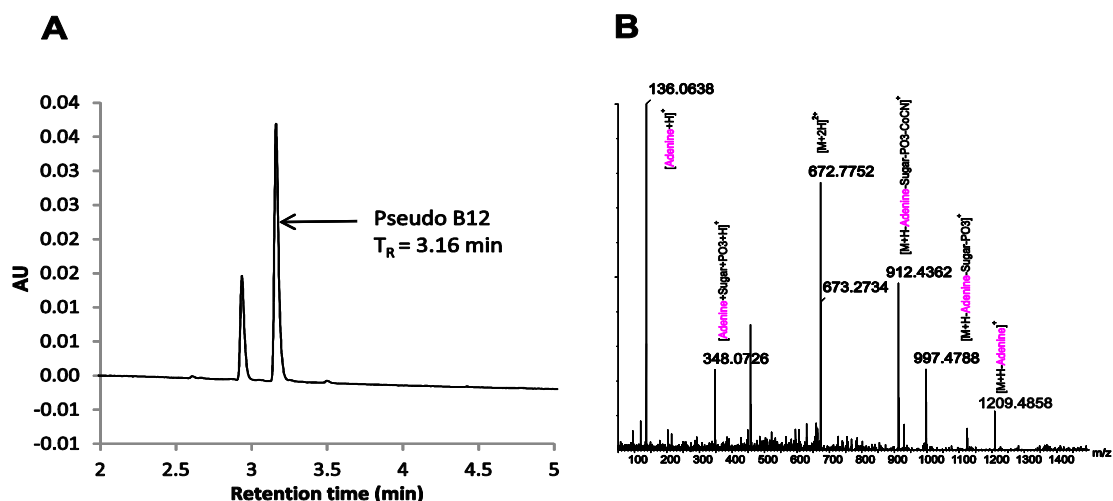The 16S rRNA gene of the type strain of *I. butyriciproducens* strain SRB521^T (NR_118554) was used for a similarity query (95 %) against all 16S rRNA samples available from the Sequence Read Archive (SRA) using the Integrated Microbial Next Generation Sequencing platform (www.imngs.org). IMNGS is an innovative web tool, which routinely checks and retrieves raw data from SRA and uniformly processes them to searchable operational taxonomic units (OTU) tables using a modified version of the UPARSE pipeline (Edgar 2013). Of note, IMNGS processes SRA samples individually, generating a list of OTUs for each sample, leading to redundancy of OTUs between samples. This feature is essential for flexibility of the system and easy comparison of OTU profiles between samples. Hence, the term sample-specific OTUs is used to refer to OTUs generated by IMNGS, and these OTUs can be used to address targeted microbial diversity.

Based on RDP classification (Cole*, et al.* 2014), all 16S rRNA reads that matched the sequence of *I. butyriciproducens* but were not categorized as unknown *Ruminococcaceae* (the current lineage of *Intestinimonas*) were removed. A local BLAST (Altschul*, et al.* 1990) database was built with the remaining reads, and those with >95 % similarity to representative sequences of the two neighbouring genera *Flavonifractor* and *Pseudoflavonifractor* were removed. Remaining reads (referred to as true SRA-derived *Intestinimonas* sequences) were aligned using SINA (Pruesse*, et al.* 2012) in SILVA (Quast*, et al.* 2013). A segment of highest coverage spanning the V4 hypervariable region was identified and the alignment was trimmed around this region (approx. 300 nt). Pairwise similarity across all remaining sequences covering this region was calculated in ESPRIT (Sun*, et al.* 2009). Based on clusters calculated at 5 % dissimilarity (considered as genus level cut-off in the present analysis) using average linkage, reads that fell into the cluster containing the reference *Intestinimonas* sequence were selected and their pairwise distance matrix was used to produce a multidimensional scaling (MDS) plot in the R programming environment. For the sake of clarity, points in the MDS plot were clustered in hexagonal bins with the number of

sequences in each bin visualized as a color gradient. A representative amplicon sequence together with the longest 100 % identical clone-derived sequences from GenBank were selected for each of the most abundant bins, and their pairwise similarity to *I. butyriciproducens* was calculated. Additional clones classified as *Intestinimonas* were collected from SILVA and those that formed a monophyletic group with the reference *Intestinimonas* sequence were kept. These clones, the clones aforementioned matching most abundant bins in the MDS plot, and 16S rRNA sequences from the two neighbouring genera and the outgroup *Faecalibacterium prausnittzii* were used for calculation of a phylogenetic tree using Maximum Likelihood based on the General Time Reversible model with invariant sites and 500 bootstraps in MEGA6 (Tamura*, et al.* 2013). All positions containing gaps and missing data were eliminated, leaving a total of 1,316 positions in the final dataset. For ecological reconstruction, each SRA-derived read from the filtered *Intestinimonas* dataset was assigned to the closest sequence out of the 5 selected clones mentioned above and the reference *I. butyriciproducens* sequence using BLAST. Sample-specific environmental information corresponding to the reads was represented as histograms next to the tree. Finally, all short reads (5,732) that were closest to the sequences within the monophyletic *Intestinimonas* branch were clustered to species-like groups (approx. 3% similarity) using CROP (Hao*, et al.* 2011).

Chapter 6

Reclassification of *Eubacterium hallii* as *Anaerobutyricum hallii* gen. nov., comb. nov., and description of *Anaerobutyricum soehngenii* sp. nov., a butyrate and propionate-producing bacterium from infant faeces

Sudarshan A. Shetty, Simone Zuffa, Thi Phuong Nam Bui, Steven Aalvink,

Hauke Smidt, Willem M. De Vos
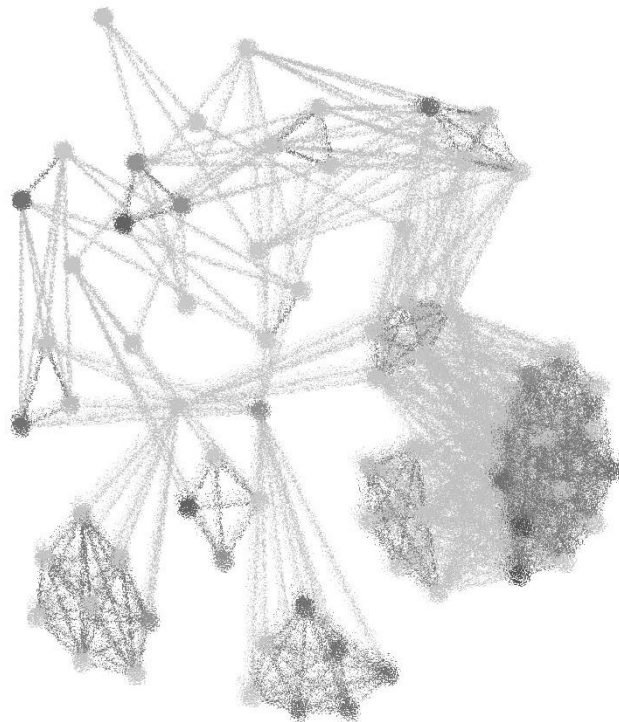
Abstract

A bacterial strain designated L2-7, phylogenetically related to *Eubacterium hallii* DSM3353[T], was previously isolated from infant faeces. The complete genome of strain L2-7 contains eight copies of the 16S rRNA gene with only 98.0-98.5% similarity to the 16S rRNA gene of the previously described type strain *Eubacterium hallii*. The next closest validly described species is *Anaerostipes hadrus* DSM 3319[T] (90.7% 16S rRNA gene similarity). A polyphasic taxonomic approach showed strain L2-7 to be a novel species, related to type strain *E. hallii* DSM3353[T]. The experimentally observed DNA–DNA hybridization value between strain L2-7 and *E. hallii* DSM3353[T] was 26.25%, close to that calculated from the genomes (34.3%). The G+C content of the chromosomal DNA of strain L2-7 was 38.6 mol%. The major fatty acids were $C_{16:0}$, $C_{16:1}$ cis 9 and a component with summed feature 10 ($C_{18:1}$c11/t9/t6c). Strain L2-7 had higher amounts of $C_{16:0}$ (30.59%) compared to *E. hallii* DSM3353[T] (19.50%) and its membrane contained phosphatidylglycerol and phosphatidylethanolamine, which were not detected in *E. hallii* DSM3353[T]. Furthermore, 16S rRNA gene phylogenetic analysis advocates that *E. hallii* DSM3353[T] is misclassified, and that its reclassification as a member of the family *Lachnospiraceae* is necessary. Using a polyphasic approach, we propose that *E. hallii* (=DSM3353[T] =ATCC27751[T]) be reclassified as the type strain of a novel genus *Anaerobutyricum* gen. comb. nov., sp. nov., comb. nov. and we propose that isolate L2-7 should be classified within the genus *Anaerobutyricum* as a novel species, *Anaerobutyricum soehngenii sp. nov.* The type strain is L2-7[T] (=DSM 17630[T] =KCTC 15707[T]).

Key words: *Eubacterium hallii, Anaerobutyricum hallii, Anaerobutyricum soehngenii*

Introduction

In a previous study investigating the phylogenetic diversity of butyrate-producing bacteria in the human intestinal tract, strain L2-7 with highest 16S ribosomal RNA (rRNA) similarity to *Eubacterium hallii* DSM3353^T was isolated from infant faeces (Barcenilla*, et al.* 2000). This isolate, deposited as DSM17630, was capable of producing butyrate from glucose, and later it was shown to convert both D- and L- lactate to butyrate in the presence of acetate (Duncan*, et al.* 2004). Subsequent culture-independent analysis based on 16S rRNA gene profiling revealed bacteria related to *Eubacterium hallii* to be part of the core microbiota of the human intestinal tract, and oral administration of cells of strain L2-7 was observed to improve insulin sensitivity in db/db mice (Engels*, et al.* 2016b, Udayappan*, et al.* 2016b, Shetty*, et al.* 2017a). Importantly, the revised roadmap of 2009 for taxonomic classification of the phylum *Firmicutes* proposed that *Eubacterium stricto sensu* includes only *Eubacterium limosum, E. aggregans, E. barkeri, and E. callanderi* as well as the genera *Acetobacterium*, *Alkalibacter*, *Anaerofustis*, *Garciella* and *Pseudoramibacter* (Ludwig*, et al.* 2009). In contrast, the species *Eubacterium hallii* has been placed in the *Lachnospiraceae* family. In order to have a better understanding of the taxonomic affiliation of strains L2-7 and DSM3353^T, we carried out a polyphasic taxonomic approach.

Materials and Methods

Bacterial strains and culture conditions

Strain L2-7 was kindly provided by Dr. Harry Flint, while strain DSM3353^T was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ, Braunschweig, Germany). Both strains were grown routinely in a medium with following composition: yeast extract (4.0 g/L), casitone (2.0 g/L), soy peptone (2.0 g/L), NaHCO$_3$ (4.0 g/L), KH$_2$PO$_4$ (0.41 g/L), MgCl$_2$.6H$_2$O (0.1 g/L), CaCl$_2$.2H$_2$O (0.11 g/L), Cysteine-HCL (0.5 g/L), vitamin K1 (0.2ml), hemin (1ml), and trace elements I, trace elements II and vitamin solutions. The trace elements I (alkaline) solution contained the following (mM): 0.1 Na$_2$SeO$_3$, 0.1 Na$_2$WO$_4$, 0.1 Na$_2$MoO$_4$ and 10 NaOH. The trace elements II (acid) solution was composed of the following (mM): 7.5 FeCl$_2$, 1 H$_3$BO$_4$, 0.5 ZnCl$_2$, 0.1 CuCl$_2$, 0.5 MnCl$_2$, 0.5 CoCl$_2$, 0.1 NiCl$_2$ and 50 HCl. The vitamin solution had

the following composition (g/L): 0.02 biotin, 0.2 niacin, 0.5 pyridoxine, 0.1 riboflavin, 0.2 thiamine, 0.1 cyanocobalamin, 0.1 p-aminobenzoic acid and 0.1 pantothenic acid. The medium was supplemented with 30mM of sodium acetate. For routine use, the medium was distributed in 35 ml serum bottles sealed with butyl-rubber stoppers and incubated at 37 °C under a gas phase of 1.7 atm (172 kPa) N_2/CO_2 (80 : 20, v/v). The pH of the medium was 7.0. The vitamin solution was filter sterilized and added after autoclaving. Similarly, filter sterilized D-Glucose or other carbon sources for substrate utilization tests, were added to a final concentration of 40 mM, unless stated otherwise. All growth experiments with strains L2-7 and DSM3353[T] were performed in duplicate. The experiments were repeated independently to confirm the observed differences. Resistance to porcine (Sigma-Aldrich, GmbH) and Oxgall bile (Fluka Analytical) was tested in the standard medium.

Cell morphology, motility and spore formation of strain L2-7 was observed using a Leica DM 2000 microscope and a JEOL-6480LV scanning electron microscope (SEM). The sample processing for SEM was done as previously described (Jarzembowska*, et al.* 2016).

For the pH experiment, the buffers used were as follows; 20 mM Tris to buffer at ≥pH 8.0, 20 mM PIPES for pH 6.0–7.5 and 20 mM sodium citrate for <pH 6.0 and adjusted with 2 M HCl and 2 M NaOH. To check for the ability to grow in the presence of oxygen, the strains were inoculated in media without reducing agent (cysteine-HCl) and checked for growth after 24 h.

The cellular fatty acid contents of strain L2-7 and *E. hallii* DSM 3353[T] were compared from cells grown for 12 h on the standard medium. The cellular fatty acids were then extracted by following the standard protocol of the Sherlock Microbial Identification Systems (MIDI, 1999) and identified using the Sherlock version 4.0 and BHIBLA 3.80 libraries at DSMZ, Braunschweig (Kuykendall*, et al.* 1988, Kämpfer and Kroppenstedt 1996, Miller 1982).

The type of peptidoglycan membrane for L2-7 and DSM 3353[T] was determined at DSMZ, using the methods described previously (Schumann 2011).

Molecular analysis

For genomic comparison, we used the recently deposited complete genome of strain L2-7 (accession number LT907978) and the draft genome sequence of the type strain *E. hallii* DSM3353[T] (accession number PRJNA18177) (Sayers, *et al.* 2012, Shetty, *et al.* 2017b). A total of eight full-length 16S rRNA gene sequences were obtained from the complete genome of strain L2-7 (Shetty, *et al.* 2017b). For a more detailed 16S rRNA gene based phylogenetic analysis, sequences were aligned using ClustalW and a phylogenetic tree was constructed using MEGA6 (for details, see legend to Figure 1).

Results and Discussion

Genome comparison suggests novel strain L2-7 and DSM3353 are two distinct species

The 16S rRNA gene phylogenetic analysis indicated that the closest relatives of strain L2-7 were members of the family *Lachnospiraceae* (Figure 1), with *E. hallii* DSM3353[T] being the nearest neighbour. The pairwise comparison of the almost-complete 16S rRNA gene sequences (1400 nt) to that of *E. hallii* DSM3353[T] revealed approximately 1.0 to 1.4 % sequence divergence depending on the copy of L2-7 16S rRNA gene. The validly described species second-most closely related to strains L2-7 and DSM3353[T] is *Anaerostipes hadrus* DSM 3319[T] (Allen-Vercoe, *et al.* 2012), with a mean 16S rRNA gene sequence similarity of 90.7%. Next, using the genome sequence of strains L2-7 DSM17630 and DSM3353[T] we calculated the average nucleotide identity (ANI) and *in-silico* DNA-DNA hybridization using the Genome-to-Genome Distance Calculator (GGDC, DSMZ) (Goris, *et al.* 2007, Rodriguez-R and Konstantinidis 2014, Henz, *et al.* 2004). The mean ANI between the two isolates was 87.81 and the *in-silico* DNA-DNA hybridization estimate (GLM-based) was 34.3% [31.9 - 36.8%]), suggesting that isolate L2-7 could be a novel species. Moreover, supporting evidence is provided by *in vitro* DNA-DNA hybridisation between strains L2-7 and *E*. *hallii* DSM3353[T], which was only 26.5%. This is well below the 70 % cut-off point generally accepted for species classification (Wade 2006). These data warranted a more detailed characterization using a polyphasic taxonomy approach.
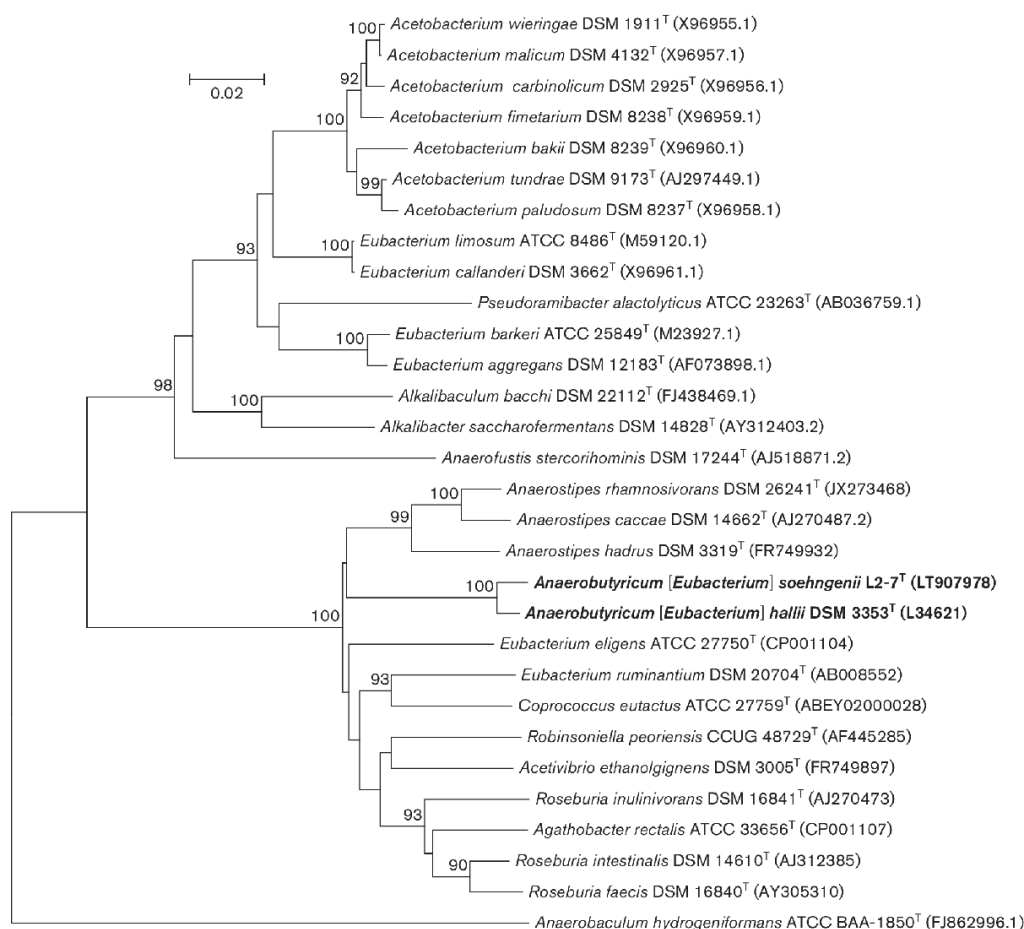
Figure 1: Rooted phylogenetic tree showing the relationship of *Anaerobutyricum* [*Eubacterium*] *hallii comb. nov.*, DSM3353^T and strain L2-7 with selected members of the families *Lachnospiraceae* and *Eubacteriaceae* constructed in MEGA6 (Tamura*, et al.* 2013). The 16S rRNA gene of *Anaerobaculum hydrogenifromans* ATCC BAA-1850 was used as an outgroup. The tree was constructed using the Neighbor-Joining method based on pairwise distances obtained using the Kimura 2-parameter. The bootstrap test included 1000 replicates, and only bootstrap values >90% are shown (Saitou and Nei 1987, Felsenstein 1985).

Biochemical and physiological comparisons

For identifying the optimum growth temperature, the isolates were incubated at different temperatures (10, 20, 25, 30, 35, 37, 40, 45, 50 and 55 ˚C). Strains L2-7 and DSM3353^T both grew at temperatures from 35 to 45 ˚C with optimum growth at 37 ˚C. Strain L2-7 is a non-motile and strictly anaerobic, straight rod-shaped organism. While we observed sub-terminal and terminal swellings in both the isolates L2-7 and DSM3353^T as previously reported for DSM3353^T (Wade 2006, Vos*, et al.* 2011), the cultures did not survive after heat treatment (80°C for 10 min). Young cells of both strains stained Gram-positive (Figure S1 A and B). The cells were approximately 0.70–4µm in size after 12h of incubation in glucose-containing growth media (Figure S1 C

and D). Cells differed slightly in length depending on the growth medium and appeared sometimes elongated under nutrient, temperature and pH stress conditions. After 48 h of incubation on Reinforced Clostridia Medium (solidified with 1 % agar; Difco) at 37 °C, both strains L2-7 and DSM3353^T produced white centred colonies, surrounded by a colourless layer, circular, raised with entire margins and 4.5mm diameter. Under the microscope cells are elongated. Strains L2-7 and DSM3353^T were grown at pH 4.5–8.5, and growth was observed in the pH range 6.0- 8.0, and the optimum growth was observed at pH 6.5 for strain L2-7 and pH 7.5 for DSM3353^T. No growth was observed in presence of oxygen confirming that they are strictly anaerobic bacteria. Aesculin is not hydrolysed. Indole is not produced.

The biochemical characterization of the isolate L2-7 and the type strain DSM3353^T gave inconsistent results on API ZYM and hence those data were not included in the results. This issue with use of commercial rapid identification kits for testing biochemical properties for this group of bacteria is already known (Dworkin 2006). In addition, we also tested fermentation end products on various carbohydrates (Table 1). After 24 h of incubation, the major fermentation products of glucose metabolism for strain L2-7 were butyrate (23.2 mM) and formate (11.9 mM). Similar values for butyrate (22.6 mM) and formate (10.2 mM) were observed for DSM3353^T as end products of glucose metabolism. For L2-7, the major end products of 1,2 propanediol (80 mM) utilisation were propionate (23.2 mM), propanal (15.4 mM) and 1-propanol (14.5 mM). For DSM3353^T, the end products of 1,2 propanediol utilisation were propionate (26.3 mM), propanal (19.1 mM) and 1-propanol (20.5 mM). The end products observed for 1,2-propanediol usage are in accordance with a previous study (Engels*, et al.* 2016b). After 24 h, strain L2-7 utilized 37.8mM of D,L-lactate and 16.8 mM of acetate and produced 24.3 mM of butyrate while minor amounts of formate were detected. On the other hand after 24 h, strain DSM3353^T utilized 26.8 mM of D,L-lactate and 10 mM acetate to produce 22.9 mM of butyrate. CO_2 was not measured. Strain DSM3353^T was resistant to 0.1 to 1% (w/v) porcine bile and 0.1 and 0.5 % (w/v) Oxgall. On the other hand, strain L2-7 (DSM17630) was not resistant to porcine bile but was resistant to 0.1% Oxgall and susceptible to higher concentrations.

Table 1: Substrate utilization for type strain DSM3353^T and strain DSM17630. The major end product for all substrates, except for 1,2 propanediol, was butyrate. For 1,2 propanediol the major end products were propionate and 1-propanol. Substrates for which differences in utilization between the two strains were observed are highlighted in bold.

	Strain L2-7 DSM17630	Type strain DSM3353^T
Glucose	+	+
Fructose	+	+
Galactose	+	+
Maltose	+	+
Sucrose	**+**	**-**
Xylose	-	-
Arabinose	-	-
Mannose	+	+
Lactose	**-**	**+**
Ribose	**-**	**+**
Maltose	+	+
Sorbitol	+	+
Mannitol	**-**	**+**
Cellbiose	-	-
Rhamnose	-	-
1,2 Propanediol	+	+

The DNA G+C content of strain L2-7 was 38.6 mol % and for DSM3353^T was 38.2 mol % as determined from the genome sequence. Gas chromatography indicated that C16:0 is the predominant fatty acid in both the isolates, with C16:0 being higher in strain L2-7 compared to *E. hallii* DSM 3353^T (Table 2). Overall similarity in the fatty acid profile indicated that both strains belong to the same genus.

The cellular fatty acid profile was composed mainly of C16:0, C16:1 cis 9 and summed feature 10 (C18:1c11/t9/t6c). The polar lipid patterns clearly differed between strains L2-7 and DSM3353^T (Figure S2) and comprised of unknown phospholipids (PL), unknown lipids (L) and phospoaminoglycolipid. DSM3353^T had additional aminolipid (AL) and phosphoglycolipid (PGNL) and two unknown lipids (L). L2-7 was observed to have phosphatidylglycerol (PGL), phosphatidylethanolamine (PE), and two unknown phospholipids (PL).

The peptidoglycan of both the strains is of the A1γ (A31) type containing meso-diaminopimelic acid (meso-Dpm). This type of peptidoglycan has also been reported recently to be present in isolates of the genus *Faecalicatena* that were reclassified from *Eubacteriaceae* to *Lachnospiraceae* (Sakamoto*, et al.* 2017).

Table 2: Cellular fatty acid profiles for type strain DSM3353[T] and strain DSM17630. Fatty acids that for which differences between both strains were observed are highlighted in bold.

Fatty acids	Strain L2-7 DSM17630	Type strain DSM3353[T]
10:0	0.1	0.2
12:0	0.4	0.4
14:0	5.0	5.7
14:0 FAME	ND	5.7
14:0 DMA	ND	0.5
16:0 ALDE	0.5	1.5
16:1 CIS 7 FAME	3.2	5.2
16:1 CIS 9 FAME	**5.5**	**11.7**
16:1 CIS 11 FAME	0.9	1.4
16:0 FAME	**30.4**	**19.5**
16:1 CIS 9 DMA	2.6	5.0
16:0 DMA	1.9	5.5
18:0 ADLE	0.2	ND
16:0 3OH FAME	0.4	0.5
18:1 CIS 9 FAME	4.3	3.2
18:1 CIS 13 FAME	0.6	0.3
18:0 FAME	3.7	1.2
18:1 CIS 9 DMA	4.6	3.2
18:1 CIS 11 DMA	11.4	8.8
18:0 DMA	0.8	0.5
20:1 cl3/t11 FAME	0.5	ND
19:0 CYC 11,12 DMA	ND	0.3
Summed feature 1	ND	0.2
Summed feature 4	1.0	2.6
Summed feature 6	0.5	2.4
Summed feature 7	1.1	0.8
Summed feature 8	3.1	2.5
Summed feature 10	16.9	16.8

Conclusion

The notion that the initial classification of DSM3353[T] as a species of *Eubacterium* is incorrect has been confirmed by phylogenetic analyses presented here, and thus we propose that it should be reclassified as *Anaerobutyricum hallii* comb. nov. within the family *Lachnospiraceae* (Wade 2006). In addition, based on the results of our comparative polyphasic characterization, it is appropriate to assign strain L2-7 to a novel species within the novel genus *Anaerobutyricum*, for which the name *Anaerobutyricum soehngenii* sp. nov. is proposed.

Description of *Anaerobutyricum* gen. nov. (Holdeman and Moore 1974)

Anaerobutyricum (*an. ae. ro. bu.ty'ri.cum Gr. pref. an* not*; Gr. n. aer* air*; L. masc. n. butyricum,* butyrate producing, anaerobic butyrate producing) is Gram-positive-staining, non-motile, with rod-shaped cells. Obligately anaerobic and catalase- and oxidase-negative. Optimal growth is observed at 37 ˚C. Monosaccharides galactose, fructose and glucose, disaccharides such as maltose and the sugar alcohol sorbitol are fermented. Mannitol can be fermented by *A. hallii*, but not by *A. soehngenii*. The major end-products of sugar fermentation are butyrate, formate, H_2 and CO_2. When grown on 1,2 propanediol, propionate, propanal and 1-propanol are produced. It requires acetate for better growth on substrates such as D,L-lactate and produces butyrate as major end product and small amounts of formate. The G+C content of the DNA is 38.2 ±4 mol%. The type species of the genus is *Anaerobutyricum hallii* (Holdeman and Moore 1974) (Moore and Holdeman 1974). Known species within this genus tolerate 0.1% oxgall.

Description of *Anaerobutyricum hallii* comb. nov.

Basonym: *Eubacterium hallii* (Holdeman and Moore 1974) Approved Lists 1980

Description is as given by Holdeman and Moore 1974, in addition to the original description of *Eubacterium hallii* (Holdeman and Moore 1974), the species requires acetate to grow on lactate. It can grow well on lactose, ribose and mannitol but not on sucrose. In accordance with Holdeman and Moore 1974, weak growth for *A. hallii* is observed in ribose and growth is seen after a 24 h lag phase. The major cellular fatty acids are C14:0 FAME, iso-C19 : 1, C16 : 0 and C18:1 CIS 11 DMA. The optimum pH for growth is 7.5 at 37ºC. The type strain for this species is *Anaerobutyricum hallii* (=DSM3353^T =ATCC27751^T).

Description of *Anaerobutyricum soehngenii* sp. nov.

Anaerobutyricum soehngenii (*soehn.ge`ni.i.N.L. gen.n. soehngenii* in honor of the first Wageningen Professor of Microbiology, Nicolaas L. Soehngen, for his contribution to anaerobic microbiology).

In addition to the characteristics of the genus, *A. soehngenii* exhibits the following features. Indole and catalase are not produced and it is negative for oxidase. Aesculin

is not hydrolysed. Acid is produced from D-glucose, maltose, galactose, sucrose, D-mannose, D-fructose and sorbitol. There is a net utilization of acetate when growing on sorbitol. It can produce butyrate when D,L-lactate and acetate are present in the medium. No growth was observed in D-xylose, L-arabinose, cellobiose, D-mannitol, lactose and D-rhamnose. The major cellular fatty acids are C16 : 0 FAME and 18:1 CIS 9 FAME, 18:0 FAME, and 18:1 CIS 11 DMA.

The type strain, L2-7^T (=DSM 17630^T =KCTC 15707 ^{T}), was isolated from human infant faeces. The DNA G+C content of the type strain is 38.6 mol%.

Acknowledgements

Supporting information
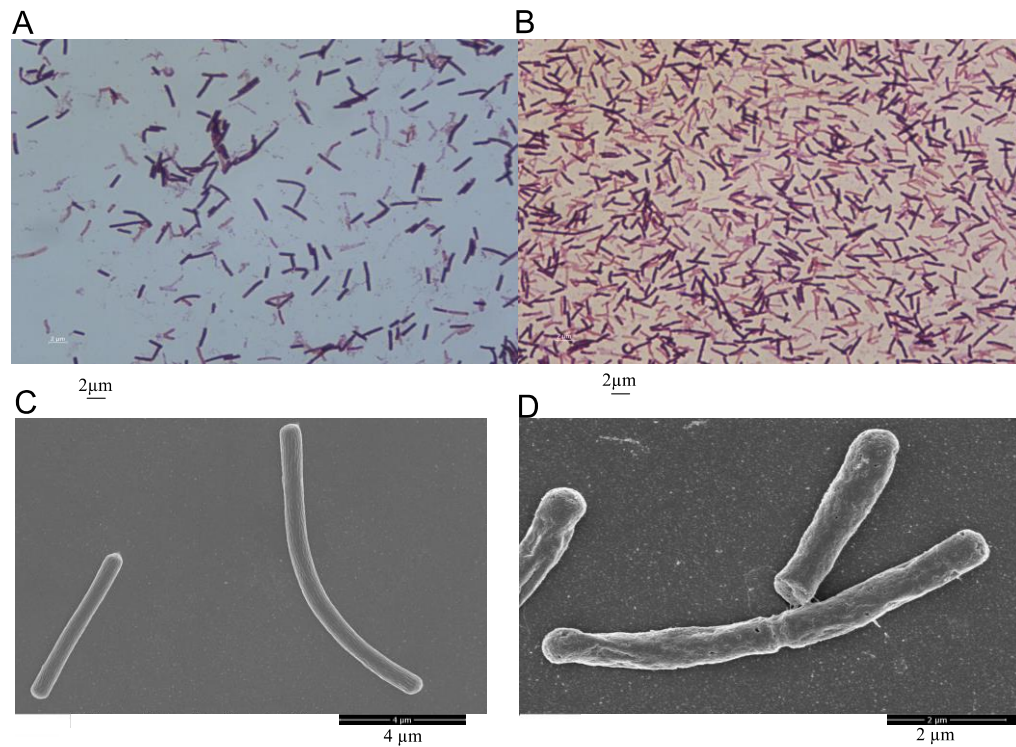

Figure S1: Gram staining for L2-7 DSM17630 (**A**) and DSM3353[T] (**B**) and scanning electron microscopy images of L2-7 DSM17630 (**C**) and DSM3353[T] (**D**).
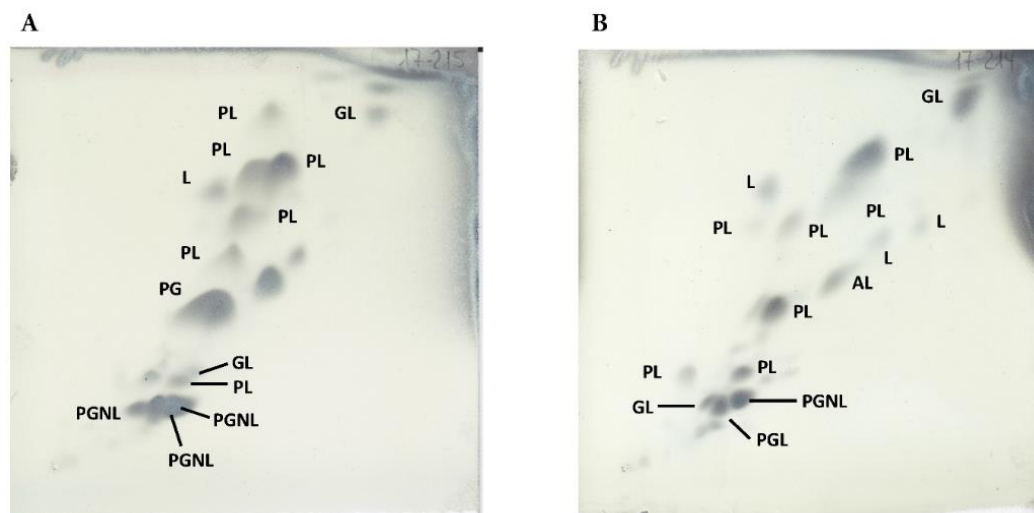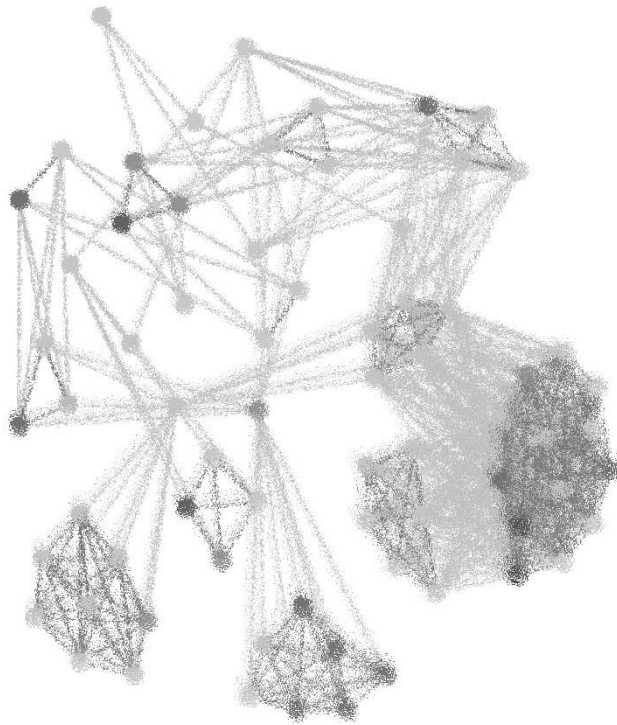

Figure S2: Polar lipid patterns for L2-7 (**A**) and DSM3353[T] (**B**).

Chapter 7

Ecophysiology of *Anaerobutyricum* in the human intestinal tract

Sudarshan A. Shetty, Sjef Boeren, Hauke Smidt, Willem M. de Vos

Abstract

More than 1000 microbial species of human intestinal tract origin have been successfully cultured in the laboratory. Yet, only few have been investigated in depth to understand their metabolic potential. Bacteria belonging to the genus *Anaerobutyricum* form part of the core microbiota in the human intestinal tract and are known butyrate producers that have been linked to host health. Here, we employ a combination of genomic, proteomic and physiological approaches to improve our understanding of the physiology and metabolic capabilities of *A. soehngenii*. Proteomic analysis of cells grown under different conditions revealed the complete pathway for butyrate production from glucose, sorbitol and lactate plus acetate. We identified the active genes involved in the *de novo* and salvage pathway for vitamin B12 biosynthesis. The 1,2 propanediol (PD) gene cluster, associated with genes for microbial cell compartment production, is constitutively expressed, and corresponding gene products were detected at high abundance in the proteome when grown on glucose, sorbitol and lactate plus acetate, which explains the preferential utilisation of 1,2 PD *in vitro*. We identified the unique organisation of a new lactate utilisation gene cluster, which suggests adaptation of *Anaerobutyricum* and related species of *Anaerostipes* and other Firmicutes to growth on DL-lactate plus acetate. Our study provides the first insights into the global proteomic response of *A. soehngenii* during growth on glucose and other carbon sources while providing support for the unique genetic adaptation of *Anaerobutyricum* spp. in producing butyrate from DL-lactate and acetate.

Keywords: *Anaerobutyriricum soehngenii*, core microbiota, butyrate producer, differential protein expression

Introduction

Bacteria related to *Anaerobutyricum hallii* (previously known as *Eubacterium hallii*) are prevalent in the human intestinal tract (Shetty*, et al.* 2017a). The novel genus *Anaerobutyricum,* which is taxonomically placed in the Lachnospiraceae family, includes presently two species, *A. hallii* and *A. soehngenii* (Shetty*, et al.* 2018). *A. soehngenii* strain L2-7 was isolated from infant feces and characterized for its growth on sugars as well as acetate and lactate (Barcenilla*, et al.* 2000). A characteristic feature of this bacterium is its ability to produce butyrate from different sugars, sugar alcohols, and converting the intestinal metabolite 1,2 propanediol (1,2 PD) to propionate in a reaction dependent on vitamin B12 that *A. soehngenii* it is able to produce itself (Engels*, et al.* 2016b, Belzer*, et al.* 2017). Butyrate is one of the important short chain fatty acids (SCFAs) produced in the human intestinal tract, and the majority (approximately 95%) is used to fuel colonic enterocytes (Topping and Clifton 2001). All SCFAs have several health benefits but specifically butyrate inhibits the proliferation and induces apoptosis of tumor cells (Thangaraju*, et al.* 2009). Butyrate has also been suggested to play a role in gene expression as it is a known inhibitor of histone deacetylases, and recently it was demonstrated that butyrate promotes histone crotonylation *in vitro* (Fellows*, et al.* 2018, Davie 2003). Related to this ability, butyrate is shown to upregulate endogenous host defence peptides, which in turn was shown to increase disease resistance in piglets (Xiong*, et al.* 2016). The second important SCFA is propionate, which is known for its anti-inflammatory properties, anti-microbial activity against *Salmonella,* and its capacity to increase satiety (Sa'ad*, et al.* 2010, Galisteo*, et al.* 2008). The metabolic routes to formation of these two important SCFAs have been widely studied in a variety of intestinal bacteria (Vital*, et al.* 2017, Louis and Flint 2017, Vital*, et al.* 2014a). It is important to understand these pathways in potential health-promoting bacteria such as *A. soehngenii* strain L2-7, which has shown to improve insulin response and reduce the expression levels of liver lipogenic genes in mouse models (Udayappan*, et al.* 2016a).

Another important small molecule metabolite produced uniquely by prokaryotes is cobalamin (vitamin B12) which is used as a co-factor for at least 15 different intestinal

bacterial enzymes (Degnan*, et al.* 2014a, Degnan*, et al.* 2014b). The most common enzyme is B12-dependent methionine synthase, which is present in 67% of all known intestinal bacteria (Degnan*, et al.* 2014b). Another relevant pathway is the methylmalonyl CoA pathway involved in propionate production (Degnan*, et al.* 2014b, Reichardt*, et al.* 2014). Nevertheless, only few bacteria are known to produce vitamin B12, whereas the majority depends on exogenous sources (Degnan*, et al.* 2014a, Magnúsdóttir*, et al.* 2015). Thus, vitamin B12 producing bacteria are considered to play an important role in shaping the microbial community (Degnan*, et al.* 2014b, Magnúsdóttir*, et al.* 2015). The functionality of the B12 produced by *A. soehngenii* L2-7 in promoting propionate formation by *Akkermansia muciniphila* via the methylmalonyl CoA pathway has recently been demonstrated (Belzer*, et al.* 2017). In addition, a recent study identified vitamin B12-dependent glycerol/diol dehydratases being encoded in the genomes of *A. hallii* and *A. soehngenii,* which was linked to their ability to produce propionate from 1,2 PD. However, based on genome annotation, a homologue of the gene *pduL*, the product of which converts propionyl-CoA to propionyl phosphate, was not detected (Engels*, et al.* 2016b). Since the ability to utilize 1,2 PD to produce propionate and the ability to produce vitamin B12 are key features of *A. hallii* and *A. soehngenii*, it is crucial to have a better understanding of active genes involved in these pathways. Additionally, the genes involved in the metabolic pathway that leads to the conversion of lactate and acetate into butyrate have not been described yet in *Anaerobutyricum* spp. and other related bacteria that share this property, such as *Anaerostipes* spp. (Barcenilla*, et al.* 2000). Knowledge of lactate utilization is particularly important because it is an end product of primary fermentation and energetically challenging to convert anaerobically (Weghoff*, et al.* 2015).

Here, we investigated the basic metabolic features of *Anaerobutyricum* spp. with an emphasis on *A.soehngenii* L2-7. We used a combination of physiological, genomic and proteomic approaches to reveal the active genes involved in the pathway for butyrogenesis from sugar, sugar alcohol and the common cross-feeding metabolite lactate, vitamin B12 biosynthesis pathways, and 1,2 PD utilization to produce propionate. The results provide support for the unique genetic adaptation of *Anaerobutyricum* spp. in producing butyrate from DL-lactate and acetate.

Materials and Methods

Bacterial strains used in this study

The two *Anaerobutyricum* species examined were *A. hallii* (DSM3353) and *A. soehngenii* (DSM17630). *A. hallii* (DSM3353) was purchased from Leibniz-Institute DSMZ-German Collection of Microorganisms and Cell Cultures (Braunschweig, Germany). *A. soehngenii* L2-7 (DSM17630) was kindly provided by Prof. Harry J. Flint from the Rowett Institute (Aberdeen, UK). The strains were grown routinely in a medium with following composition: yeast extract (4.0 g/L), casitone (2.0 g/L), soy peptone (2.0 g/L), $NaHCO_3$ (4.0 g/L), KH_2PO_4 (0.41 g/L), $MgCl_2.6H_2O$ (0.1 g/L), $CaCl_2.2H_2O$ (0.11 g/L), cysteine-HCL (0.5 g/L), vitamin K1 (0.2ml), hemin (1ml), and trace elements I, trace elements II and vitamin solutions. The trace elements I (alkaline) solution contained the following (mM): 0.1 Na_2SeO_3, 0.1 Na_2WO_4, 0.1 Na_2MoO_4 and 10 NaOH. The trace elements II (acid) solution was composed of the following (mM): 7.5 $FeCl_2$, 1 H_3BO_4, 0.5 $ZnCl_2$, 0.1 $CuCl_2$, 0.5 $MnCl_2$, 0.5 $CoCl_2$, 0.1 $NiCl_2$ and 50 HCl. The vitamin solution had the following composition (g/L): 0.02 biotin, 0.2 niacin, 0.5 pyridoxine, 0.1 riboflavin, 0.2 thiamine, 0.1 cyanocobalamin, 0.1 p-aminobenzoic acid and 0.1 pantothenic acid. The medium was supplemented with 30mM of sodium acetate. For routine use, the medium was distributed in 35 ml serum bottles sealed with butyl-rubber stoppers and incubated at 37 °C under a gas phase of 1.7 atm (172 kPa) N_2/CO_2 (80: 20, v/v). The pH of the medium was 7.0. The vitamin solution was filter sterilized and added to the media bottles after autoclaving.

Comparison of growth on different carbon sources

The two strains were grown on combinations of two carbon sources (40mM each) including lactose and glycerol (*A. hallii*), and sucrose and glycerol (*A. soehngenii*). Moreover, both strains were grown on 1,2 PD (80 mM, as this concentration provided exponential growth) and glycerol (40 mM). Samples (1ml) were taken at different stages of growth (0, 5, 8, 10, 24 h) for analysis of substrate utilization and measurement of SCFAs using high-performance liquid chromatography (HPLC) as described before (Bui, *et al.* 2016).

Identification of gene neighbourhood and phylogeny

Gene clusters were identified using the Joint Genome Institute Integrated Microbial Genomes & Microbiome System (JGI-IMG/MER) webserver (Markowitz, *et al.* 2012). The amino acid sequences for EHLA_0974 and EHLA_0978 were searched against the IMG database (Markowitz, *et al.* 2012). The BLASTp hits were limited to those with an E-value cut-off of 1e-10 and at least 60% identity. The amino acid sequences were downloaded in fasta format from the IMG website. The sequences were aligned in MEGA6 using ClustalW. The aligned amino acid sequences were used to reconstruct their phylogenetic relationship based on the Jones-Taylor-Thornton (JTT) substitution model.

Global proteomic profiling

Actively growing *A. soehngenii* pre-cultured in medium with 60mM glucose was inoculated (5%) in 450 ml of media as specified above to which 60 mM of filter sterilized glucose was added after autoclaving. The experiment consisted of biological duplicates. Samples were collected immediately after inoculation of bacteria (T0), early exponential phase at 16 h (T2), mid-exponential phase at 20 h (T3) and late exponential phase at 22 h (T4). A total of 50 ml was sampled at each time point. For comparison of proteomes in cultures grown on sorbitol, sucrose and lactate, *A. soehngenii* was grown anaerobically in triplicate in 450 ml volumes as described above in media containing 20 mM sucrose, 40 mM D-sorbitol or 80 mM DL-lactate, respectively. Cells were harvested at late-exponential phase. The samples were centrifuged at 4700 rpm for 30 min at 4°C. The cells were stored immediately in -80ºC until protein extraction. The cells were thawed on ice and washed with reduced phosphate buffer solution (PBS) twice before extraction of proteins. The reducing agent used for PBS was Titanium (III) citrate solution. The cell pellet was resuspended in SDT-Lysis buffer (Supplementary data), and cells were disrupted using a French pressure cell (1500 psi). For every sample, the cells were passed through the French pressure cell three times (cell disruption checked under light microscope). The samples were placed immediately on ice after cell disruption. The extracted proteins were denatured by heat treatment (95°C for 30 min). Protein quantification was done using the Qubit™ Protein Assay Kit (ThermoFisher Scientific, cat.no. Q33211). A total of 50 µg of proteins plus 5 µl of loading dye was

mixed and electrophoresed on 10% SDS gels (10% Mini-PROTEAN® TGX™ Precast Protein Gels) for 40 min at 20mA. Staining was done using Coomassie Brilliant Blue R250 (2.5 g/L) (in 45% methanol and 10% glacial acetic acid) for 3 h and destained with de-staining (25% methanol and 10% glacial acetic acid in ultrapure water) solution overnight.

The gels were treated for reduction and alkylation using 50 mM ammonium bicarbonate, 20 mM beta-mercaptoethanol (pH 8) and 20 mM acrylamide (pH 8). Each lane was cut into 3 even slices, which were further cut into small pieces of approximately 1–2 mm^2. Trypsin digestion was performed by adding 50 μL of sequencing grade trypsin (5 ng/μL in 50 mM ammonium bicarbonate) and incubated at room temperature overnight while shaking. The resulting tryptic peptide samples were desalted and concentrated using the μColumn (made with LichroprepC18) and subjected to nanoLC-MS/MS using a Proxeon Easy nanoLC and an LTQ-Orbitrap XL instrument.

Downstream data processing, analysis and visualisation

Raw data was initially processed using MaxQuant 1.5.2.854, and was further analysed using the DEP bioconductor/R package (Zhang*, et al.* 2018). All proteins that were detected in both duplicates of at least one condition were used for further analysis. Missing data was imputed using random draws from a Gaussian distribution centred around a minimal value to control for proteins missing not at random (MNAR). Mostly, these proteins were potentially missed in one condition as a consequence of their intensities being below the detection limit. These pre-processed data were then subjected to variance stabilization, and pair-wise differentially abundant proteins were identified using the limma R package (Ritchie*, et al.* 2015). Here, the tested contrasting conditions were early vs mid, early vs late and mid vs late exponential growth stages, and lactate vs sorbitol, lactate vs sucrose, and sorbitol vs sucrose conditions for *A. soehngenii*.

Results and Discussion

Genomic and proteomic reconstruction of the butyrogenic pathway in *A. soehngenii*

The differential proteome content of *A. soehngenii* L2-7 during growth on glucose was determined by analysing replicate samples taken after inoculation (T0), early exponential growth phase at 16 h (T2), mid-exponential growth phase at 20 hrs (T3) and late exponential growth at 22 hrs (T4) (Figure 1). Based on these proteomic data in combination with advanced annotation of the complete genome of *A. soehngenii* strain L2-7 (Shetty*, et al.* 2017c), the pathway from glucose to butyrate production was reconstructed. All the enzymes involved in the glycolytic pathway that were encoded in the genome were confirmed in the expressed proteome. Of particular interest was the constitutive and high production of each of the two phosphofructokinases (PFK) encoded by two copies of the *pfk* gene (EHLA_0314, and EHLA_3223). PFK is a key enzyme in the glycolysis, and its apparent duplication illustrates the importance of this step. Additionally, diphosphate-fructose-6-phosphate 1-phosphotransferase (PFP) encoded by *pfp* (EHLA_0586) was also present at high levels in the proteome. PFP uses diphosphate as the phosphate donor, while PFK utilizes ATP for conversion of fructose 6-phosphate to D-fructose 2,6-bisphosphate. Only single isomers of the enzymes involved in the subsequent steps in the glycolysis leading to pyruvate were observed in the proteome data although gene duplications were observed in the genome for phospholglycerate mutase (3 copies, EHLA_0716, EHLA_0912 (active), EHLA_1387), phosphypyruvate hydrates (2 copies, EHLA_0921, EHLA_1257 (active)) and pyruvate kinase (2 copies, EHLA_0785, EHLA_2893(active)). The generated pyruvate is then converted to acetyl-CoA by pyruvate:ferredoxin oxidoreductase core domain II (EHLA_0273). This enzyme carries out a reversible reaction and utilizes two oxidized ferrodoxin (Fd_{ox}) and one CoA per one pyruvate molecule to produce two reduced ferrodoxin (Fd_{rd}), one acetyl-CoA and CO_2 and two reducing equivalents (H^+). The observed levels of pyruvate:ferredoxin oxidoreductase core domain II were not significantly different between the growth phases. The recycling of Fd is most likely carried out by ferredoxin hydrogenase (EHL_0882), which was detected in the expressed proteome of *A. soehngenii*.
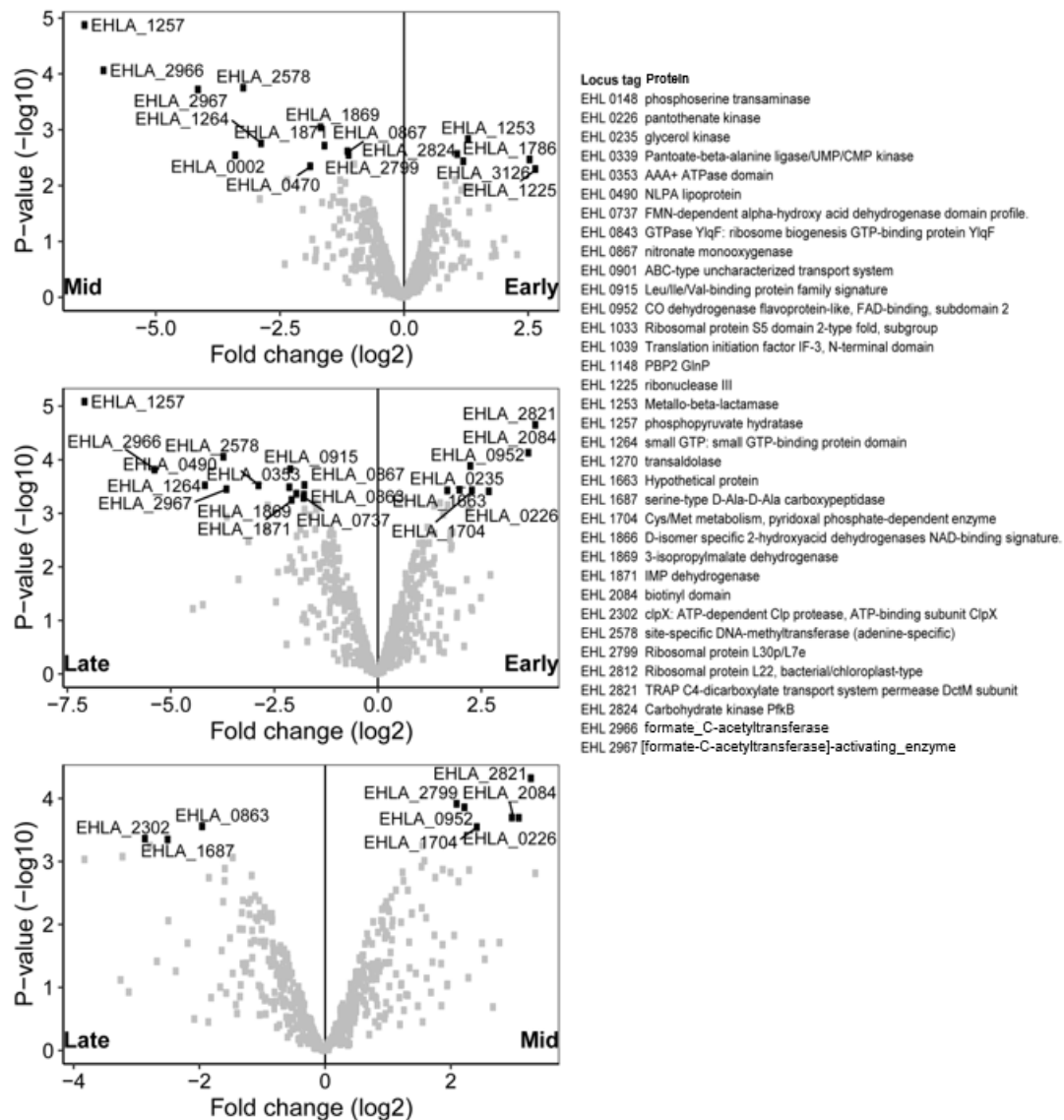
Figure 1: Volcano plot highlighting the proteins significantly up- and down-regulated in early, late and mid exponential growth stage for *A. soehngenii* when grown on glucose. The proteins for which levels were significantly different ($p < 0.05$) after correcting for false discovery rate are labelled.

Except for one, all genes encoding enzymes involved in the conversion from acetyl-CoA to butyrate were detected in the earlier annotated genome, and the corresponding gene products were found to be expressed in the proteome (Figure 2) (Shetty, *et al.* 2017c). The missing enzyme involved in last step i.e. butyryl-CoA to butyrate was found upon manual inspection of the annotated genome and investigation of its active domain as predicted by the InterPro database.
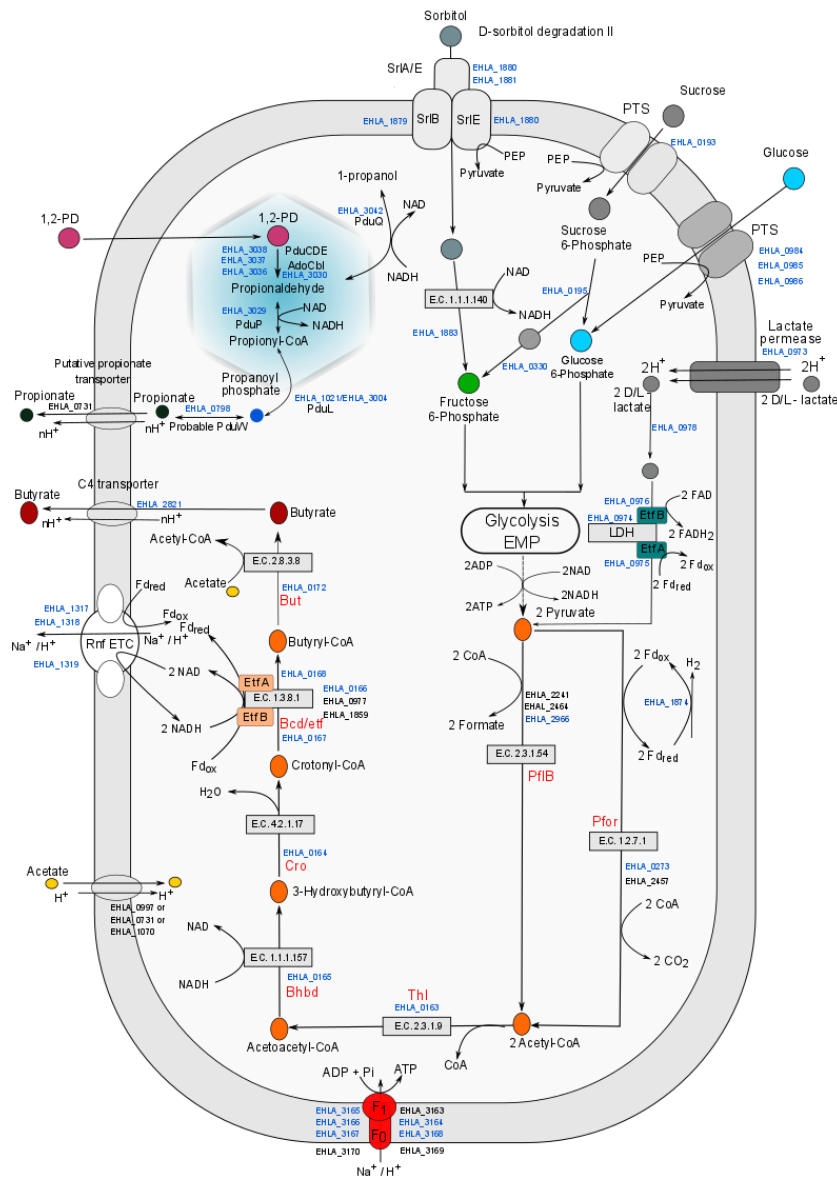
Figure 2: Overview of the key metabolic features of *A. soehngenii*. Relevant locus tags with the prefix EHLA_ are indicated and those in blue relate to gene products that are detected in the proteome, while those in black are only identified in the genome but not detected in the proteome at any of the growth conditions. SCFA transporters most likely belong to the C4 TRAP family of transporters. The symport is with a proton and *n* indicates the number of protons transported across the cell membrane which may vary under various growth conditions.

This key enzyme in butyrate production, butyryl CoA transferase, was identified as the protein encoded by the gene with locus tag EHLA_0172 and was found to be actively expressed and present in all proteome datasets.

Cobalamin biosynthesis and cobalamin-dependent glycerol/diol dehydratase

Cobalamin (Vitamin B12) is an important modulator of structure and composition of intestinal microbiota (Degnan*, et al.* 2014b). There is a high competition for cobalamin

among intestinal microbes, and bacteria that produce cobalamin may influence other members in their immediate surrounding. Recently, in a cross-feeding experiment with *A. soehngenii* and *Akkermansia muciniphila*, the latter was demonstrated to benefit from cobalamin produced by *A. soehngenii* (Belzer*, et al.* 2017). On the other hand, *A. soehngenii* used 1,2 PD produced by *A. muciniphila* for its growth and propionate production. To further elucidate the pathway for cobalamin biosynthesis, we investigated the genome sequence and the protein expression profile of *A. soehngenii* at different growth stages. Firstly, the genome was investigated for the presence of AdoCbl (adenosylcobalamin) and cobalamin riboswitches. An AdoCbl riboswitch was identified upstream of the gene encoding adenosylcobinamide amidohydrolase (EHLA_2015, CbiZ, 2116744bp to 2116884bp), which converts adenosylcobinamide (AdoCbi) to adenosylcobyric acid (AdoCby). The *de novo* B12 biosynthesis and salvage pathways were identified in the genome. L-glutamate is most likely the initial substrate for the *de novo* biosynthesis pathway in *A. soehngenii* (Figure S1). Proteome data confirmed that both pathways were active (Figure S1). The PocR regulatory protein (EHLA_3044) was also detected in the expressed proteome. The PocR regulator is known to modulate propanediol utilization (encoded by the *pdu* genes) and vitamin B12 biosynthesis in *Salmonella* and *Lactobacillus* species (Chen*, et al.* 1995, Santos*, et al.* 2011). Out of the 22 genes, the products of 15 of these were detected in the proteome in absence of 1,2 PD. Previously, analysis of the *Anaerobutyricum* genome did not identify a homologue to *pduL*, the product of which converts propionyl-CoA to propionyl-PO_4^{2-}, which suggested that there is another enzyme carrying out this reaction (Engels*, et al.* 2016b).
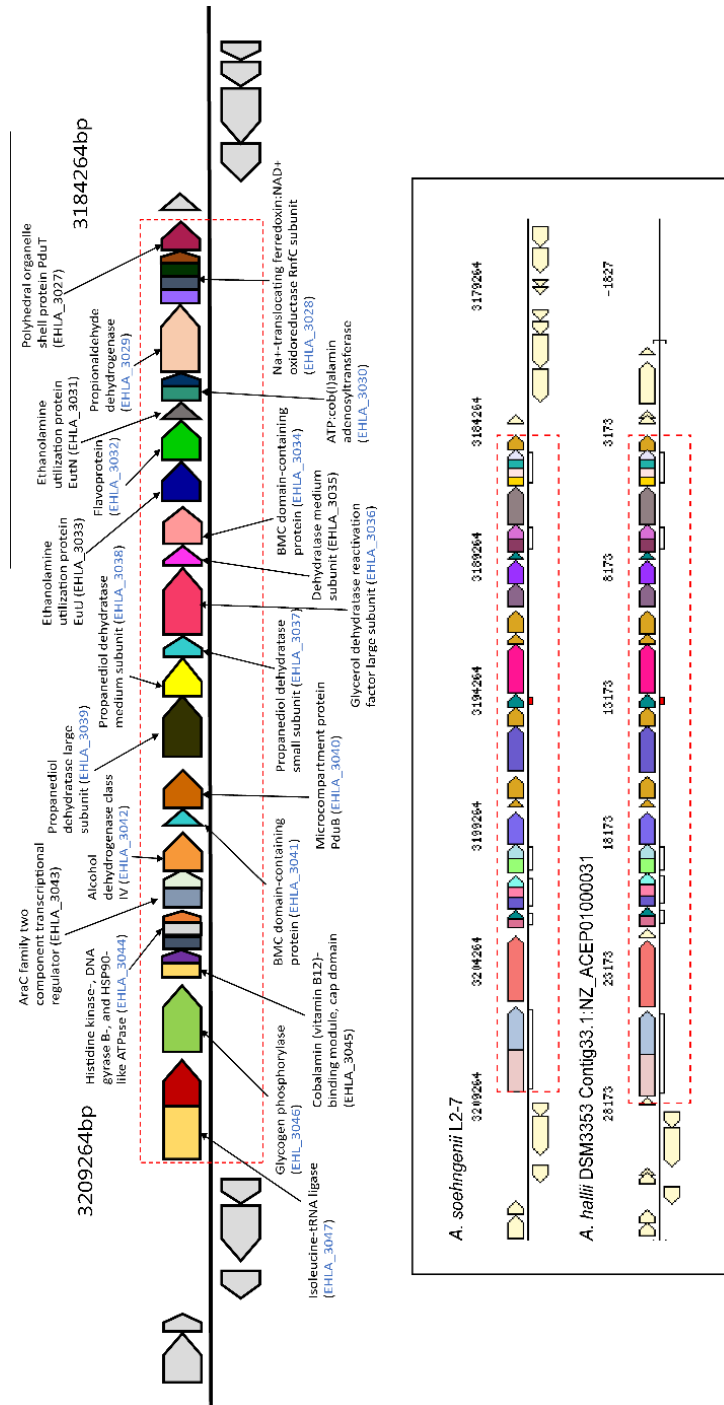
Figure 3: Organisation of 1,2 propanediol utilization genes in the *A. soehngenii* genome. The locus tags encoding proteins detected in the proteome are coloured in blue. The box depicts the organisation of 1, 2 PD utilization genes in the genomes of *A. soehngenii* and the closely related *A. hallii*.

Based on the reanalysis of the complete genome and proteome of *A. soehngenii*, we hypothesize that this enzyme could be the one encoded by the gene with the locus tag EHLA_3004 that is now annotated as a phosphate acetyltransferase. The gene for this phosphate acetyltransferase is located further upstream of the *pduCDE* gene cluster, and the protein was found in the proteome data in relatively low amounts. The limited

activity of PduL may explain the weak growth of *Anaerobutyricum* spp. in presence of 1,2 PD as the sole carbon source (Liu*, et al.* 2007).

Anaerobutyricum spp. readily convert 1,2 PD to propionate

In a recent study, *A. hallii* and *A. soehngenii* were reported to produce propanal, propanol and propionate when grown in presence of 1,2 PD (Engels*, et al.* 2016b). This study reported that in presence of 1,2 PD, neither *A. hallii* nor *A. soehngenii* utilized glucose. The 1,2 PD present in the intestinal tract derives from the metabolism of deoxy sugars, such as fucose and rhamnose, by other intestinal bacteria. Recently, 1,2 PD was reported as a cross-feeding metabolite between *Akkermansia muciniphila* (producer) and *A. soehngenii* (utilizer) (Belzer*, et al.* 2017). These findings raised an intriguing question regarding the substrate preference of these strains. Here, we tested the carbon source utilization in combinations *viz.* sucrose plus glycerol for *A. soehngenii*, lactose plus glycerol for *A. hallii* and 1,2 PD plus glycerol for both strains. Both *A. hallii* and *A. soehngenii* were able to utilize 1,2 PD before utilization of glycerol and produced propionate, propanal and propanol (Figure 4A and B). When *A. soehngenii* was grown in sucrose plus glycerol, the latter was the preferred substrate while sucrose was not utilized after 24 h (Figure 4D and F). Previously, it was demonstrated that 3-hydroxypropionaldehyde (3-HPA) was produced from glycerol by *A. soehngenii*, which inhibited its growth at low concentrations (7–8 mM 3-HPA) (Engels*, et al.* 2016b). This suggests that the end products of glycerol fermentation have bacteriostatic effect on *Anaerobutyricum* spp. (Engels*, et al.* 2016b)*.* Similarly, the growth of *A. hallii* was inhibited in media with glycerol plus lactose after the utilisation of glycerol while lactose was not utilized anymore (Figure 4C and E). However, when 1,2 PD and glycerol were present, both substrates were utilized. Since, 1,2 PD was observed to be a preferred substrate *in vitro* for *Anaerobutyricum*, we were interested in knowing whether the ability to utilize 1,2 PD is wide spread in commensal bacteria isolated from the human intestinal tract. For this, we searched 677 bacterial and archaeal genomes annotated in the IMG database for cobalamin-dependent glycerol/diol dehydratase (PduCDE). The selection included bacteria that are part of the human intestinal phylogenetic core and related bacteria isolated from human intestine (Atashgahi*, et al.* 2018, Shetty*, et al.* 2017a).
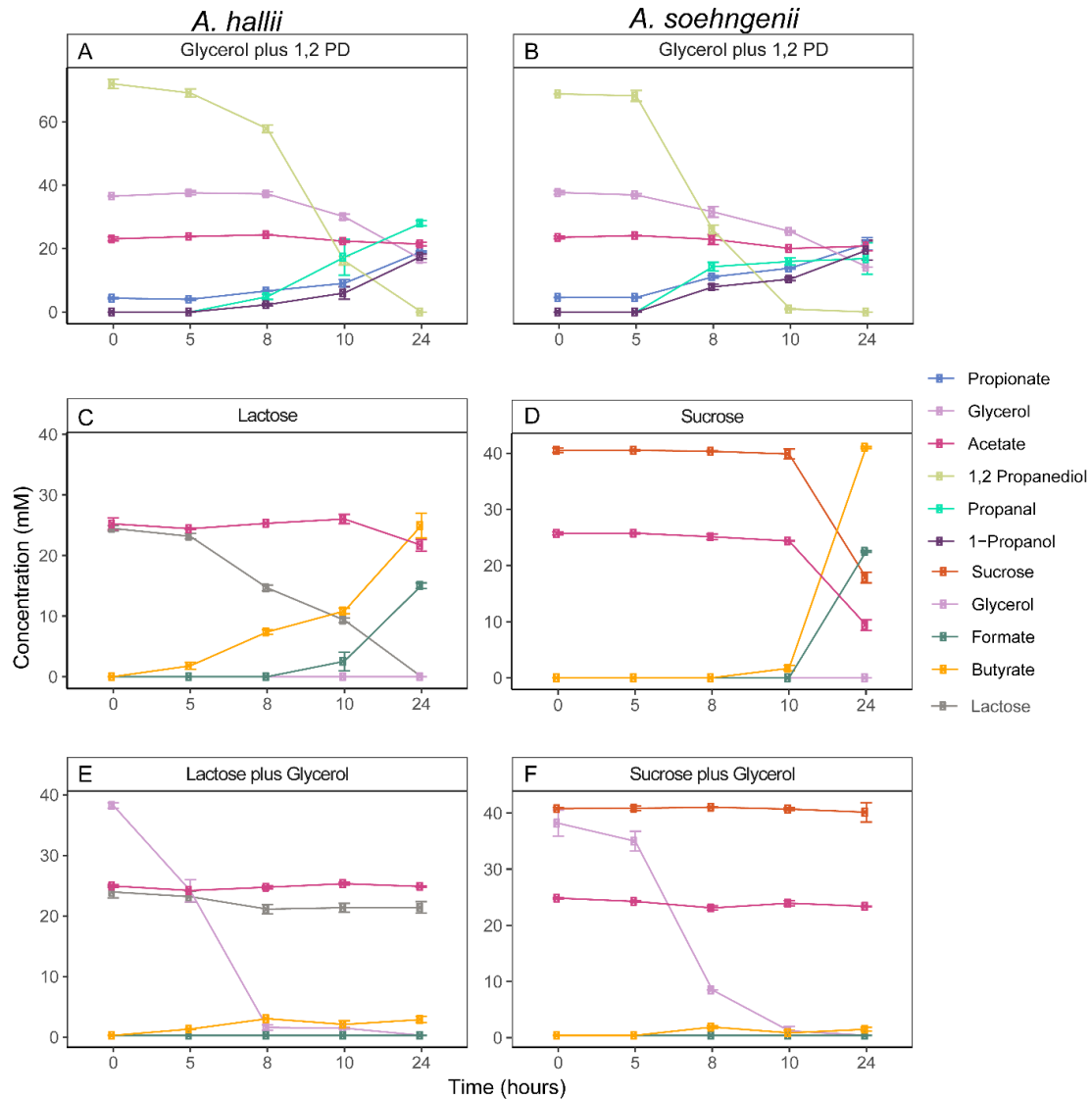
Figure 4: Metabolic end products of carbon metabolism. **A]** *A. hallii* grown in presence of glycerol and 1,2 PD; **B]** *A. soehngenii* grown in presence of glycerol and 1,2 PD; **C]** *A. hallii* grown in presence of lactose; **D]** *A. soehngenii* grown in presence of sucrose; **E]** *A. hallii* grown in presence of lactose plus glycerol; **F]** *A. soehngenii* grown in presence of sucrose plus glycerol.

The complete list of bacteria that have *pduCDE* genes is summarized in Table S1. These genes were found to be common in bacteria from the Enterobacteriaceae family, such as *Klebsiella*, *Citrobacter* and *Listeria*, most of which are known human pathogens. In contrast, only few obligately anaerobic commensal bacteria have these genes, such *A. hallii*, *Blautia obeum,* and a not yet characterized *Lachnospiraceae* bacterium sp. 7_1_58FAA, which all belong to the phylogenetic core (Shetty, *et al.* 2017a). Thus, utilization of 1,2 PD could be a specialized niche in the human intestinal microbiome,

and some of the obligately anaerobic core bacteria can compete for 1,2 PD under anoxic conditions with bacteria from the Enterobacteriaceae family.

Differential proteomes of *A. soehngenii* during growth on sorbitol and sucrose

Sugar alcohols are commonly used as replacement of glucose for controlling the blood glucose levels in humans. One such common sugar alcohol is sorbitol, which is metabolized slowly by the human body and hence can be available as a carbon and energy source for bacteria in the large intestine (Deis and Kearsley 2012). Furthermore, the ability to utilize sucrose is one of the differentiating features for *A. soehngenii* compared to *A. hallii.* Therefore, we sought to identify the active proteins in the metabolic pathway that leads to conversion of sorbitol and sucrose to butyrate.

In total, we identified 67 proteins that were significantly different between sorbitol and sucrose growth conditions. Among these, 27 were overexpressed and 40 were less expressed when cells were grown on sorbitol (Figure 5 A). Further inspection of the expressed proteome revealed that *A. soehngenii* has an inducible operon for sorbitol uptake. The genes encoding the phosphotransferase system (PTS) for sorbitol are located between the glucitol/sorbitol operon activator (*gutM*) and the lichenan operon transcriptional antiterminator (Figure S1A). The significantly higher (> 7-fold) abundance of these proteins during growth on sorbitol when compared to that on sucrose suggests that this gene cluster is co-ordinately expressed and highly induced in presence of sorbitol. The promoter region is located before the gene EHLA_1883 that encodes sorbitol-6-phosphate dehydrogenase, while the terminator is located immediately after the transcriptional terminator EHLA_1878. Comparison of the expressed proteome observed in cells grown on sucrose and sorbitol revealed that the product of EHLA_0193 annotated as a Npi-phosphohistidine-sugar phosphotransferase (PTS) was present with more than 8-fold higher abundance in the presence of sucrose versus glucose. This transporter is most likely a sucrose PTS that shares similarity with maltose-PTS . Upstream of the gene encoding the sucrose PTS is a LacI-type HTH domain (EHLA_0192), which was not detected in the proteome data, but is likely a transcriptional regulator controlling the expression of the sucrose-PTS gene.
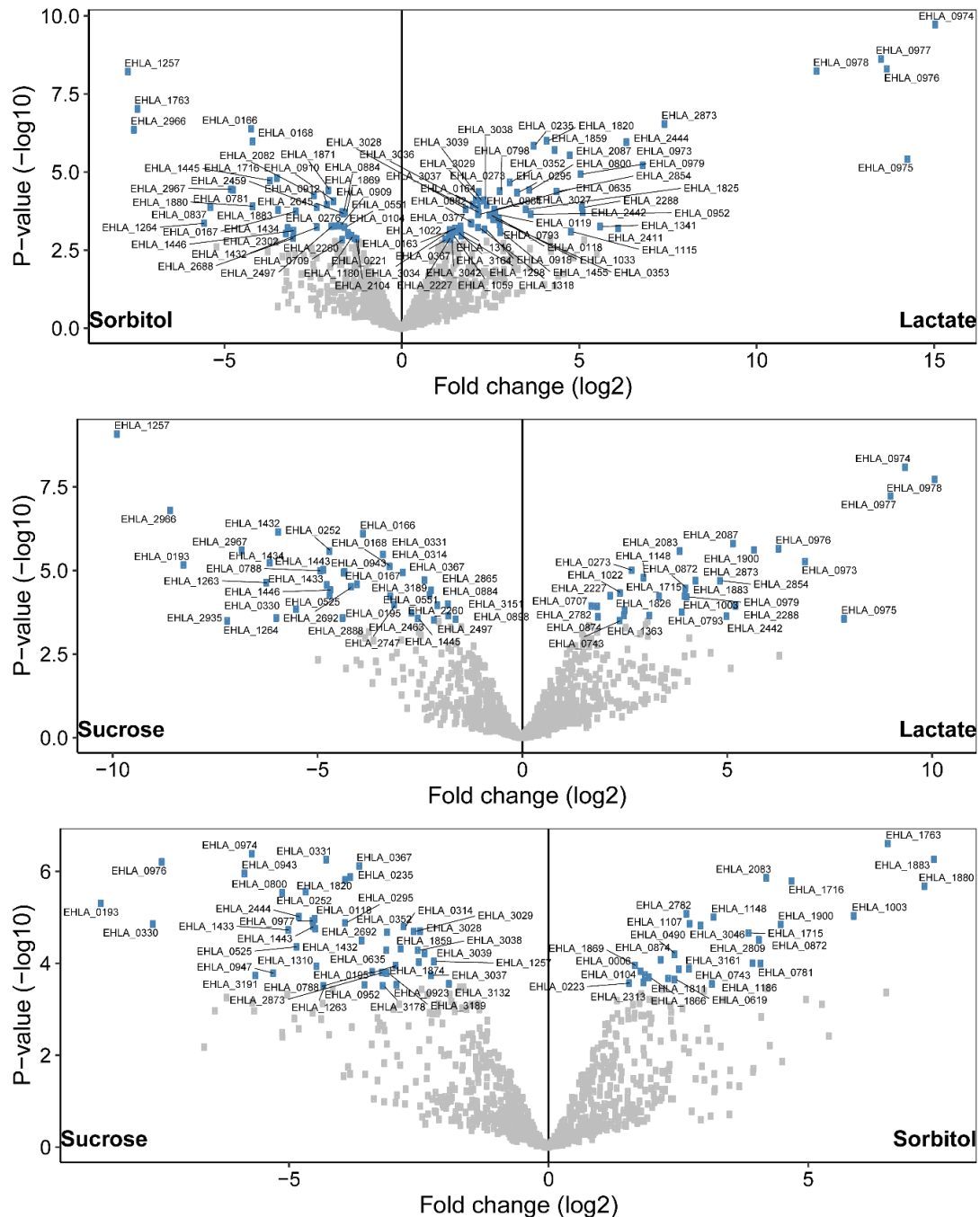
Figure 5: Differentially abundant *A. soehngenii* proteins in cells grown on sorbitol, sucrose and lactate. **A]** Comparison of proteomes when grown on sucrose or sorbitol, **B]** Comparison of proteomes when grown on sucrose or D, L-lactate, **C]** Comparison of proteomes when grown on sorbitol or D, L-lactate. The locus tags that are significantly different (P < 0.01) in abundance between the two conditions in the proteome are coloured in blue. For clarity, only the locus tags are labelled, and a complete list with product names for locus tags is given in Table S2.

Downstream of the gene coding for the sucrose PTS is a gene encoding maltose 6'-phosphate phosphatase (EHLA_0194) and a gene encoding 4-alpha-glucanotransferase (EHLA_0195). The latter belongs to glycoside hydrolase family 77 (GH77), members of

which breaks the alpha-1,2 glycosidic bond between glucose and fructose. Interestingly, lactate dehydrogenase (EHLA_0974) was 5-fold more abundant in cells grown on sucrose compared to sorbitol, which could be a result of an excess of fructose-1,6- bisphosphate.

Similarly, levels of formate-acetyltransferase (EHL_2967) were significantly higher ($P < 0.05$) when cells were grown on sucrose compared to lactate and 2-fold higher ($P > 0.05$) in the presence of sucrose as compared to sorbitol. Next, we did a BLASTp search for the active proteins involved in sucrose utilization in the publically available genome of *A. hallii*. The following genes encoded by *A. soehngenii*, EHLA_0192 and EHLA_0193, were not identified in *A. hallii*, while a homologue of EHLA_0195 was identified, which shared 52% identity with the gene encoding 4-alpha-glucanotransferase in *A. hallii*. Additionally, the gene encoding maltose 6'-phosphate phosphatase (EHLA_0194) did not have a homologue in the genome of *A. hallii*.

Proteomics guided identification of the active DL-lactate utilization gene cluster

Another important metabolite that is available as a growth substrate for intestinal bacteria is lactate (Duncan, *et al.* 2004). Several intestinal bacteria such as those belonging to the genera *Bifidobacterium* and *Lactobacillus* are known lactate producers. Lactate is one of the major cross-feeding metabolites that is converted to butyrate by butyrate producing bacteria such as *A. soehngenii* (Duncan, *et al.* 2004). Till date, only bacteria related to *Anaerobutytricum hallii* and *Anaerostipes* have been shown to utilize both D and L- forms of lactate to produce butyrate (Duncan, *et al.* 2004). However, to the best of our knowledge, the active genes involved in butyrogenesis from DL-lactate and acetate have not been investigated in detail for members of *Anaerobutyricum* and related genera.

During growth on lactate plus acetate, lactate permease (EHLA_0973), lactate dehydrogenase (EHLA_0974), electron transfer flavoprotein beta subunit (EHLA_0975), electron transfer flavoprotein (EFT) alpha subunit (EHLA_0976), short-chain acyl-CoA dehydrogenase (EHLA_0977), and lactate racemase (EHLA_0978) showed significantly high abundance (Figure 5B and C). Remarkably, lactate dehydrogenase (EHLA_0974) is co-occurring with EtfA and EtfB, which suggests the presence of a D-lactate dehydrogenase/electron-transferring flavoprotein (Etf) (LDH/Etf) complex allowing

growth on lactate, which is a low energy substrate (Weghoff, *et al.* 2015). Interestingly, a gene coding for a short-chain acyl-CoA dehydrogenase, orthologous to that coding for butyryl–CoA dehydrogenase, was identified to be located downstream of the gene encoding lactate dehydrogenase, and corresponding gene products were found produced together with other lactate-specific proteins. Consistent with a previous study (Duncan, *et al.* 2004), formate was not detected when *A. soehngenii* was grown on lactate, and formate-acetyltransferase (pyruvate formate lyase) was present at significantly lower levels (-6.16 fold) in cells grown on lactate as compared to sorbitol (Table 1). In addition, when the proteome of cells grown on glucose and harvested at different growth phases was compared, the enzymes formate-C-acetyltransferase-activating and formate-acetyltransferase were found to be significantly less abundant in mid (-4 and -6-fold, respectively) and late exponential (-3.5 and -5.3-fold, respectively) phase compared to early exponential growth phase (Figure 1A and B). These observations support the previous hypothesis concerning the activation of a formate-acetyltransferase in the presence of glucose only (Duncan, *et al.* 2004).

Table 1: Substrate utilisation and fermentation end products for *A. soehngenii*.
*ND: not detected.

C-source	Consumption (mM)		Production (mM)	
	Substrate	Acetate	Butyrate	Formate
DL-Lactate	48.7 ± 0.2	18.8 ± 0.3	24.6 ± 0.7	ND*
Sorbitol	26.4 ± 3.9	10.6 ± 2.7	22.8 ± 1.0	5.4 ± 1.9
Sucrose	14.8 ± 0.3	7.0 ± 1.2	20.0 ± 1.2	17.6 ± 1.2

Comparison of genomes of *Anaerobutyricum* and related species

Next, we searched for homologues of genes encoding lactate dehydrogenase (EHLA_0974), lactate permease (EHLA_0973) and lactate racemase (EHLA_0978) in the genomes of bacteria closely related to *Anaerobutyricum* (Figure 6). *A. hallii* and *Anaerostipes caccae* have homologues of genes coding for lactate dehydrogenase, lactate racemase and lactate permease.
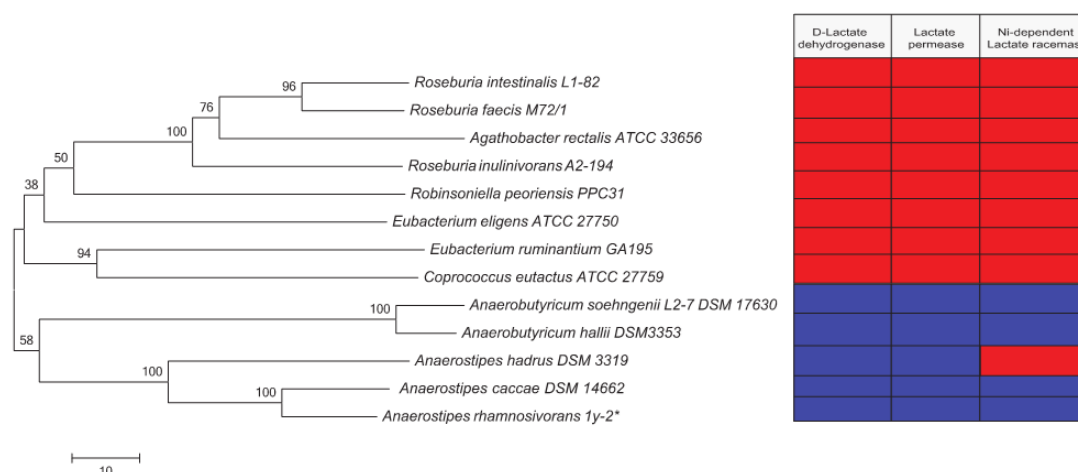
Figure 6: Phylogenetic tree based on the 16S rRNA gene sequences for *A. soehngenii* and closely related species. The evolutionary history was inferred using the Neighbour-Joining method based on 1000 bootstrap replicates. The amino acid sequences for D-lactate dehydrogenase (EHLA_0974), lactate permease (EHLA_0973) and nickel dependent lactate racemase (EHLA_0978) were searched against the genomes of closely related species. Presence in genome is indicated by blue coloured cells, absence by red coloured cells. *The genome sequence of *Anaerostipes rhamnosivorans* strain 1y-2 was kindly shared by Dr. TP Nam Bui (Bui 2016).

However, the genome of *Anaerostipes hadrus* lacked the gene encoding lactate racemase, which is consistent with its inability to utilize L-lactate (Duncan*, et al.* 2004). We further expanded our search to other bacterial genomes present in the IMG database (Markowitz*, et al.* 2012).

The lactate racemase encoded in the genome of *Anaerobutyricum* is predicted to be a Ni-dependent enzyme (Desguin*, et al.* 2014). A BLASTp search for amino acid sequences similar to EHLA_0978 revealed their presence in genomes of members of genera *Intestinimonas* and *Flavonifractor*, as well as species *Eubacterium barkeri*, *E. aggregans*, *Pseudoramibacter alactolyticus*, *Clostridium merdae*, *Clostridium jeddahense* and related species*.* A phylogenetic analysis suggests that lactate racemase (EHLA_0978) in *A. soehngenii* shares similarity to the lactate racemases present in related *E. barkeri*, *E. aggregans*, *Intestinimonas* spp. and *Flavonifractor* (Figure 7). *E. aggregans* (isolated form methanogenic bioreactors) and *Intestinimonas* spp. are known for their ability to convert lactate plus acetate to butyrate, however, the growth is weak (Kläring*, et al.* 2013, Tahar Mechichi and Woo 1998). *Flavonifractor* species are closely related to *Intestinimonas* and can potentially utilize lactate and acetate to

169

produce butyrate. Interestingly, the lactate racemase identified in *Anaerostipes* has low identity to EHLA_0978 (e-value: 4e-158, and identity:56%).

A BLASTp search for amino acid sequences of lactate dehydrogenase (EHLA_0974) revealed a wide spread distribution of homologues in this genome database (Supplementary Figure S3). As with the lactate racemase, the phylogenetically related (based on 16S rRNA gene) *Anaerobutyricum* and *Anaerostipes* did not share high identity with each other with respect to lactate dehydrogenase encoding genes, and were phylogenetically placed in two distinct clusters.
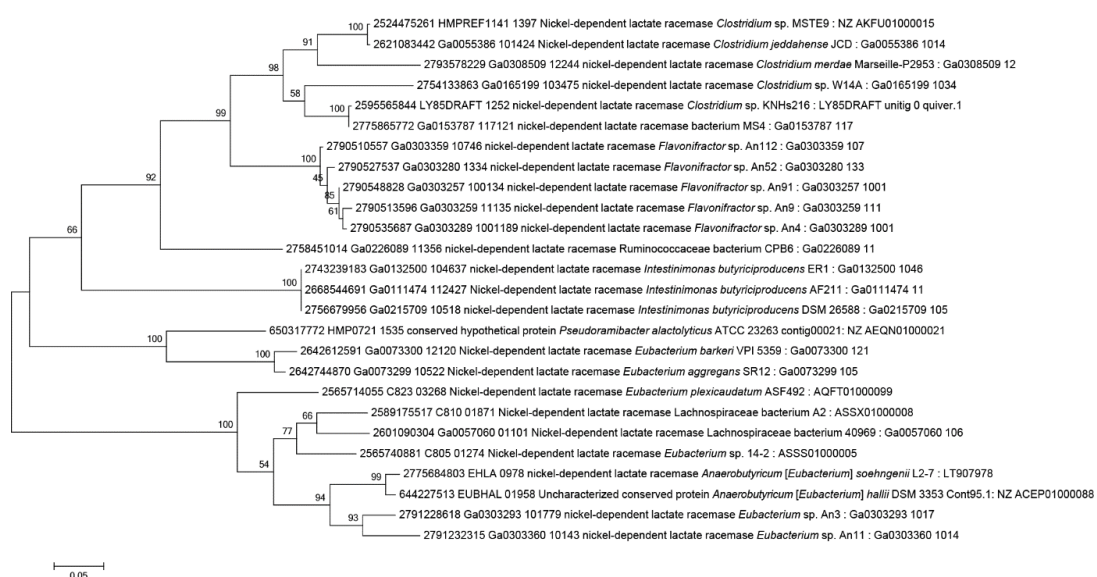
Figure 7: Maximum likelihood phylogenetic tree of 26 lactate racemase homologues. Phylogenetic tree of 26 lactate racemase homologs identified by searching against 55,499 isolate genomes of the IMG/ER database (as of 2 October 2018). Labels represent the IMG gene id, locus tag, IMG annotation for the gene product, taxonomic identity, strain name, and assembly and/or contig/scaffold and NCBI/DDJB/ENA accession number of genomes which contains this gene. Where several genomes for a species were present, the branch of the tree was collapsed for clarity. The numbers on branches represent the bootstrap values from 1000 replicates. See methods for details on calculation.

However, further analysis of neighbourhood genes indicated that both the *Anaerobutyricum* and *Anaerostipes* lactate dehydrogenase gene is followed by genes coding for Etf alpha and beta subunits (Supplementary Figure S4). Overall, these observations provide evidence for unique adaptation of *A. soehngenii, A. hallii* and *Anaerostipes caccae* to utilize an important low energy cross-feeding metabolite, lactate.

Conclusions

The genomic and proteomic data presented in this study allowed elucidating key metabolic features of *A. soehngenii*. We confirmed the active expression of the gene cluster previously predicted to catalyze the conversion of 1,2 PD to propionate, and demonstrated its expression in absence of the substrate 1,2 PD. The products of gene cluster *pduCDE* were detected in the proteome of *A. soehngenii* irrespective of the carbon source used in the growth medium. This activity is linked to the ability to *A. soehngenii* to synthesize vitamin B12 as it has a B12-dependent diol hydratase. Additionally, we identified an inducible operon encoding a sorbitol transporter, which was abundant in the expressed proteome of *A. soehngenii* when grown in presence of sorbitol. The ability to utilize sucrose is conferred by two proteins that are not encoded in the genome of the phylogenetically related *A. hallii*, *viz.* a sucrose PTS and LacI-type HTH domain, which both were highly induced in presence of sucrose. The specialized ability of *A. soehngenii* to convert DL-lactate and acetate to butyrate is conferred by the unique genomic organisation of a lactate utilization gene cluster (Thauer*, et al.* 1977, Weghoff*, et al.* 2015). Genome analysis demonstrated that *Anaerobutyricum hallii, A. soehngenii* and *Anaerostipes caccae* have well adapted lactate utilization gene clusters that encode all necessary proteins (transcriptional regulator, transporter, racemase (two copies), LDH/Etf complex, a homolog of butyryl-CoA dehydrogenase) that were also detected in high amounts in the expressed proteome data. In conclusion, data presented here considerably improve our understanding of the metabolic lifestyle of *Anaerobutyricum soehngenii*.

Acknowledgements

Supplementary information

Table S1: Bacteria which encode the *pduCDE* gene cluster in their genome.

Genome	*pduC* (KO:K01699)	*pduD* (KO:K13919)	*pduE* (KO:K13920)
Acetomicrobium hydrogeniformans ATCC BAA-1850 (647000209)	2	2	1
Listeria monocytogenes FSL J2-071 (641736262)	2	0	1
Aneurinibacillus aneurinilyticus ATCC 12856 (2563366762)	1	0	0
Bacillus massiliosenegalensis JC6 (2547132161)	1	1	1
Blautia obeum ATCC 29174 (640963024)	1	1	1
Citrobacter amalonaticus FDAARGOS_165 (2718218027)	1	1	1
Citrobacter freundii 4_7_47CFAA (2513237266)	1	1	1
Citrobacter koseri ATCC BAA-895 (640753015)	1	1	1
Citrobacter sp. 30_2 (646206256)	1	1	1
Citrobacter sp. FDAARGOS_155 (2718218443)	1	1	1
Citrobacter youngae ATCC 29220 (2562617093)	1	1	1
Clostridium methylpentosum R2, DSM 5476 (643886206)	1	1	1
Clostridium orbiscindens 1_3_50AFAA (2597489878)	1	1	1
Enterococcus avium ATCC 14025 (2541047449)	1	1	1
Enterococcus raffinosus cftri2200 (2554235108)	1	1	1
Escherichia albertii GTC 14781 (2602041909)	1	1	1
Escherichia albertii TW07627 (641736163)	1	1	1
Escherichia coli LF82 (651053021)	1	1	1
Escherichia coli MS 115-1 (648276650)	1	1	1
Escherichia coli sv. O139:H28 E24377A (ETEC) (640753024)	1	1	1
Escherichia fergusonii UMN026, ATCC 35469 (643692022)	1	1	1
Anaerobutyricum [Eubacterium] hallii DSM 3353 (643886203)	1	1	1
Anaerobutyricum [Eubacterium] soehngenii L2-7 (2775506742)	1	1	1
Flavonifractor plautii ATCC 29863 (2513237287)	1	1	1
Fusobacterium nucleatum animalis 7_1 (646206252)	1	1	1
Fusobacterium nucleatum subsp. *animalis* 11_3_2 (651324032)	1	1	1
Fusobacterium nucleatum subsp. *animalis* D11 (645058833)	1	1	1
Fusobacterium periodonticum 1_1_41FAA (647533168)	1	1	1
Fusobacterium varium ATCC 27725 (646206275)	1	1	1
Hafnia alvei ATCC 51873 (2513237377)	1	1	1
Hafnia alvei FDAARGOS_158 (2721755583)	1	1	1

Intestinimonas butyriciproducens AF211 (2667527550)	1	1	1
Intestinimonas butyriciproducens ER1 (2740892378)	1	1	1
Klebsiella pneumoniae RYC492 (2565956817)	1	1	1
Klebsiella pneumoniae TH1 (2721755650)	1	1	1
Klebsiella sp. 1_1_55 (647533174)	1	1	1
Klebsiella sp. 4_1_44FAA (2513237265)	1	1	1
Klebsiella sp. MS 92-3 (651324045)	1	1	1
Lachnospiraceae bacterium sp. 7_1_58FAA (2513237329)	1	1	1
Lactobacillus brevis ATCC 14869 (2597490301)	1	1	1
Lacto*bacillus coryniformis coryniformis* KCTC 3167 (651324058)	1	1	1
Lactobacillus coryniformis torquens 30, KCTC 3535 (651324059)	1	1	1
Lactobacillus reuteri CF48-3A (2502171173)	1	1	1
Lactobacillus reuteri DSM 20016 (2671180761)	1	1	1
Lactobacillus reuteri DSM 20016 (640427118)	1	1	1
Lactobacillus reuteri JCM 1112 (642555135)	1	1	1
Lactobacillus reuteri SD2112, ATCC 55730 (650716048)	1	1	1
Listeria innocua ATCC 33091 (2537561540)	1	1	1
Listeria monocytogenes 4b H7858 (638341118)	1	1	1
Listeria monocytogenes Finland 1988 (2511231203)	1	1	1
Listeria monocytogenes FSL N1-017 (640963059)	1	1	1
Listeria monocytogenes FSL R2-561 (2511231184)	1	1	1
Listeria monocytogenes sv. 1/2a 10403S (2511231174)	1	1	1
Listeria monocytogenes sv. 1/2a J0161 (2511231190)	1	1	1
Listeria monocytogenes sv. 1/2a1 FSL N3-165 (640963060)	1	1	1
Listeria monocytogenes sv. 1/2b FSL J2-064 (641736263)	1	1	1
Listeria monocytogenes sv. 4a HCC23 (643348561)	1	1	1
Listeria monocytogenes sv. 4a L99 (650377957)	1	1	1
Listeria monocytogenes sv. 4b F2365 (637000154)	1	1	1
Listeria monocytogenes sv. 4b HPB2262 (640963047)	1	1	1
Metakosakonia massiliensis JC163 (2547132115)	1	1	1
Pediococcus acidilactici 7_4 (647533194)	1	1	1
Salmonella enterica enterica sv. *Typhi* ST0208 (2548877057)	1	1	1
Salmonella enterica enterica sv. *Typhimurium* ST1660/06 (2548876846)	1	1	1
Shigella sonnei 19904011_1361399 (2698536368)	1	1	1
Veillonella sp. 6_1_27 (647533242)	1	1	1
Yersinia bercovieri ATCC 43970 (638341238)	1	1	1

Yersinia enterocolitica palearctica sv. O:3 bt. 4 Y11 (651053079)	1	1	1

Table S2: Significantly different proteins identified for each pairwise comparison.

Locus tag-gene product	Lactate vs Sorbitol significant	Lactate vs Sucrose significant	Sorbitol vs Sucrose significant	Lactate vs Sorbitol ratio	Lactate vs Sucrose ratio	Sorbitol vs Sucrose ratio
EHLA_0006 DNA topoisomerase (ATP-hydrolysing)	No	No	Yes	-0.68	1.12	1.8
EHLA_0104 Serine-tRNA ligase	No	No	Yes	-1.48	0.405	1.89
EHLA_0118 Ribose/Galactose Isomerase	No	No	Yes	3.01	-1.71	-4.71
EHLA_0119 Carbohydrate kinase, FGGY, C-terminal	Yes	No	No	2.68	0.577	-2.1
EHLA_0163 acetyl-CoA C-acetyltransferase	Yes	No	No	1.54	1.49	-0.0473
EHLA_0164 enoyl-CoA hydratase	Yes	No	No	2.04	1.33	-0.703
EHLA_0166 acyl-CoA dehydrogenase	Yes	Yes	No	-4.2	-3.86	0.343
EHLA_0167 Electron transfer flavoprotein, alpha/beta-subunit, N-terminal	Yes	Yes	No	-2.95	-4.33	-1.38
EHLA_0168 Electron transfer flavoprotein, alpha/beta-subunit, N-terminal	Yes	Yes	No	-4.17	-3.21	0.954
EHLA_0193 protein-Npi-phosphohistidine-sugar phosphotransferase	No	Yes	Yes	0.282	-8.24	-8.52
EHLA_0195 4-alpha-glucanotransferase	No	Yes	Yes	0.125	-3.09	-3.21
EHLA_0223 ornithine carbamoyltransferase	No	No	Yes	-1.18	0.399	1.58
EHLA_0235 glycerol kinase	Yes	No	Yes	3.75	-0.0391	-3.79

EHLA_0252 BMC domain	No	Yes	Yes	0.273	-5.7	-5.97
EHLA_0273 Pyruvate:ferredoxin oxidoreductase core domain II	Yes	Yes	No	2.2	2.7	0.494
EHLA_0276 Glutamate synthase, alpha subunit, C-terminal	Yes	No	No	-1.87	-2.01	-0.14
EHLA_0314 6-phosphofructokinase	No	Yes	Yes	-0.12	-2.89	-2.77
EHLA_0330 1-PFK: hexose kinase, 1-phosphofructokinase family	Yes	Yes	Yes	3.22	-5.49	-8.71
EHLA_0331 protein-Npi-phosphohistidine-sugar phosphotransferase	No	Yes	Yes	0.888	-3.37	-4.26
EHLA_0352 Bacterial extracellular solute-binding protein	Yes	No	Yes	3.07	-0.00806	-3.08
EHLA_0353 AAA+ ATPase domain	Yes	No	No	2.62	0.493	-2.12
EHLA_0367 starch synthase (glycosyl-transferring)/glycogen(starch) synthase	No	Yes	Yes	1.26	-2.36	-3.62
EHLA_0377 Diol dehydratase-reactivating factor alpha subunit	Yes	No	No	2.01	1.04	-0.965
EHLA_0490 NLPA lipoprotein	No	No	Yes	-0.939	1.51	2.45
EHLA_0525 shikimate kinase	No	Yes	Yes	0.75	-4.83	-5.58
EHLA_0529 Protein of unknown function DUF1700	No	No	Yes	2.9	-3.36	-6.25
EHLA_0551 Pyridoxal 5-phosphate synthase subunit PdxS [pdxS].	Yes	Yes	No	-1.58	-2.23	-0.655

EHLA_0619 DNA-directed RNA polymerase	No	No	Yes	-1.3	1.15	2.45
EHLA_0635 adenosinetriphosphatase	Yes	No	Yes	3.19	-0.453	-3.64
EHLA_0707 Iron only hydrogenase large subunit, C-terminal domain	No	Yes	No	1.06	1.71	0.656
EHLA_0743 Endopeptidase La	No	Yes	Yes	-0.327	3.09	3.42
EHLA_0781 Glutamate-ammonia ligase	Yes	No	Yes	-3.06	0.713	3.77
EHLA_0788 P-type cation-transporting ATPase superfamily signature	No	Yes	Yes	-1.53	-4.9	-3.37
EHLA_0793 3-isopropylmalate dehydratase	Yes	Yes	No	3.56	3.3	-0.264
EHLA_0798 acetate kinase	Yes	No	No	2.8	1.72	-1.08
EHLA_0800 Patatin-like phospholipase domain	Yes	No	Yes	3.77	-1.48	-5.26
EHLA_0837 30S ribosomal protein S16 [rpsP].	Yes	No	No	-5.36	-3.15	2.21
EHLA_0874 acetyl-CoA carboxylase	No	Yes	Yes	-0.31	1.87	2.18
EHLA_0881 NADH:ubiquinone oxidoreductase, 51kDa subunit, conserved site	Yes	No	No	2.42	1.85	-0.569
EHLA_0882 ferredoxin hydrogenase	Yes	No	No	1.99	1.36	-0.629
EHLA_0884 Solute-binding protein family 3, conserved site	Yes	Yes	No	-1.63	-1.78	-0.15
EHLA_0898 glutamate dehydrogenase (NADP+)	No	Yes	No	-1.03	-1.59	-0.563
EHLA_0903 Regulator of K+ conductance, C-terminal	No	Yes	No	2.38	3.89	1.51

EHLA_0909 Glyceraldehyde-3-phosphate dehydrogenase signature	Yes	No	No	-1.72	-1.73	-0.00474
EHLA_0910 phosphoglycerate kinase	Yes	No	No	-1.89	-1.46	0.427
EHLA_0912 phosphoglycerate mutase	Yes	No	No	-2.08	-1.21	0.871
EHLA_0918 adenosinetriphosphatase	Yes	No	No	2.37	1.21	-1.16
EHLA_0943 Hypothetical protein	No	Yes	Yes	0.24	-5.27	-5.51
EHLA_0947 preQ1 synthase	No	No	Yes	2.62	-2.15	-4.77
EHLA_0973 lctP: transporter, lactate permease (LctP) family	Yes	Yes	No	7.64	6.36	-1.28
EHLA_0974 CO dehydrogenase flavoprotein-like, FAD-binding, subdomain 2	Yes	Yes	Yes	15.1	9.38	-5.69
EHLA_0975 Electron transfer flavoprotein, alpha/beta-subunit, N-terminal	Yes	Yes	No	14.1	7.88	-6.26
EHLA_0976 Electron transfer flavoprotein, alpha/beta-subunit, N-terminal	Yes	Yes	Yes	13.7	6.28	-7.42
EHLA_0977 short-chain acyl-CoA dehydrogenase	Yes	Yes	Yes	13.8	9.02	-4.78
EHLA_0978 Domain of unknown function DUF2088	Yes	Yes	No	12.1	10.1	-1.98
EHLA_0979 GntR bacterial regulatory protein HTH signature	Yes	Yes	No	4.62	4.29	-0.333
EHLA_1003 Peptidylprolyl isomerase	No	No	Yes	-1.98	3.02	5

EHLA_1022 glycerol dehydrogenase	Yes	Yes	No	1.69	2.41	0.725
EHLA_1033 Ribosomal protein S5 domain 2-type fold, subgroup	Yes	No	No	2.51	1.58	-0.928
EHLA_1107 Initiation factor 2 signature.	No	No	Yes	-1.07	1.67	2.74
EHLA_1115 NADPH-dependent FMN reductase	Yes	No	No	4.7	0.287	-4.41
EHLA_1148 PBP2 GlnP	No	Yes	Yes	-0.21	2.99	3.2
EHLA_1186 Sporulation protein YtfJ (Spore YtfJ)	No	No	Yes	-0.0829	3.09	3.17
EHLA_1257 phosphopyruvate hydratase	Yes	Yes	Yes	-7.68	-9.86	-2.18
EHLA_1263 FeoA domain	No	Yes	No	-1.97	-6.28	-4.31
EHLA_1264 small GTP: small GTP-binding protein domain	Yes	Yes	No	-5.51	-5.95	-0.439
EHLA_1316 Respiratory-chain NADH dehydrogenase 51 Kd subunit	Yes	No	No	1.7	0.801	-0.901
EHLA_1318 rnfG: electron transport complex, RnfABCDGE type, G subunit	Yes	No	No	2.18	1.85	-0.329
EHLA_1341 PurE domain	Yes	No	No	5.73	2.64	-3.08
EHLA_1363 Hypothetical protein	No	Yes	No	0.995	2.51	1.52
EHLA_1432 AAA+ ATPase domain	No	Yes	Yes	-2.21	-5.78	-3.58
EHLA_1433 adenosinetriphosphatase	No	Yes	Yes	0.147	-4.81	-4.96
EHLA_1434 adenosinetriphosphatase	No	Yes	No	-2.72	-5.68	-2.96
EHLA_1443 Alpha/Beta hydrolase fold	No	Yes	Yes	-0.348	-4.97	-4.63

EHLA_1445 Methyltransferase domain	Yes	Yes	No	-3.68	-2.52	1.15
EHLA_1446 AMP-dependent synthetase/ligase	Yes	Yes	No	-2.78	-4.42	-1.64
EHLA_1455 Type III restriction enzyme, res subunit	Yes	No	No	2.82	-0.0356	-2.85
EHLA_1715 Restriction alleviation protein Lar	No	Yes	Yes	-0.509	3.34	3.84
EHLA_1716 Alkyl hydroperoxide reductase subunit C/ Thiol specific antioxidant	Yes	No	Yes	-3.84	0.318	4.16
EHLA_1763 Oxidoreductase family, NAD-binding Rossmann fold	Yes	No	Yes	-7.41	-0.854	6.56
EHLA_1811 consensus disorder prediction	No	No	Yes	-0.913	1.42	2.33
EHLA_1820 OmpR/PhoB-type DNA-binding domain profile.	Yes	No	Yes	3.99	0.227	-3.76
EHLA_1825 glutamate N-acetyltransferase/ amino-acid N-acetyltransferase	Yes	No	No	3.35	3.19	-0.158
EHLA_1826 N-acetyl-gamma-glutamyl-phosphate reductase	No	Yes	No	0.303	2.53	2.23
EHLA_1859 Acyl-CoA dehydrogenase/oxidase C-terminal	Yes	No	Yes	4.6	1.52	-3.07
EHLA_1866 D-isomer specific 2-hydroxyacid dehydrogenases NAD-binding signature.	No	No	Yes	-0.971	0.986	1.96
EHLA_1869 3-isopropylmalate dehydrogenase	Yes	No	Yes	-1.56	0.135	1.69

EHLA_1871 IMP dehydrogenase	Yes	No	No	-2.03	-0.685	1.34
EHLA_1874 Cytochrome-c3 hydrogenase, gamma subunit	No	No	Yes	1.53	-0.944	-2.47
EHLA_1880 protein-Npi-phosphohistidine-sugar phosphotransferase	Yes	No	Yes	-4.17	3.11	7.28
EHLA_1883 Short-chain dehydrogenase/reductase, conserved site	Yes	Yes	Yes	-3.45	4	7.44
EHLA_1900 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	No	Yes	No	1.19	5.24	4.05
EHLA_2082 Conserved carboxylase domain	Yes	No	No	-2.44	-0.79	1.65
EHLA_2083 Na pump decarbB: sodium ion-translocating decarboxylase, beta subunit	No	Yes	Yes	-0.34	4.28	4.62
EHLA_2087 Phosphoribosyltransferase domain	Yes	Yes	No	4.28	5.03	0.756
EHLA_2227 Lactaldehyde reductase	No	Yes	No	1.38	2.17	0.796
EHLA_2260 aspartate carbamoyltransferase	No	Yes	No	-1.46	-2.05	-0.59
EHLA_2288 Domain of unknown function DUF2088	Yes	Yes	No	5.85	5.41	-0.436
EHLA_2302 clpX: ATP-dependent Clp protease, ATP-binding subunit ClpX	Yes	No	No	-3.1	-1.13	1.97
EHLA_2313 secA: preprotein translocase, SecA subunit	No	No	Yes	-0.329	1.52	1.85

EHLA_2411 CRISPR type III-associated RAMP protein	Yes	No	No	6.09	0.49	-5.6
EHLA_2442 Carbohydrate kinase PfkB	Yes	Yes	No	4.81	4.29	-0.52
EHLA_2444 protein-Npi-phosphohistidine-sugar phosphotransferase	Yes	No	Yes	7.17	1.59	-5.58
EHLA_2459 Aspartyl/glutamyl-tRNA amidotransferase subunit B-related	Yes	No	No	-2.36	-1.49	0.871
EHLA_2463 Penicillin binding protein transpeptidase domain	No	Yes	No	-0.259	-2.68	-2.42
EHLA_2497 phosphoribosylaminoimidazolesuccinocarboxamide synthase	Yes	Yes	No	-1.95	-2.13	-0.181
EHLA_2645 Ribosomal protein S5 domain 2-type fold	Yes	No	No	-1.61	-0.724	0.887
EHLA_2692 adenosinetriphosphatase	No	Yes	Yes	-0.44	-5.21	-4.77
EHLA_2747 Peptidase S1C	No	Yes	No	-1.78	-3.1	-1.33
EHLA_2782 DNA-directed RNA polymerase	No	Yes	Yes	-0.82	1.86	2.68
EHLA_2809 L29: ribosomal protein uL29	No	No	Yes	-1.04	2.91	3.95
EHLA_2854 Cobalt uptake substrate-specific transmembrane region	Yes	Yes	No	4.59	4.18	-0.402
EHLA_2865 tryptophan synthase	No	Yes	No	-0.58	-2.19	-1.61
EHLA_2873 Periplasmic binding protein domain	Yes	Yes	Yes	7.44	4.26	-3.19

EHLA_2966 formate C-acetyltransferase	Yes	Yes	No	-7.51	-8.57	-1.05
EHLA_2967 [formate-C-acetyltransferase]-activating enzyme	Yes	Yes	No	-4.29	-6.33	-2.05
EHLA_2987 Phosphotransferase enzyme family	Yes	No	Yes	2.27	0.165	-2.1
EHLA_3027 Polyhedral organelle shell protein PduT	Yes	No	No	2.19	0.39	-1.8
EHLA_3028 Respiratory-chain NADH dehydrogenase 51 Kd subunit	Yes	No	Yes	1.84	-0.611	-2.45
EHLA_3029 Aldehyde dehydrogenase family	Yes	No	Yes	2.13	-0.442	-2.57
EHLA_3034 BMC domain	Yes	No	No	1.38	0.0741	-1.31
EHLA_3036 Diol dehydratase-reactivating factor, alpha subunit, swiveling domain	Yes	No	No	2.19	1.23	-0.96
EHLA_3037 Dehydratase small subunit	Yes	No	Yes	2.2	-0.0374	-2.24
EHLA_3038 Propanediol/glycerol dehydratase, medium subunit	Yes	No	Yes	2.35	-0.143	-2.5
EHLA_3039 propanediol dehydratase	Yes	No	Yes	2.25	-0.105	-2.36
EHLA_3046 glycogen phosphorylase	No	No	Yes	-1.37	1.58	2.95
EHLA_3132 leucyl aminopeptidase/bacterial leucyl aminopeptidase	No	No	Yes	1.42	-0.478	-1.9
EHLA_3142 CBS domain	No	No	Yes	3.58	-1.56	-5.14
EHLA_3151 ribose-phosphate diphosphokinase	No	Yes	No	-0.1	-1.78	-1.68
EHLA_3161 consensus disorder prediction	No	No	Yes	-1.05	1.48	2.54

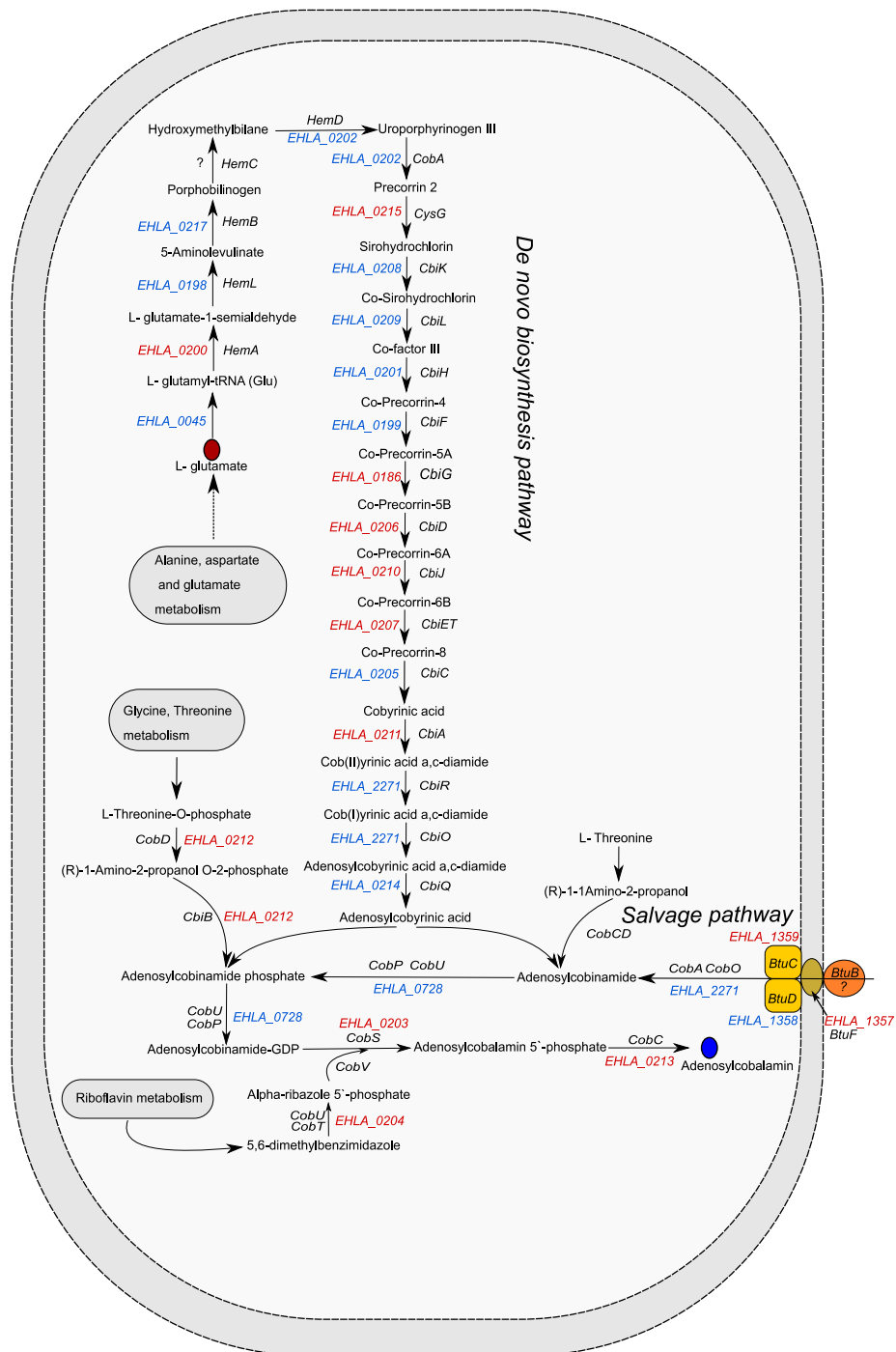|						
EHLA_3164 H+-transporting two-sector ATPase	Yes	No	No	1.59	1	-0.583
EHLA_3189 4-alpha-glucanotransferase	No	Yes	Yes	0.993	-2.94	-3.93
EHLA_3191 consensus disorder prediction	No	No	Yes	3.09	-1.97	-5.06

Figure S1: *De novo* and salvage pathways for vitamin B12 biosynthesis in *A. soehngenii*. The locus tags are with the prefix EHLA_. Those in blue were detected in the proteome. Locus tags in red were identified in the genome but not detected in proteome. No homologue for HemC was detected.
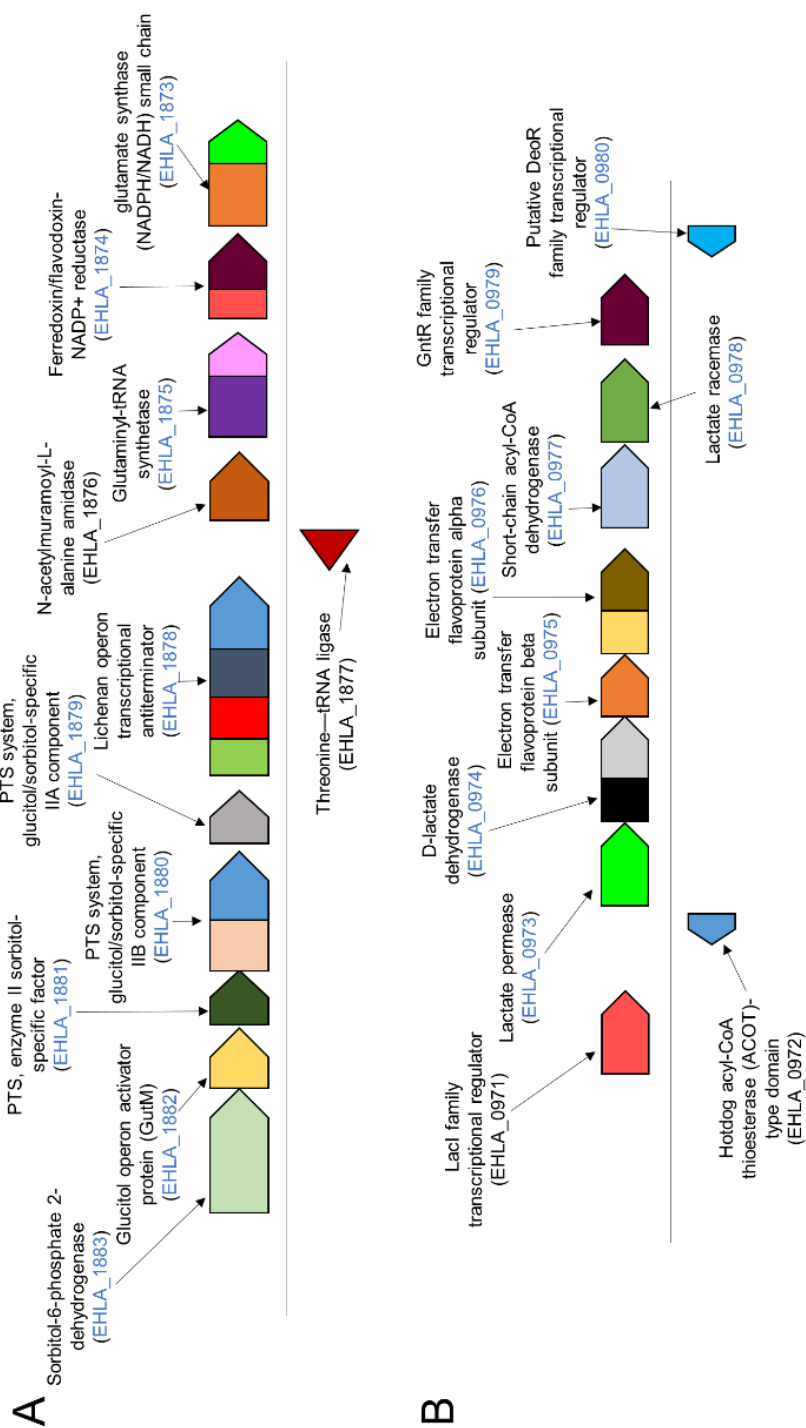
Figure S2: Organisation of gene clusters related to sorbitol and lactate utilization in the genome of *A. soehngenii*. **A]** Sorbitol and **B]** Lactate. The locus tags detected in the proteome are coloured in blue.
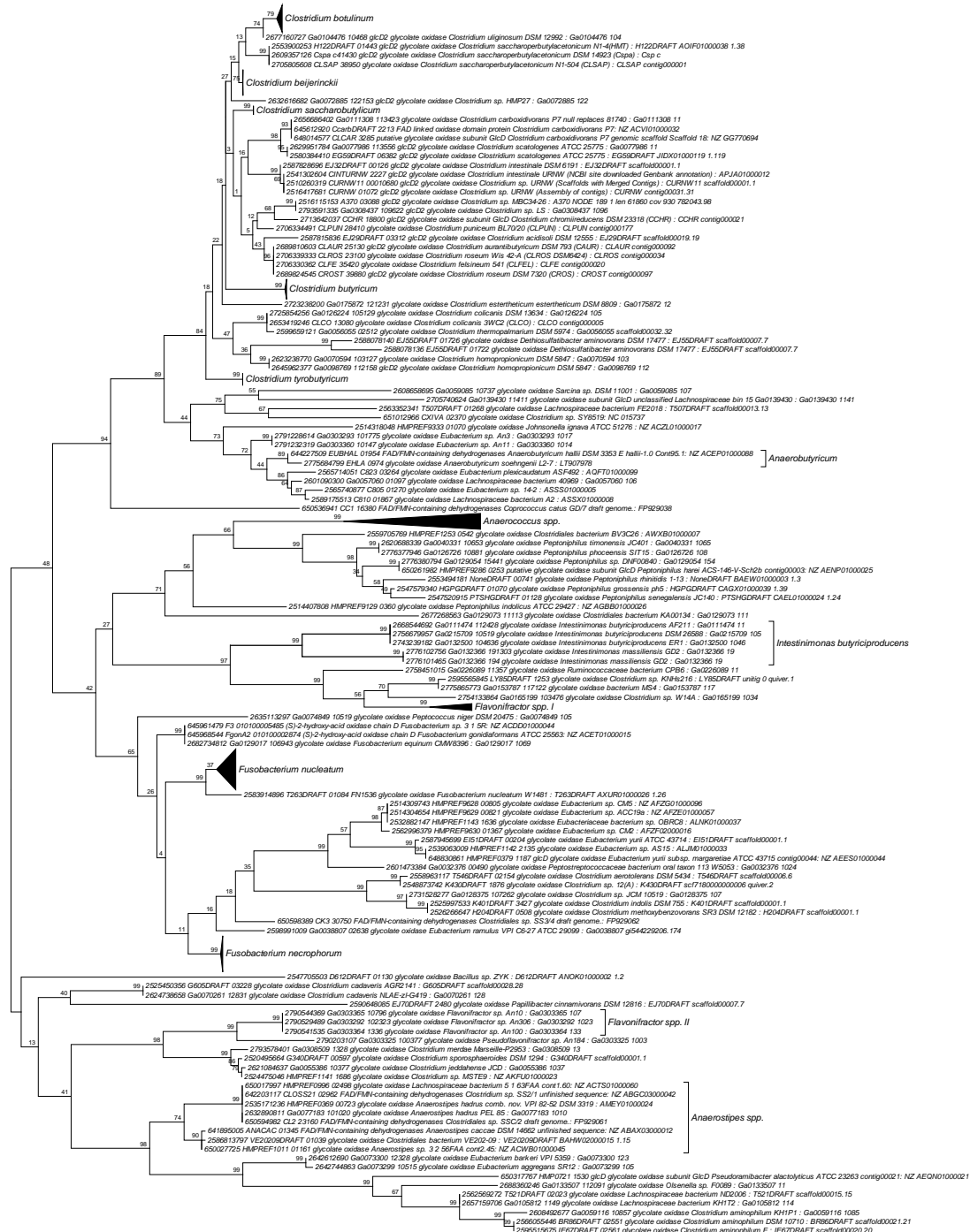
Figure S3: Maximum likelihood phylogenetic tree of 285 lactate dehydrogenase homologues. Phylogenetic tree of 285 LDH homologs identified by searching against 55,499 isolate genomes of the IMG/ER database (as of 2 October 2018). Labels represent the IMG gene id, IMG annotation for the gene product, taxonomic identity, strain name, and assembly and/or contig/scaffold which contains this gene. Where several genomes for a species were present, the branch of the tree was collapsed for clarity. The numbers on branches represent the bootstrap values from 1000 replicates. See methods for details on calculation.
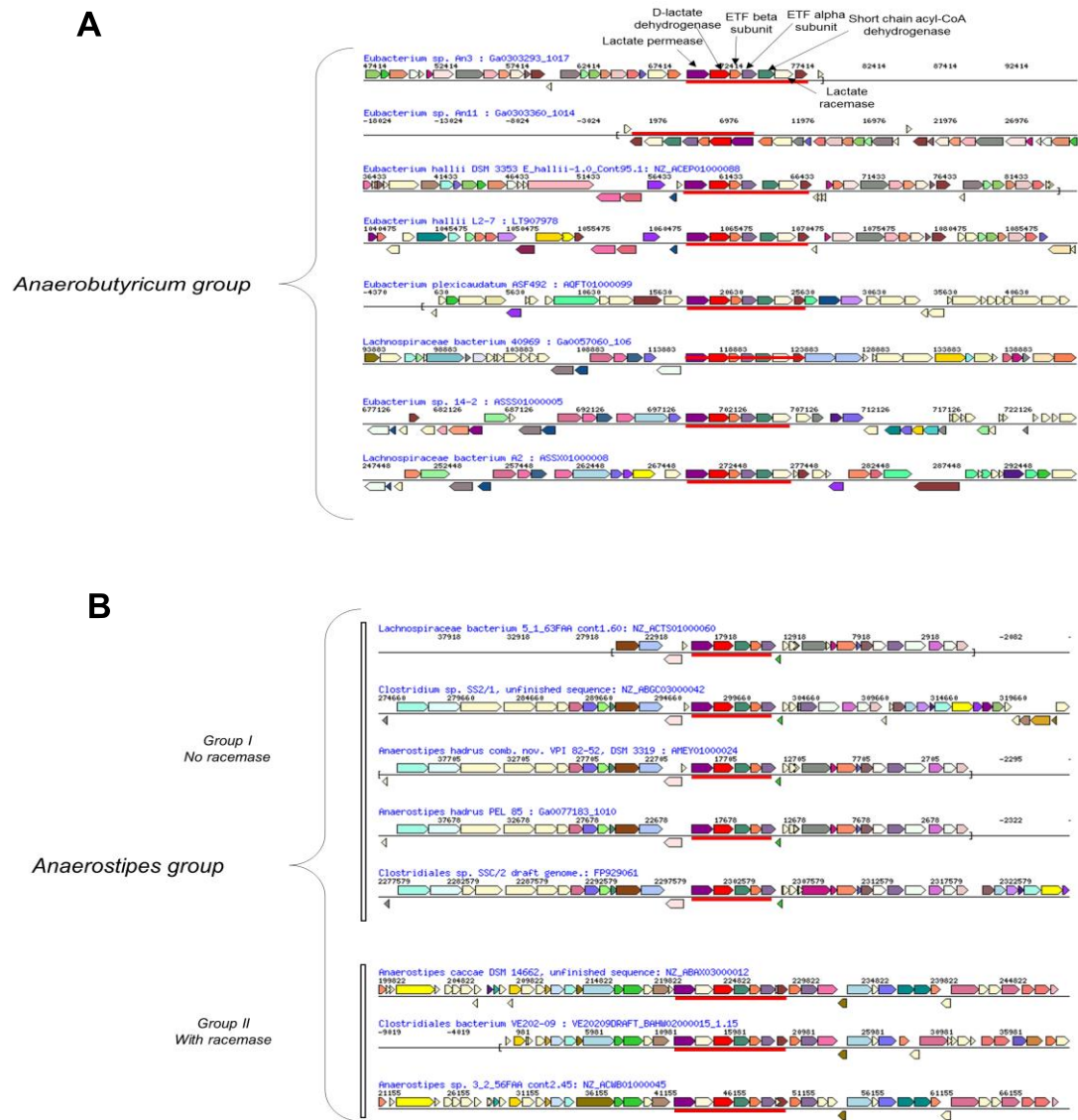
Figure S4: Comparison of the neighbourhood of genes involved in lactate utilization in genomes of *A. soehngenii* and related lactate utilizing bacteria. The red line highlights the gene cluster involved in lactate utilization. Neighbourhoods of genes in other genomes with the same top cluster of orthologs (COG) hit and roughly same matching length are shown below using the IMG gene neighbourhood search. Genes of the same colour (except light yellow) are from the same orthologous group (top COG hit). **A]** Selected bacterial genomes that share similar gene organisation as *A. soehngenii* lactate utilization gene cluster. **B]** Selected bacterial genomes that share similar gene organisation as *Anaerostipes caccae* lactate utilization gene cluster.

Chapter 8

Towards a minimal microbiome of the human intestinal tract

Sudarshan A. Shetty, Luuk Mommersteeg, Leo Lahti, Siavash Atashgahi,

Hauke Smidt, Willem M. de Vos
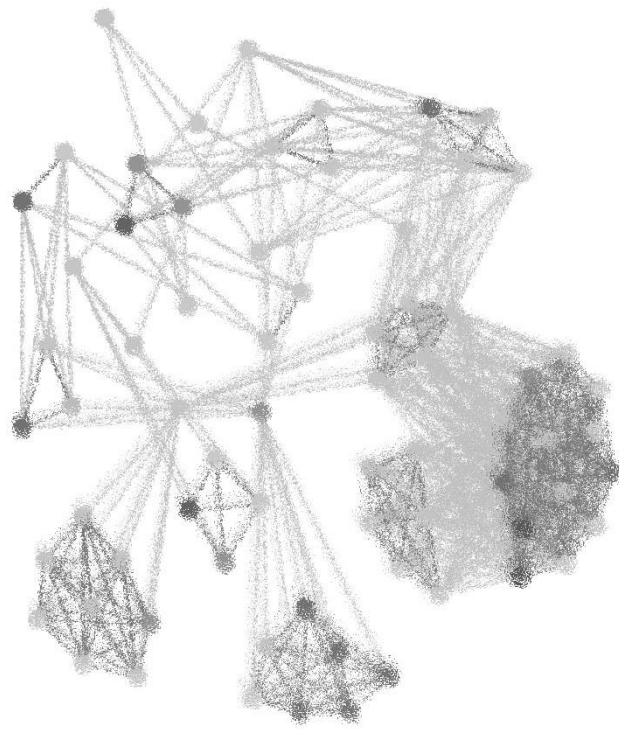
Abstract

The human intestinal tract microbiome has an intricate relationship with the host. One of the major challenges in understanding the ecological processes and metabolic interactions in the human microbiome is the high level of complexity. Defined microbial communities with known composition offer a unique opportunity to understand metabolic interactions in the human microbiome. Here, we designed a diet-based minimal microbiome (Db-MM), a defined microbial community comprising nine bacterial strains from the phyla Firmicutes and Bacteroidetes for butyrate and propionate production from dietary carbohydrates. The Db-MM was designed to incorporate four features that are relevant for the functioning of the colonic intestinal microbiome: a) being representative of the prevalent and abundant core bacteria; b) having a higher proportion of competitive than complementarity interactions; c) having functional redundancies at each trophic level; d) having trophic interactions *via* breakdown of complex dietary carbohydrates and fermentation by-products. We tested the ability of individual strains to utilize either a combination of cellobiose-xylan or starch-inulin as growth substrates in batch cultures. When grown in mixed cultures, the Db-MM community structure was shaped by the combination of available dietary carbohydrates. This signifies the strong deterministic role of nutrient availability on Db-MM community assembly. However, irrespective of the combinations of growth substrates used, a similar short chain fatty acid profile was observed with butyrate and propionate being produced in high amounts compared to mono-cultures. Together, the Db-MM designed in this study is a crucial step towards developing a functionally and phylogenetically diverse and experimentally accessible representation of the human intestinal tract microbiome.

Keywords: Minimal microbiome, Dietary carbohydrates, Core microbiota

Introduction

The microbial community assembly in the human intestinal tract has been proposed to be shaped by deterministic as well as stochastic factors (Ley*, et al.* 2006). A major deterministic factor in the human intestinal microbiome is the host diet (Cotillard*, et al.* 2013, David*, et al.* 2014, De Filippo*, et al.* 2010, Salonen and de Vos 2014, Salonen*, et al.* 2014). This is manifested already in early life as breastmilk-fed infants have a different microbial community compared to formula-fed infants (Bäckhed*, et al.* 2015, Baumann-Dudenhoeffer*, et al.* 2018). In adults, changes in diet can rapidly alter the microbiota composition and gene richness (Cotillard*, et al.* 2013, David*, et al.* 2014, Salonen*, et al.* 2014). For example, a diet rich in plant polysaccharides has been shown to enrich for bacteria from the genera *Roseburia*, *Eubacterium rectale* and *Ruminococcus bromii* (Chung*, et al.* 2016, Cotillard*, et al.* 2013, De Filippo*, et al.* 2010, Salonen and de Vos 2014), whereas animal-based diets have been shown to be associated with bacteria belonging to *Alistipes*, *Bilophila* and *Bacteroides* (David*, et al.* 2014). Thus, diet composition can support growth of certain bacteria and consequently impact co-existence of complimentary species *via* trophic interactions (Rothschild*, et al.* 2018, Vital*, et al.* 2018).

The trophic interactions in the intestinal tract facilitate co-existence of complementary species that share resources (Kovatcheva-Datchary*, et al.* 2009). However, competition is often more prevalent than complementarity in the human intestinal microbiome (Levy and Borenstein 2013). Such competition in the intestinal microbiome can have a stabilizing effect wherein a bacterial species can be replaced by another which provides the same ecosystem function (Coyte*, et al.* 2015, Li*, et al.* 2016). In order to understand how competing species, co-exist, it is important to understand their metabolic lifestyle in presence of other species in systems that are sufficiently well characterized. However, studying metabolic interactions as well as identifying emergent biosynthetic pathways resulting from multiple interacting species is challenging due to the complexity of natural microbiota.

Reducing the complexity of natural microbiomes by growing few representative species in defined mixtures has been recently employed for investigating metabolic

interactions and other ecological properties such as assembly in the intestinal microbiomes (D'hoe*, et al.* 2018, Venturelli*, et al.* 2018). However, most of the research was focused on defined co-cultures of two or three species, testing the occurrence of metabolic cross-feeding or competition (Belenguer*, et al.* 2006, Sonnenburg*, et al.* 2010, Belzer*, et al.* 2017, Chia*, et al.* 2018, D'hoe*, et al.* 2018).

For instance, in a two species model with *Bacteroides ovatus* (an inulin degrader) and *B. vulgatus* (a non-inulin degrader), cooperation can evolve if there are reciprocal benefits to *B. ovatus* by feeding *B. vulgatus* (Rakoff-Nahoum*, et al.* 2016). It was further reported that *B. ovatus* had a higher fitness when it provided a simple carbon source to *B. vulgatus*. Hence, the presence of other bacteria can influence the metabolism and fitness of the primary degrader. Incorporating such interaction in a defined microbial community can provide insights into factors governing *in situ* observed co-existence and/or competition of bacteria. In addition, including competition at different trophic levels can provide insights into the microbial interaction network and assembly into distinct clusters of co-occurring bacteria or functional guilds arising from high order interactions. This can be approached by applying the concept of the minimal microbiome that was coined several years ago (de Vos 2013). To this end, we here developed a consortium of a limited number of well-studied bacterial species that together provide the vital functions of the intestinal microbiome.

Since diet is a major deterministic factor governing community assembly, composition and function it is important to understand the metabolic interactions between members of the intestinal microbiome and how diet governs intestinal ecology (Flint*, et al.* 2012a, Korpela*, et al.* 2014, Salonen*, et al.* 2014). In the present study, we applied an ecophysiology guided approach by taking into account the ecological properties and physiological roles of core intestinal bacteria and designed a butyrate- and propionate-producing diet-based minimal microbiome (Db-MM) as a representative of the intestinal microbiome. We use the term Db-MM because the minimal microbiome designed in this study reflects the colonic conversions of dietary carbohydrates. The Db-MM consists of representative bacterial species from the two major phyla Firmicutes and Bacteroidetes that were selected based on their ability to break down specific carbon sources (in our study starch, inulin, xylan, cellobiose) and/or produce

specific metabolites such as acetate, propionate and butyrate. Using batch *in-vitro* culturing approaches, we investigated community assembly, dynamics and activity of the Db-MM using different sets of common carbon sources as a deterministic factor.

Materials and Methods:

An overview of the experimental workflow is shown in Figure 1.

Selection of bacterial strains

Taxonomic composition data from metagenomic studies was obtained from the curatedMetagenomicData data package (Pasolli*, et al.* 2017). The data was analyzed at species level to identify taxa that are part of the core microbiota. A total of 54 metagenomic species were selected, which were present in at least 50% of all samples analyzed. For all selected bacteria either a high-quality draft or complete genome sequence was available.

Bacterial strains and culturing conditions

The following strains were obtained from the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures (Braunschweig, Germany): *Eubacterium rectale* (DSM 17629), *Eubacterium siraeum* (DSM 15702), *Roseburia intestinalis* (DSM 14610), *Subdoligranulum variabile* (DSM 15176). *Eubacterium hallii* (DSM 1736) was kindly provided by Prof. Harry J. Flint's group (University of Aberdeen, UK). The strains selected from the human microbiome project (HMP) catalogue were *Bacteroides* sp. 3_1_23 (*Bacteroides ovatus*), *Bacteroides* sp. 2_1_22 (*B. xylanisolvens*) and *Lachnospiraceae* bacterium 7_1_58FAA. These strains obtained from HMP and *Faecalibacterium prausnitzii* (A2-165) were kindly provided by Dr. Clara Belzer (Laboratory of Microbiology, Wageningen University & Research, the Netherlands). Bacterial cells were stored at -80°C in pre-reduced 20% polyethylene glycol in phosphate buffer saline (PBS). The following media composition was used for all experiments.
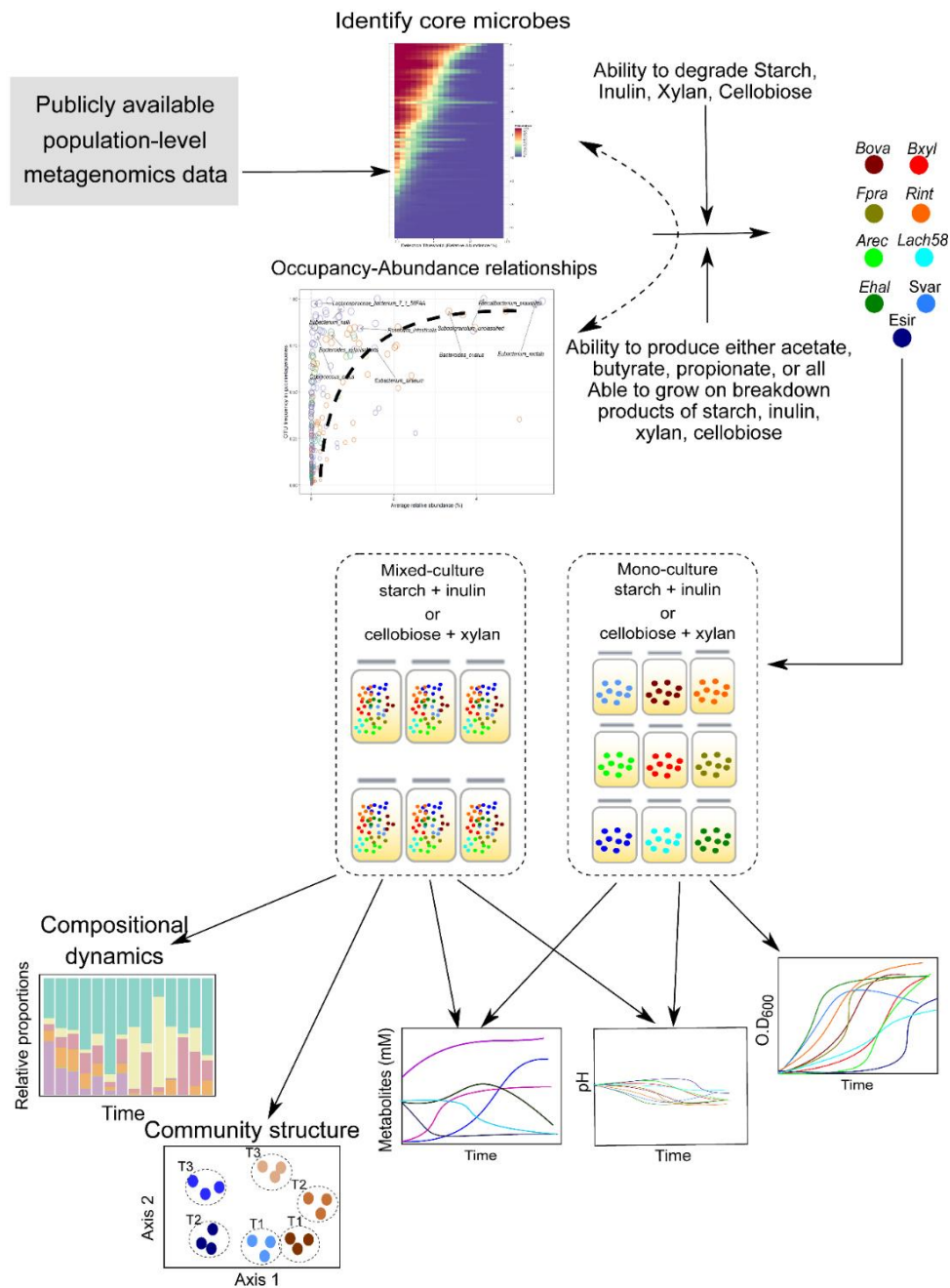
Figure 1: Experimental workflow for selection of bacterial strains, design and ecophysiological study of a diet-based minimal microbiome (Db-MM). The core microbiota was identified based on public metagenomic datasets, followed by rational selection of bacteria to represent primary and secondary consumers. First, mono-cultures were tested for growth on two mixtures of carbohydrates, to confirm primary and secondary consumer strains. The strains were subsequently combined to an approximate proportion of 1:1 based on the OD600 values and grown on the same carbohydrate mixtures. Samples were taken at 0, 24, 48 and 72 h to measure growth, pH, metabolites and community composition. Abbreviations: Bova, *Bacteroides ovatus*; Bxyl, *B. xylanisolvens*; Fpra, *Faecalibacterium prausnitzii*; Rint, *Roseburia intestinalis*; Erec, *Eubacterium rectale*; Lach58; *Lachnospiraceae* sp. 58_7_FAA; Ehal, *E. hallii*; Svar, *Subdoligranulum variabile*; Esir, *Eubacterium siraeum*.

The media components were KH_2PO_4 (27.2 g/L), $Na_2HPO_4.2H_2O$ (35.6 g/L), NH_4Cl, (24 g/L), NaCl, (24 g/L), $MgCl_2.6H_2O$ (8 g/L), $NaHCO_3$ (4 g/L), Yeast extract (2 g/L), Casitone (2 g/L), Beef extract (2 g/L), Sodium acetate (2.46 g/L), Peptone (2 g/L), 1ml of trace elements in acid (HCl, 50 mM; H_3BO_3, 1 mM; $MnCl_2.4H_2O$, 0.5 mM; $FeCl_2.4H_2O$, 7.5 mM; $CoCl_2.6H_2O$, 0.5 mM; $NiCl_2.6H_2O$, 0.1 mM; $ZnCl_2$, 0.5 mM; $CuCl_2.2H_2O$, 0.1 mM), 1 ml of trace elements in alkaline (NaOH, 10 mM; Na_2SeO_3, 0.1 mM; $Na_2WO_4.2H_2O$, 0.1 mM; $Na_2MoO_4.2H_2O$, 0.1 mM), 0.5 g Cysteine-HCl, 1 mL Haemin solution (50 mg haemin; 1 mL NaOH, 1N; 99 mL dH2O), 0.2 mL Vitamin K1 solution (0.1 mL vitamin K1; 20 mL ethanol, 95%). The medium was boiled and cooled under N_2 gas, and 45 ml aliquots were distributed in 120 ml bottles. The bottles were sealed with rubber stoppers and aluminum caps, and the headspace gas was exchanged with an N_2/CO_2 gas mixture (4:1 ratio). After autoclaving and before inoculation, 1% vitamin in $CaCl_2$ solution was added. The vitamin solution contained biotin (20 mg/L), nicotinamide (200 mg/L), p-aminobenzoic acid (100 mg/L), thiamine (B1) (200 mg/L), pantothenic acid (100 mg/L), pyridoxamine (500 mg/L), cyanocobalamin (B12) (100 mg/L) and riboflavin (100 mg/L) dissolved in deionized water. For pre-cultures, the bacteria were grown on combinations of different carbon sources (Table S1). The carbon sources, prepared as 1% stock solutions of xylan, soluble starch (from potato), inulin (from chicory) and cellobiose in anoxic water, were added by filter (0.22μM, MDI, India) sterilization after autoclaving the media. The pre-cultures were incubated non-shaking at 37ºC in the dark for 24h.

Testing of growth in mono-cultures and in consortia

For testing growth of mono-cultures on mixtures of complex carbohydrates, 0.5 ml of actively growing culture in mid exponential phase was used to inoculate duplicate bottles containing 50ml fresh medium. Growth (OD_{600}), pH and short chain fatty acid (SCFA) production were measured at 0, 24, 48 and 72h after inoculation. For testing growth of mixed cultures on mixtures of complex carbohydrates, the OD_{600} of the actively growing pure pre-cultures was measured, and strains were combined at an approximate 1:1 ratio, and used for inoculation of triplicate bottles containing 50ml fresh medium. Samples (2ml) were taken for measuring growth (OD_{600}), pH, SCFA

production and DNA extraction for community analysis 0, 24, 48 and 72h after inoculation.

A Shimadzu Prominence-i LC-2030c HPLC liquid chromatograph equipped with a Shodex SUGAR SH1011 column was used for the analyses of SCFA (formate, acetate, lactate, propionate, iso-butyrate and butyrate) and simple sugars (glucose, xylose and fructose). For the mobile phase, 0.01 N H_2SO_4 was used. Samples were prepared by centrifuging 1 mL of the bacterial culture at 13,000 rpm for 10 min and adding 400 µL of the supernatant to 600 µL of 30 mM crotonate as an internal standard. Standard curves for SCFA's and simple sugars were prepared using four different concentrations (2.5, 5, 10 and 20 mM).

DNA extraction and sequencing

Genomic DNA was extracted using the FAST DNA Spin Kit (MP Biomedicals, Fisher Scientific, Netherlands) following the manufacturer's instructions. The concentration of gDNA was measured using a Nanodrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). The hypervariable region V5-V6 (~280bp) of the 16S rRNA gene was amplified with Phusion Hot start II DNA polymerase (2U/µl) for 25 cycles using 0.05 µM of each primer (784F - 1064R) that both contained sample-specific barcodes at their 5'-end. The amplification program for PCR included 30 s initial denaturation step at 98 °C, followed by 25 cycles of denaturation at 98 °C for 10 s, annealing at 42 °C for 10 s, elongation at 72 °C for 10 s, and a final extension at 72 °C for 7 min. Purified PCR products were quantified using the Qubit dsDNA BR Assay Kit (Life Technologies, USA) and were pooled in equimolar amounts into one single library and sequenced on an Illumina HiSeq platform in 2×150 bp paired-end mode at GATC Biotech (Constance, Germany).

Data processing and community analysis

The raw reads were processed in R using the DADA2 R package (Callahan, *et al.* 2016). A total of 6,150,598 reads were obtained from 24 samples with a minimum of 42,952 and maximum of 7,82,125 reads after filtering the reads with low quality, removal of reads with more than 2 errors and those matching the PhiX (filterAndTrim function), chimeric sequences (removeBimeraDenovo, consensus method). The taxonomic assignment was done using a custom database containing the full length 16S rRNA gene

sequence for strains used in the present study. The classification was done using the RDP classifier (Wang, *et al.* 2007b). The sequence variants that were unclassified were removed from further analyses. The amplicon sequence variants identified by DADA2 were collapsed at species level and normalized for the differences in copy numbers. The total counts for each species were divided by the number of 16S rRNA gene copies identified in its genome (Table 1). Further analysis of the community composition and structure was done using the microbiome R package (Lahti and Shetty 2018).

***In silico* prediction of pairwise competition and complementarity**

Genome annotations for the selected bacterial strains were downloaded from the Joint Genome Institute Integrated Microbial Genomes & Microbiome System (JGI-IMG/ER) database. Using the KEGG ortholog annotations, a metabolic network was reconstructed. This was used to identify a set of seed compounds which are acquired by the bacterium from the extracellular environment and non-seed compounds which are produced by the bacterium. Using the seed and non-seed compounds for each bacterium, the pairwise metabolic complementarity and competition were calculated following the reverse ecology method described previously (Levy and Borenstein 2013) in R using the RevEcoR package (Cao, *et al.* 2016).

Results and Discussion

Prevalent bacterial strains with key roles in the human intestinal tract

Metagenomic data from 494 stool samples of individuals from the USA, Denmark, Spain, Italy, France and Germany were analyzed. Out of a total of 2573 metagenomic species, 54 were identified as members of the core microbiota based on their presence in at least 50% of the subjects at a minimum relative abundance of 0.0001 (Figure 2). From these 54 core bacterial metagenomic species, nine were selected for the experiments described here. The 54 members of the core intestinal microbiota share high functional similarity and therefore, representative strains from the three predominant families Bacteroidaceae, Ruminococcaceae and Lachnospiraceae were selected.
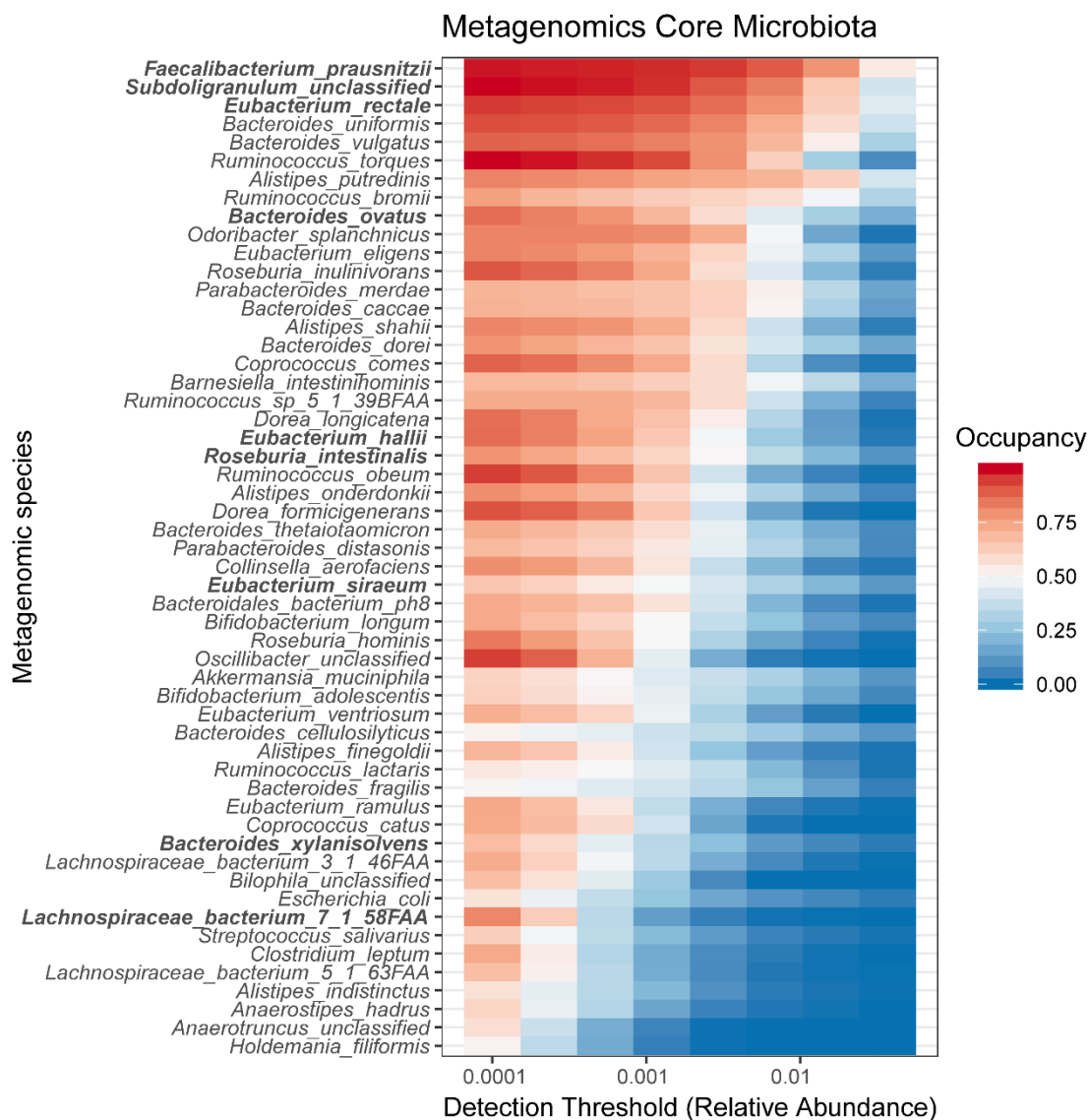
Figure 2: Heat map depicting the core microbiota in 494 human fecal metagenomic samples. The bacterial candidates for Db-MM are shown in boldface.

We selected nine bacterial species that represent the dominant and rare fractions of highly prevalent core species and have a relative high and low mean abundance, respectively (Figure S1). The occupancy (65-99 %) and known metabolic properties of the bacterial strains used as representative of the metagenomic species in the present study are given in Table 1. Among the selected bacterial strains, *Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Bacteroides ovatus*, *Roseburia intestinalis*, *Bacteroides xylanisolvens* are all capable of utilizing complex carbohydrates and can compete for resources (Table 1). *Subdoligranulum* sp. and *Eubacterium hallii* are not known to degrade complex carbohydrates.

Table 1: Occupancy and metabolic features of the selected bacterial strains.

Sr. No.	Bacterial strain	Strain used (16S rRNA gene copy number[#])	Occupancy (%)	Carbon source breakdown/utilization§, (reference)
1	*Eubacterium rectale*	*Eubacterium rectale* DSM 17629 (1)	94.9	Starch, cellobiose, inulin Butyrate and lactate production (Cockburn*, et al.* 2015, Duncan and Flint 2008)
2	*Bacteroides ovatus*	*Bacteroides ovatus* 3_8_47FAA (2)	86.2	Starch, xylan Acetate and propionate production (Sheridan*, et al.* 2016)
3	*Bacteroides xylanisolvens*	*Bacteroides* sp 2_1_22 (1)	67.6	Xylan, cellobiose (Chassard*, et al.* 2008)
5	*Eubacterium hallii*	*Eubacterium hallii* DSM 1736 (8)	86.4	Glucose, lactate plus acetate (Engels*, et al.* 2016b)
6	*Eubacterium siraeum*	*Eubacterium siraeum* DSM 15702 (4)	65.0	Starch*, fructose, cellobiose (Moore*, et al.* 1976)
7	*Faecalibacterium prausnitzii*	*Faecalibacterium prausnitzii* A2-165 (3)	98.8	Starch, inulin (Duncan*, et al.* 2002b)
8	Lachnospiraceae bacterium 7_1_58FAA	Lachnospiraceae bacterium 7_1_58FAA (1)	81.0	Glucose, cellobiose (This study)
9	*Roseburia intestinalis*	*Roseburia intestinalis* DSM 14610 (1)	78.3	Starch, cellobiose*, fructose, xylan (Duncan*, et al.* 2002a)
10	*Subdoligranulum* unclassified	*Subdoligranulum variabile* DSM 15176 (3)	99.8	Cellobiose, glucose (Holmstrøm*, et al.* 2004)

*Weak growth reported previously. § Information based on representative strains for each of the species. [#] The 16S rRNA gene copy numbers were calculated using the genome sequence.

No published information was available regarding metabolic capabilities of Lachnospiraceae bacterium 7_1_58FAA but we determined these in the course of this study (Table 1).

To further investigate if the selected bacteria can compete and/or complement each other, we used a reverse ecology approach, which uses genome-predicted metabolic networks (Levy and Borenstein 2013). We observed that the strains belonging to the

two *Bacteroides* spp. had the highest competition and the lowest complementarity between them (Figure 3). On the other hand, the majority of the bacteria belonging to the Firmicutes had higher complementarity with *Bacteroides* spp. compared to the other Firmicutes (Figure 3). *R. intestinalis* was predicted to have competition with both *Bacteroides* strains, whereas *F. prausnitzii* was predicted to have relatively higher competition with *Bacteroides* sp. 2_1_22 (*B. xylanisolvens*) than with *B. ovatus* (Figure 3).
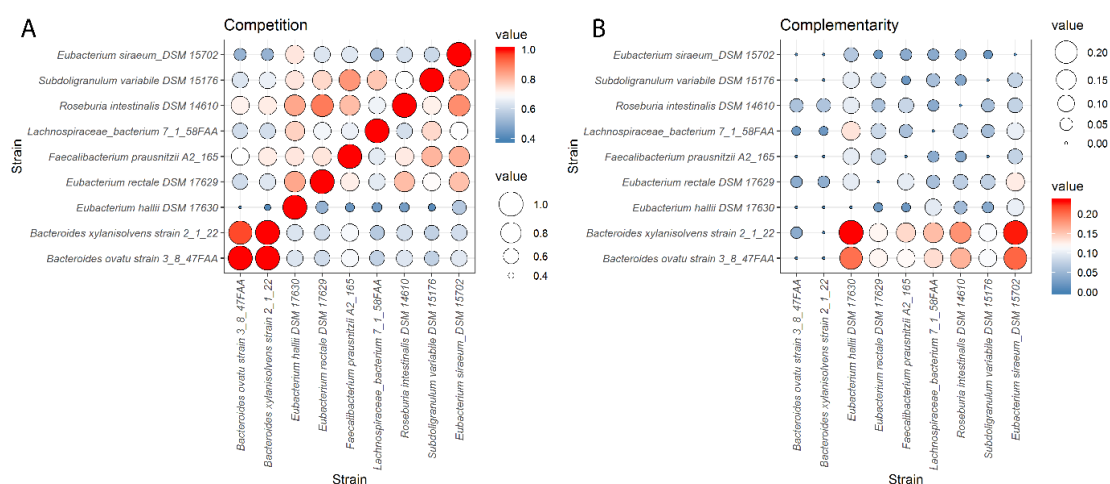

Figure 3: Reverse ecology based prediction of pairwise competition and complementarity between nine bacterial strains selected for the Db-MM based on genome based metabolic interaction networks. **A]** Predicted strength of pairwise metabolic competition. **B]** Predicted strength of pairwise metabolic complementarity.

Growth and SCFA production in mono-cultures on mixtures of carbon sources

Both *Bacteroides* spp., *B. ovatus* and *B. xylanisolvens*, showed growth on cellobiose-xylan or starch-inulin as carbon sources (Figure 4). *B. xylanisolvens* reached higher cell densities with both sets of carbon sources as compared to *B. ovatus*. Both *Bacteroides* spp. have a diverse repertoire of glycan degrading enzymes (Chassard*, et al.* 2008, Despres*, et al.* 2016). With cellobiose-xylan as carbon sources, acetate was the major end product of fermentation in *B. ovatus* and *B. xylanisolvens* cultures, followed by propionate (Figure 5). Small amounts of lactate (< 5mM) were detected in *B. ovatus* cultures. Propionate and acetate were the major end products for *B. xylanisolvens* and *B. ovatus* when grown on starch-inulin (Figure 6). Minor amounts of lactate and propionate were detected in cultures of *B. ovatus* at 24h that remained unchanged after 72h of incubation (Figure 6). Free fructose, but not glucose, was detected after

24h in the mono-cultures of *B. ovatus* and *B. xylanisolvens* when grown in presence of starch-inulin (Figure 6). *B. ovatus* and *B. xylanisolvens* are metabolically versatile and showed good growth in presence of cellobiose-xylan. This is in line with previous observations regarding the glycan degrading ability of these two species (Sonnenburg*, et al.* 2010, Chassard*, et al.* 2008, Leth*, et al.* 2018). *B. xylanisolvens* reached the highest cell densities on both cellobiose-xylan and starch-inulin compared to all other bacterial strains used in this study.

Among the bacteria belonging to the Firmicutes, *Eubacterium rectale* demonstrated higher growth with starch-inulin as compared to cellobiose-xylan. The cell density of *E. rectale* grown on starch-inulin declined after 24h. Previously, *E. rectale* strain A1-86 was reported to utilize various types of inulin with different branch lengths (Scott et al., 2014). Butyrate was the major fermentation end product for *E. rectale* on cellobiose-xylan and starch-inulin (Figures 5 and 6). Highest cell density and butyrate production was observed when *E. rectale* was grown on starch-inulin, confirming its prominent role for butyrogenesis from starch (Cockburn*, et al.* 2015). *F. prausnitzii* grew on both cellobiose-xylan and starch-inulin (Figure 4). Higher cell density was observed in the presence of cellobiose-xylan compared to starch-inulin. Butyrate and formate were the major end products with both carbon source combinations (Figures 5 and 6). However, the weak growth on complex carbohydrates suggests its lack of adaptation to these substrates even after 48h. *R. intestinalis* grew to an OD600 comparable to *B. ovatus* in the presence of cellobiose-xylan, whereas only marginal growth was observed with starch-inulin (Figure 4). Butyrate was the major end product in both conditions, while minor amounts of lactate, propionate and formate were detected with cellobiose-xylan (Figure 5 and 6). *E. siraeum* showed weak growth on both cellobiose-xylan and starch-inulin (Figure 4). A net production of acetate was observed in *E. siraeum* cultures (Figure 5 and 6). *E. hallii* and *Subdoligranulum variabile* as other known butyrate-producing bacteria showed marginal to no growth on cellobiose-xylan, confirming their inability to utilize these substrates (Figure 4) (Holmstrøm*, et al.* 2004, Barcenilla*, et al.* 2000). The unclassified Lachnospiraceae bacterium 7_1_58FAA bacterium showed similar behavior as *E. hallii* and *S. variabile* with starch-inulin and cellobiose-xylan, confirming its inability to degrade complex polysaccharides (see Table 1).
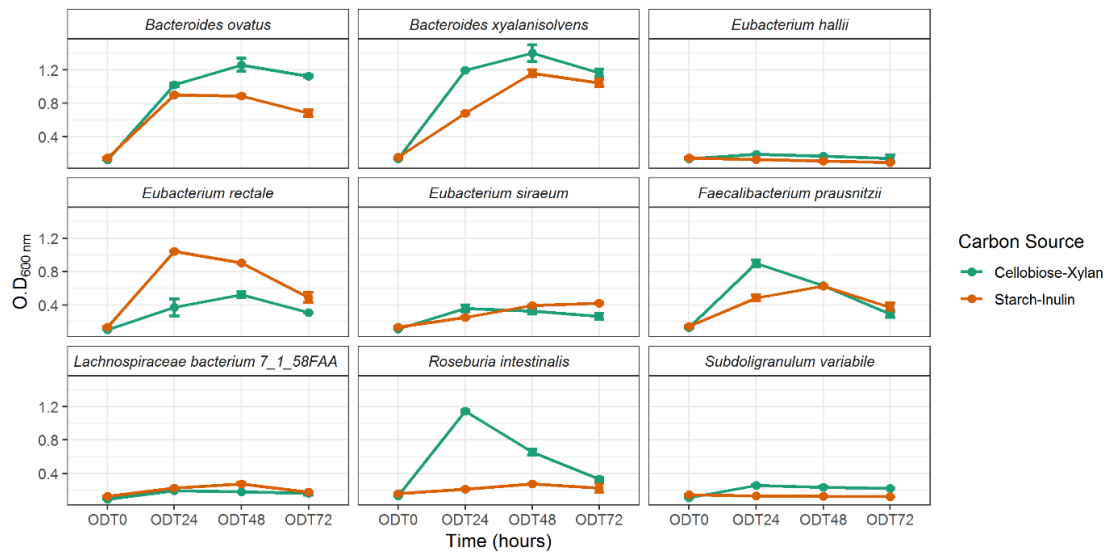
Figure 4: Growth curves of the nine selected strains in mono-cultures grown on either cellobiose-xylan or starch-inulin.
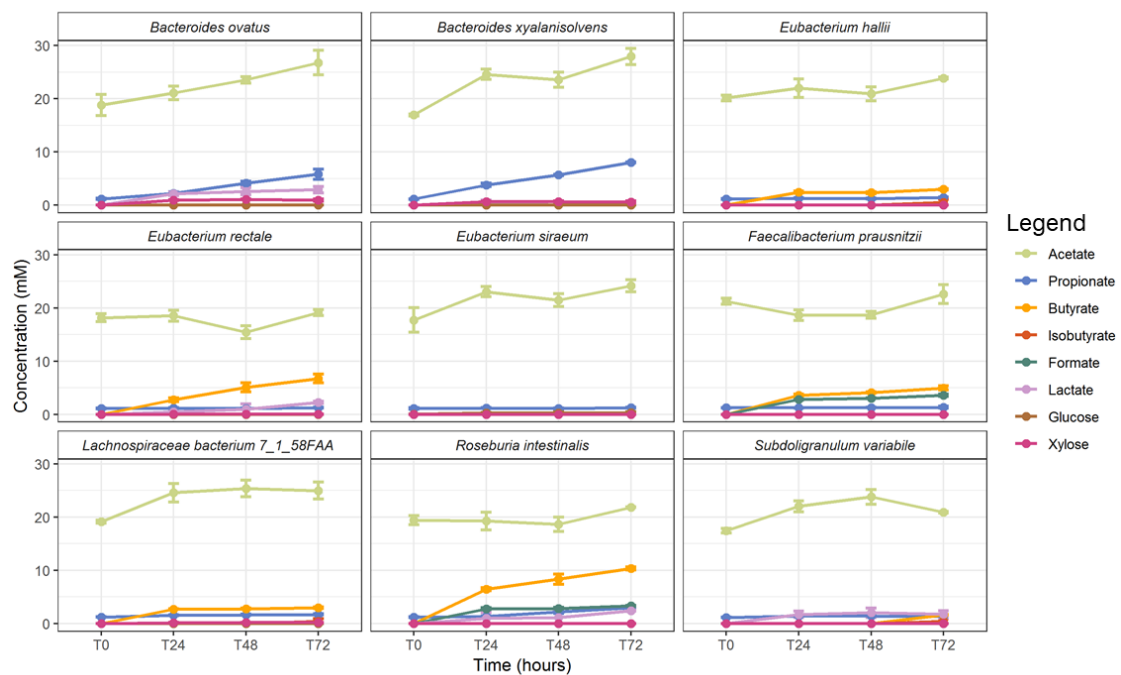

Figure 5: Short chain fatty acid profiles for the nine selected strains grown in mono-cultures on cellobiose-xylan.

Overall, the observations of mono-cultures suggest that the majority of the bacteria selected in this study are able to grow better in presence of cellobiose-xylan as compared to starch-inulin. Among the Firmicutes, *E. rectale* showed best growth on starch-inulin and was the best butyrate-producer. *R. intestinalis* produced the highest amount of butyrate on cellobiose-xylan. In contrast *E. hallii*, *S. variabile* and

Lachnospiraceae bacterium 7_1_58FAA showed marginal to no growth supporting their inability to directly utilize the main carbohydrates used in this study.
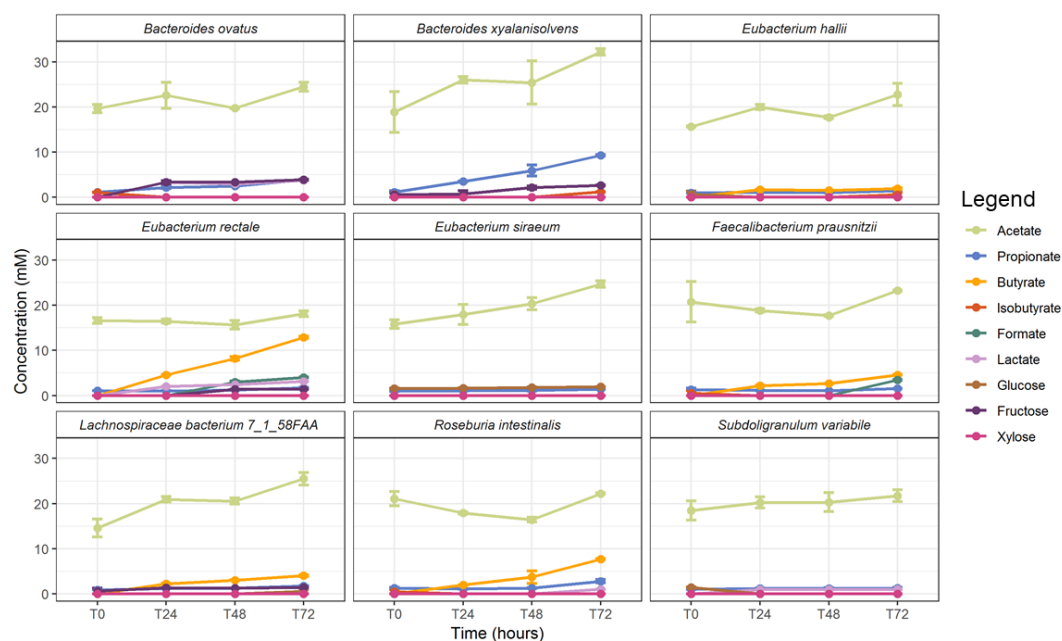

Figure 6: Short chain fatty acid profiles for the nine selected strains grown in mono-cultures on starch-inulin.

Physiological and Compositional Dynamics of Db-MM

Physiological dynamics of Db-MM

We investigated the changes in pH and SCFA profiles during the assembly process of the mixed culture that was composed of similar amounts of growing cells of each of the nine species. A marginal change in pH was observed during growth on cellobiose-xylan and starch-inulin (Figure S2). Butyrate was the major end product of carbohydrate fermentation for both substrate combinations. Propionate and formate were the second and third most abundant SCFA (Figure 7A), with propionate concentrations being higher on cellobiose-xylan compared to starch-inulin (Figure 7B). Additionally, lactate was produced at 24h in the presence of starch-inulin, but was rapidly consumed with concomitant decrease in acetate concentrations (Figure 7B).
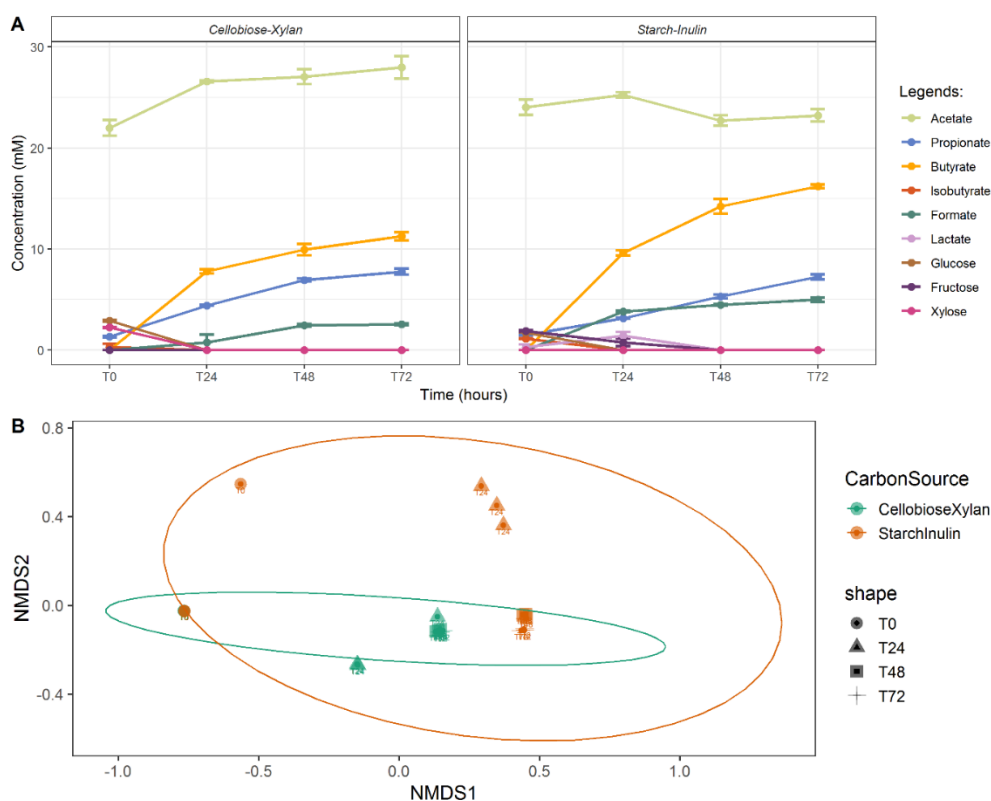
Figure 7: Short chain fatty acid profiles of Db-MM in presence of **A]** Cellobiose-xylan, or **B]** Starch-inulin. **C]** Nonmetric multidimensional scaling (NMDS) plot (Bray-Curtis dissimilarity) depicting the dynamics and convergence of the Db-MM metabolic profile over time (Stress 0.01). The ellipses represent 95% confidence intervals around the centroid.

Next, we used the SCFA profiles to identify community structure based on active metabolism. Analysis of pairwise dissimilarities between the samples highlighted similar metabolic profiles between the two growth conditions (cellobiose-xylan vs starch-inulin) after 72h of incubation, and hence high functional similarities irrespective of the different growth substrates. The convergence of community metabolic profiles was more prominent compared to that observed for species abundance profiles (Figure 7C and 8B). To the best of our knowledge, this is the first time that functional redundancy for metabolic end products was explicitly and successfully incorporated in a defined bacterial mixture.

Composition and assembly dynamics of Db-MM

The individual bacterial strains followed different growth trajectories over time with both substrate combinations. The community composition on cellobiose-xylan showed

increasing unevenness, whereas on starch-inulin the peak of unevenness was observed at 24h, after which it declined at 72h (Figure 8A, B).

The slight decline in unevenness was likely a result of reduced abundance of *E. rectale* and increased abundance of *B. xylanisolvens*, *B. ovatus*, *F. prausnitzii* and Lachnospiraceae bacterium 7_1_58FAA (Figure 9A). This change in composition was concomitant with utilization of lactate and acetate (Figure 7B). In the mono-culture studies with starch-inulin, we observed that *E. rectale* had the highest growth after 24h and produced lactate (Figure 4). Free fructose was detected in mono-cultures of *B. xylanisolvens* along with lactate (Figure 4).
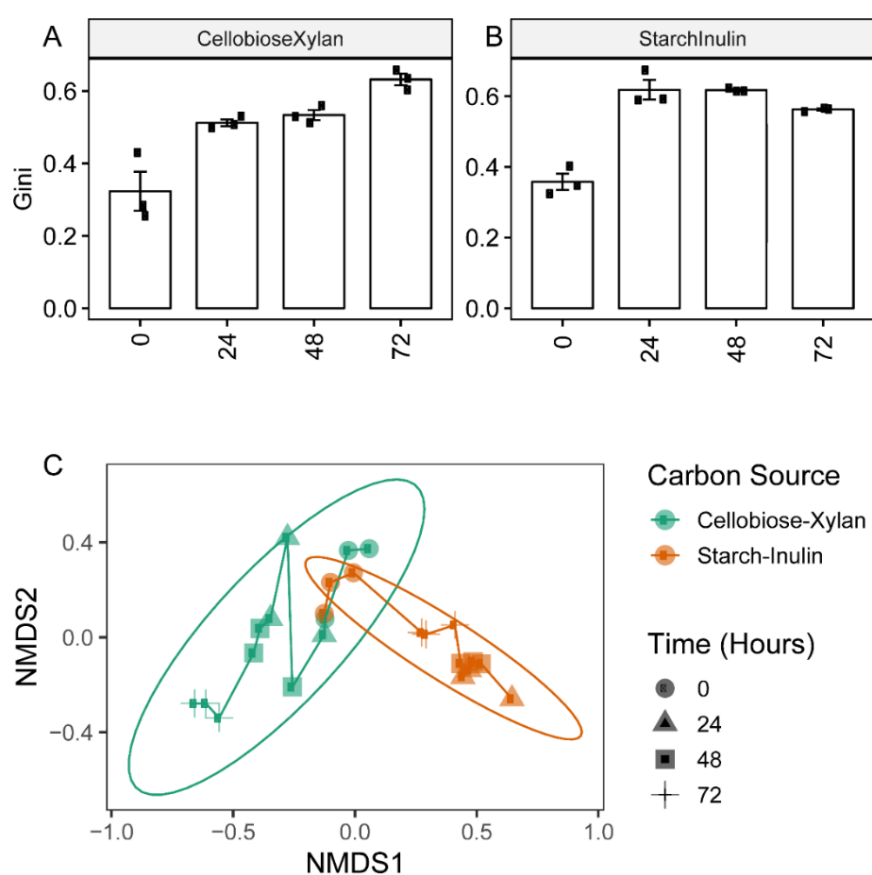

Figure 8: Community unevenness and community dissimilarity. Gini was used as a measure of unevenness in the community. **A]** Cellobiose-Xylan **B]** Starch-Inulin. **C]** Community dissimilarity was calculated using the Bray-Curtis dissimilarity index and visualized in a two dimensional space based on the non-metric multidimensional scaling method (stress 0.08). The lines connect different replicates over time in each of the growth conditions. The ellipses represent 95% confidence intervals around the centroid.

At community level, these observations imply a major role for the primary degraders and demonstrate that the change in the main carbon source and consequent changes

in identity of primary degraders can have major impact on the evenness and structure of the microbial community. In cellobiose-xylan fed cultures, *B. xylanisolvens* was the most dominant species at 72h. At 24h, *B. ovatus* showed moderately higher growth but reduced in read counts at 48h. *B. xylanisolvens* outcompeted *B. ovatus* and *R. intestinalis* in all triplicate consortia underpinning its competitiveness for xylan. *R. intestinalis* was previously reported to compete strongly with *B. ovatus* in two species co-culture experiments (Leth*, et al.* 2018). Our experimental set-up incorporates the possibility for high order interactions, *i.e.* presence of other species and their effects on pairwise competition. To the best of our knowledge, such interactions have not been explicitly simulated in studies investigating assembly dynamics of defined bacterial communities. Similar to the starch-inulin condition, there was similarity in the community function between replicates at 72h (Figure 6B and 7B).

Figure 9: Variation in normalized counts of the 16S rRNA gene amplicons representing the different strains in triplicate cultures supplemented with **A]** cellobiose-xylan and **B]** starch-inulin. The number of reads in each sample was normalized by rarefying to 11,528 reads/sample. The counts of the individual bacteria were normalized to the number of 16S rRNA gene copies in their genomes (Table 1). The individual panels represent samples taken at 0, 24, 48 and 72 h, respectively.

Previous investigations of the assembly of microbial communities derived from soil and plants and defined maize root-associated bacterial mixtures have reported similar emergent behavior in the assembly of communities (Niu*, et al.* 2017, Goldford*, et al.* 2018). Our results of compositional similarities between replicates on different carbohydrate mixtures further support the notion that closed (absence of immigration and dispersal) bacterial communities tend to converge towards a similar community

compositional structure under the same growth conditions (Goldford*, et al.* 2018). Future studies should investigate the impact of dispersal and immigration on assembly and structure of bacterial communities to have a better understanding of these processes in natural ecosystems.

Conclusions and future prospects

The aim of the present study was to design a Db-MM which mimics key ecological and physiological properties of the human intestinal microbiome, i.e. degradation of complex carbohydrates and production of beneficial SCFAs. We shortlisted a combination of nine bacterial strains representative of species that are highly prevalent across a range of high and low relative abundances in a large population. Further criteria for selection of the strains were based on some key metabolic features, and particularly their ability to breakdown complex carbohydrates, which are commonly found as fibers in the human diet. Among the complex carbohydrates selected in the present study, xylan and starch are major components of adult diets, whereas inulin is a prebiotic known for its potential to modulate the microbiota (Selvendran 1984, Macfarlane and Englyst 1986). Cellobiose was included as a disaccharide that ends up undigested in the colon due to resistance to disaccharidases of human small intestinal origin (Nakamura*, et al.* 2004). The fermentation of this substrate and its potential impact on the intestinal microbial community has received less attention, even though it is part of plant-derived dietary compounds (Ilhan*, et al.* 2017). Other key features of the human intestinal microbiome include the production of SCFAs such as acetate, butyrate, lactate, propionate, formate and other, branched chain SCFAs (Topping and Clifton 2001), functional redundancy and consequent competition (Li*, et al.* 2014, Coyte*, et al.* 2015, Levy and Borenstein 2013). We selected multiple bacterial strains at each of the main trophic levels in our system. At the level of primary degraders, we included *B. ovatus*, *B. xylanisolvens*, *E. rectale*, *R. intestinalis* and *E. siraeum* that all utilize starch (Sonnenburg*, et al.* 2010, Duncan*, et al.* 2002a, Flint*, et al.* 2012c, Sheridan*, et al.* 2016). Furthermore, *B. ovatus*, *B. xylanisolvens* and *R. intestinalis* are capable of utilizing xylan (Leth*, et al.* 2018, Sonnenburg*, et al.* 2010), whereas, *F. prausnitzi*, *E. siraeum*, *B. ovatus* and *B. xylanisolvens* can utilize inulin (Figure 5) (Duncan*, et al.* 2002b). Each of these strains had the potential to be the keystone strain

and influence the assembly of the community. At the level of secondary degraders, we included *E. hallii*, *S. variabile*, and Lachnospiraceae bacterium 7_1_58FAA (related to *Flavonifractor plautii*), which can utilize xylose, glucose and fructose as major breakdown products of xylan, starch and inulin, respectively. While *Bacteroides* strains are propionate producers, the strains belonging to the families Ruminococcaceae and Lachnospiraceae are known butyrate-producers. Hence, our Db-MM overall mimics the key features of functional redundancy, cooperation and competition while incorporating biologically relevant bacterial strains found to be highly prevalent in the human intestinal microbiome.

We observed that at each of the trophic levels, one species tended to dominate depending on the available substrate (Figure 9A and B). *B. xylanisolvens* demonstrated highest resource utilization potential reaching higher cell densities when grown in mono-culture with both starch-inulin and cellobiose-xylan. However, *E. rectale* grew faster and outcompeted *B. xylanisolvens* in the presence of starch-inulin in the mixed culture (Figure 9B). *E. rectale* reached highest cell density at 24h in mono-culture, and the observation of its dominance in the community growing on starch-inulin suggests an influence of lag times in determining identity and dominance of the key stone strain in the Db-MM. Whereas *B. xylanisolvens* grew after the decline of *E. rectale*, this was also the phase of transition from a highly uneven community composition dominated by *E. rectale* to a more even composition at 72h (Figure 9B). During this phase we observed an increased abundance of one of the secondary consumers, Lachnospiraceae bacterium 7_1_58FAA. This could be a consequence of the modification of the culture conditions due to production of lactate and production of free fructose from inulin by *Bacteroides* strains. *B. xylanisolvens* was the dominant strain at 72h which resulted in production of propionate. These results suggest the presence of active competition and complementarity *via* trophic metabolic interactions in the Db-MM.

The similar assembly and convergence of bacterial community composition and function between replicates in the presence of either starch-inulin or cellobiose-xylan at 72h demonstrated the deterministic role of carbohydrates on microbial community assembly. In addition, the low abundance of *R. intestinalis* after 24h showed that it is

competitively excluded by *E. rectale* in the presence of starch-inulin and by *B. xylanisolvens* with cellobiose-xylan.

Future studies on the Db-MM should incorporate metatranscriptomic analyses for better understanding changes in gene regulation in specific metabolic pathways during the Db-MM assembly process. Additionally, measuring of metabolites other than SCFAs will be crucial as potentially small molecules (*e.g.* those involved in quorum sensing) may play an important role in modulating the assembly process. Another important aspect is the growth medium composition. The human intestinal tract is a nutritionally rich environment supporting high densities of diverse microorganisms (Pereira and Berry 2017). The growth medium used in the present study was nutritionally rich in macro- and micro-nutrients, and included several vitamins and co-factors. Especially, vitamin B12 has been suggested to be a potential modulator of the microbial community in the human intestinal tract (Degnan*, et al.* 2014b, Belzer*, et al.* 2017). Thus, the syntrophic dependency (e.g. between B12-producers and non-producers) may not necessarily be mimicked in the growth media used in this study. However, the strain composition within the Db-MM described here allowed for investigating such potential syntrophic interactions due to the presence of multiple strains capable of *de novo* biosynthesis of vitamin B12 (*E. hallii*, Lachnospiraceae bacterium 7_1_58FAA) and non-producers and/or strains utilizing the salvage pathway and thus being dependent on external cobamides (*B. xylanisolvens*, *B. ovatus*) (Venema and Van den Abbeele 2013). Such investigations could be possible by changing the medium composition to enforce the influence of vitamins and co-factors of interest. Furthermore, the influence of bile acids on growth of the tolerant and non-tolerant strains will require further investigation. Our experiments were done under batch conditions *in-vitro*, and we suggest that future research should investigate the ecophysiology of Db-MM in *in-vitro* and/or *in-vivo* intestinal models. Moreover, the stability of the Db-MM in face of biotic and abiotic stressors should be investigated.

The Db-MM consists of important bacteria from two major phyla, Bacteroidetes and Firmicutes, but it lacks species coverage compared to the immense phylogenetic and functional diversity which exists in the human intestinal microbiome. Nevertheless, the Db-MM developed in this study is an important step towards a more comprehensive

design of the minimal microbiome, one that incorporates greater phylogenetic diversity and functional roles. In addition, strain level variation in fitness, colonization potential and metabolic capabilities is widespread in bacteria (Li*, et al.* 2016, Scholz*, et al.* 2016b). We suggest that future studies also investigate the impact of strain-level differences on microbial community assembly. The next step will be towards expanding the Db-MM to incorporate larger phylogenetic and functional diversity and a host mucosal component to create a host mucus- and diet-based minimal microbiome (MDb-MM). Potential candidate strains of special interest for MDb-MM are core mucus-degrading bacteria such as *Akkermansia muciniphila*, *Barnesiella intestinihominis* and *Bacteroides caccae* (Figure 1).

In conclusion, we have developed a butyrate- and propionate-producing Db-MM with key features such as functional redundancy, competition, and multiple trophic interactions. The rational selection of the strains was guided by the knowledge of their ecological and physiological features, integrating both top-down and bottom-up approaches. Future studies will be focused towards underpinning the molecular basis of bacterial interaction and community assembly to develop improved predictive models for the dynamics of the intestinal microbiome.

Acknowledgements

Supplementary data

Table S1: Carbon sources used for the different strains during pre-cultivation.

Sr. No.	Bacterial strain	Carbon sources
1	*Eubacterium rectale* (DSM 17629)	Glucose, 20 mM; lactose, 20 mM; xylose, 20 mM
2	*Bacteroides ovatus*	Galactose, 20 mM; xylose, 20 mM; starch, 0.2%
3	*Bacteroides xylanisolvens*	Galactose, 20 mM; xylose, 20 mM; starch, 0.2%
4	*Eubacterium hallii* (DSM 1736)	Glucose, 60 mM
5	*Eubacterium siraeum* (DSM 15702)	Glucose, 30 mM; maltose, 20 mM; starch, 0.1%
6	*Faecalibacterium prausnitzii* (A2-165)	Glucose, 30 mM; fructose, 30 mM
7	Lachnospiraceae bacterium 7_1_58FAA	Glucose, 20 mM; galactose, 20 mM; xylose, 20 mM
8	*Roseburia intestinalis* (DSM 14610)	Glucose, 30 mM; maltose, 20 mM
9	*Subdoligranulum variabile* (DSM 15176)	Glucose, 60 mM
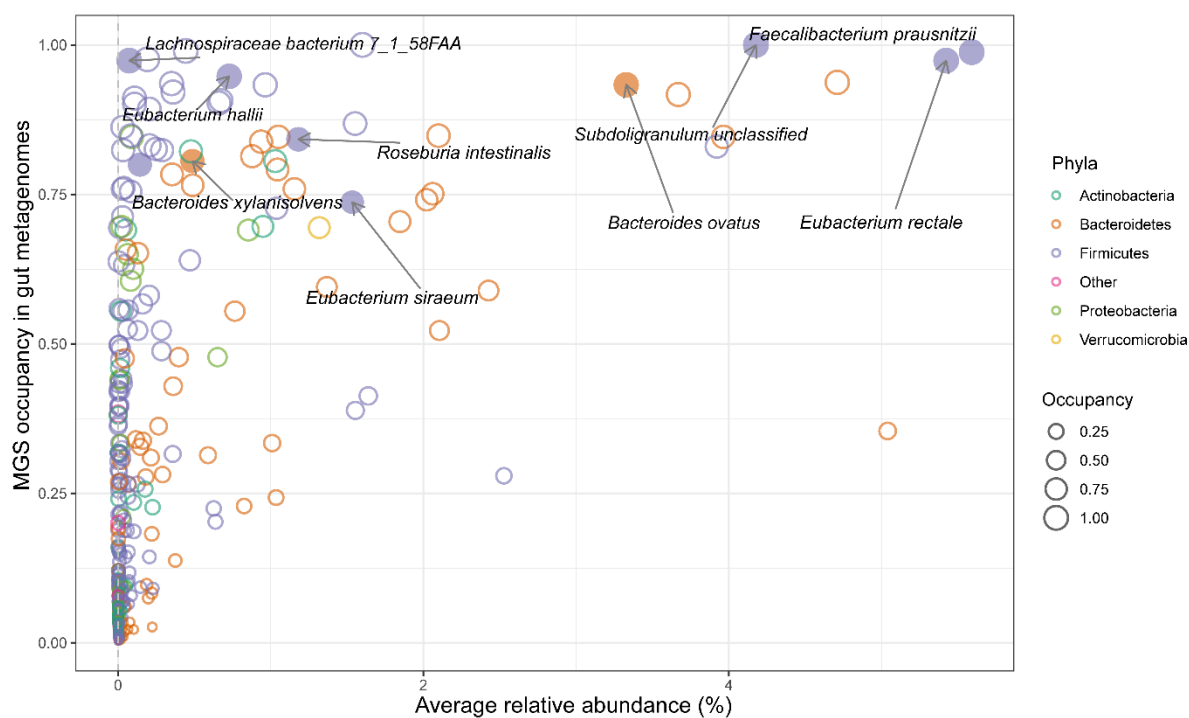
Figure S1: Occupancy-abundance relationship for metagenomic species identified in 494 human fecal metagenomic samples. The bacterial candidates for Db-MM are shown as closed circles.
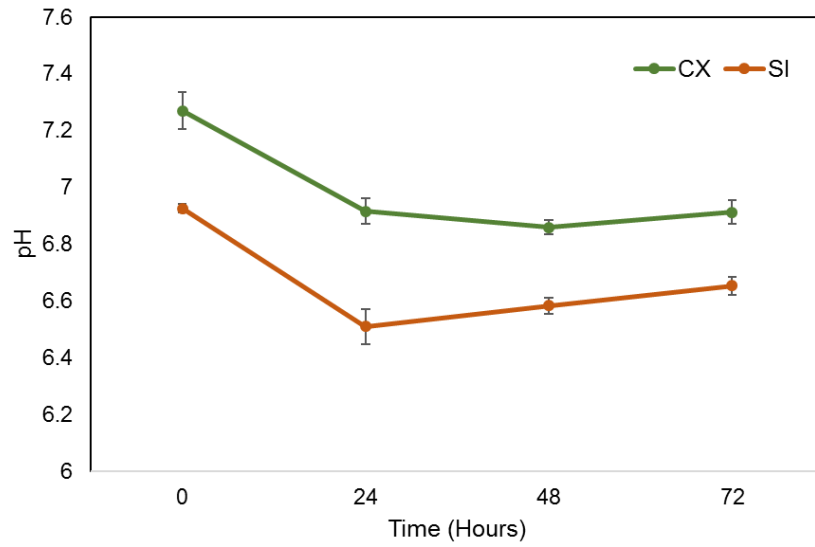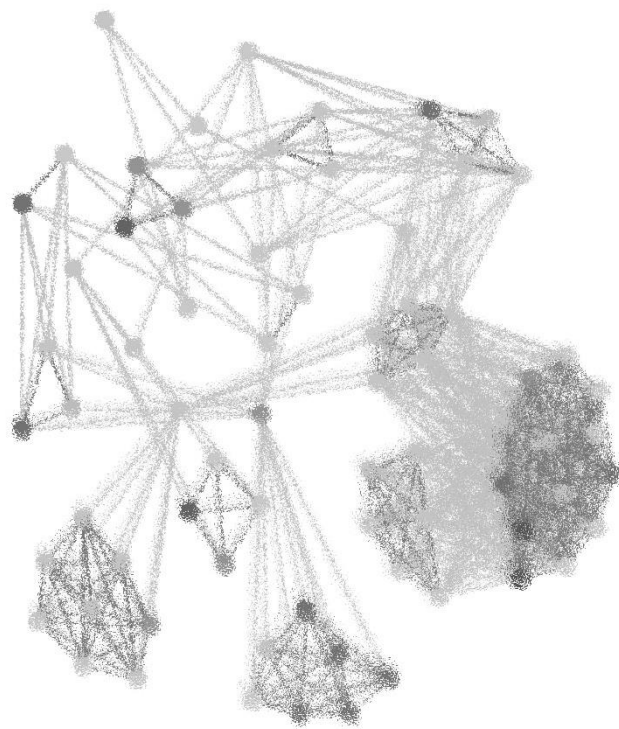
Figure S2: Changes in pH during growth of Db-MM on cellobiose-xylan (CX) and starch-inulin (SI).

Chapter 9

General Discussion

General discussion

The human intestinal microbiome is a thriving ecosystem with complex interactions between trillions of microbes. Therefore, understanding of this ecosystem requires an integration of ecology, physiology and genomics of its microbial members. Ecosystem level understanding is crucial for designing modulation strategies for the overall microbiome, while knowledge of the metabolic features of individual microorganisms is important for the development of specific microbial therapies. Here, the observations and findings of the research described in this thesis are discussed in detail. Furthermore, we highlight the challenges, opportunities and key areas that need attention for building a comprehensive understanding of the intestinal microbiome and its impact on human health.

Ecosystem level understanding of the human intestinal microbiome

Due to the immense impact of the human microbiome on health, efforts are underway to identify diagnostic biomarkers and discover new microbiome-inspired treatments. One of the important characteristics that is not to be neglected is that the microbiome is an ecosystem with more than 1000 cultured microbial species that have been officially described and deposited in culture collections and another few thousand species that have not yet been cultured and described (Ritari*, et al.* 2015, Rajilić-Stojanović and de Vos 2014). A better understanding of the ecological principles governing the compositional and functional stability and resilience of the microbiome is imperative for developing robust biomarkers and effective therapeutics (**Chapter 2**).

Co-existence of functionally redundant species

One of the major challenges in understanding the microbiome is the high variability in microbial community composition. Regardless, however, of the high individuality and variation, we observed that there is a phylogenetic core microbiota comprising 34 genus-like groups (**Chapter 2**, Figure 1, Table S1). This phylogenetic core consists of bacteria, which cover major functions such as a breakdown of complex polysaccharides present in our diet and mucus, and fermentation of simple carbon sources. We observed a high functional redundancy within the phylogenetic core, which may play a

crucial role in stabilizing the ecosystem. The high functional redundancy also suggests high competition at every level of the trophic food chain.
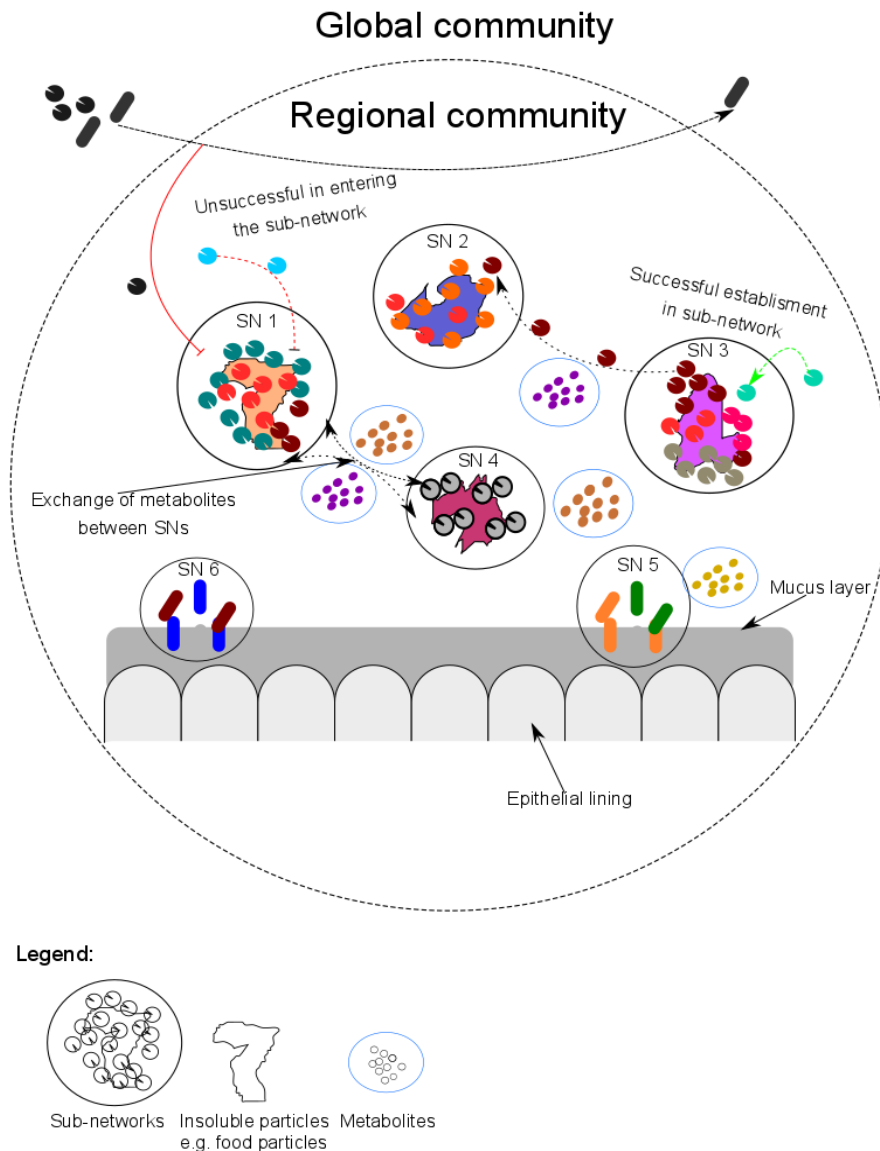

Figure 1: Multi-level organization from global communities to micro-scale sub-networks of microbes in the intestinal lumen and at the mucus layer. In the colon, the luminal microbial community is relatively dense compared to the mucosal community. Bacteria can form aggregates where complimentary bacteria can co-exist. Some may have multiple species (lumen: SN1, 2, 3 and mucus SN5, 6) or some may be dominated by single highly specialized and efficient species (SN 4). In addition, there can be exchange of metabolites between the sub-networks (SN1 and SN4). In case there are specific niches created and unoccupied there can be successful establishment of a species from the meta-community (migration). Alternatively, there can be competitive exclusion of a species (SN1). In some cases, there can be transfer of species from one sub-network (SN3) e.g. mucosal to another (SN2) e.g. luminal (dispersal). Both dispersal and migration result in higher species diversity in sub-communities.

Core bacteria such as *Ruminococcus bromii*, *Bacteroides ovatus*, *Roseburia intestinalis*, and *Eubacterium rectale* have the ability to breakdown complex carbohydrates, consequently giving rise to high competition (Ze*, et al.* 2012, Cockburn*, et al.* 2015, Sheridan*, et al.* 2016). Fecal samples are widely used for microbiota profiling, and the processing involves homogenization before DNA extraction (Salonen*, et al.* 2010, Costea*, et al.* 2017b). Hence, whereas homogenization purposely leads to representative sampling of the overall microbiota, it undermines our ability to obtain clues with respect spatial or micro-scale ecological networks. In an adult gnotobiotic mice model, local aggregates of one or few bacteria have been demonstrated (Welch*, et al.* 2017). Such spatial organization may depend on the affinity of specific bacterial strains for growth substrates and choice of complementary species. This may give rise to assembly of local and functional sub-networks of complimentary bacteria. A regional community can be comprised of several functional guilds, which may or may not carry out similar functions. Investigating such ecological processes remain a major task in human intestinal microbiome research that is technically challenging.

Importance of the few and why they dominate?

An intriguing observation when investigating microbial communities is the observation of right-skewed distribution of microbial taxa where only few seem to dominate (Lynch and Neufeld 2015). Species abundance distribution (SAD) models are widely used to investigate the patterns of commonness and rarity of species in any given macro- and micro-ecosystem (McGill*, et al.* 2007, Shoemaker*, et al.* 2017). Recently, the SADs for microbial communities were demonstrated to fit well with a log-normal distribution model (Shoemaker*, et al.* 2017). Log-normal distribution, as the name suggests, is a distribution of frequency that becomes approximately normal when it is log transformed. While it is widely accepted that dominance is common, the extent to which few species can account for a major fraction of the total community was unknown. In **chapter 3**, we investigated the phenomenon of hyperdominance, which was previously observed in tree communities in the Amazon forest (ter Steege*, et al.* 2013). Hyperdominance is defined here as the fraction of species that account for 50% of the total biomass in a given ecosystem. We adapted this measure to quantify the extent of hyperdominance in diverse microbial communities. Since biomass cannot be

directly correlated with bacterial abundances based on 16S ribosomal RNA (16S rRNA) gene surveys, we define hyperdominance as a fraction of amplicon sequence variants (ASVs) that account for a cumulative relative abundance of 50%. We observed that on average 4.6% of the total ASVs account for 50% of the community across a broad range of different environments (**Chapter 3**). Investigation of human intestinal metagenomic datasets also confirmed our theory of hyperdominance as on average 5.5% of the metagenomic species accounted for a cumulative relative abundance of 50% (**Chapter 3**, Figure 3). We observed that many of the microbial taxa related to core genus-like groups identified in **chapter 2** are hyperdominant according to this definition (**Chapter 3**, Table 1). Finally, analysis of time series data from the intestinal microbial community of an individual provided information regarding the dynamics of the hyperdominant core taxa (**Chapter 3**, Figure 4).

While hyperdominant bacteria such as *Faecalibacterium* showed least variation, others such as *Coprococcus* and *Haemophilus* revealed high variation in abundances over time (**Chapter 3**, Figure 4, Table 2). The latter represent the conditionally rare taxa (CRTs) in the community. CRTs are defined as those taxa which have periodic instances of high relative abundances in the community (Shade*, et al.* 2014). In addition, we identified some core ASVs which are hyperdominant and conditionally rare. These ASVs exhibit bimodal distribution. Previously, it was suggested that bacteria that have bimodal abundances can be potential biomarkers for observed resilience of the microbiome (Lahti*, et al.* 2014). Bacteria with bimodal abundance have been shown to represent tipping elements with alternative stable states of low and high abundance (Lahti*, et al.* 2014). These conditionally hyperdominant taxa could be important indicator taxa for identifying potential shifts from one alternative stable state to another (**Chapter 2**). In-depth investigation of these taxa and their abundance states could provide opportunities for tuning intervention strategies aimed at modulating the microbiome towards a preferred stable state (**Chapter 2 and 3**).

Another intriguing observation was the high functional redundancy amongst the hyperdominant taxa (**Chapter 3**, Table 1). *Ruminococcus bromii*, *Eubacterium rectale*, *Bacteroides ovatus* and *B. uniformis* are known degraders of complex polysaccharides (Sonnenburg*, et al.* 2010, Flint*, et al.* 2012a, Cockburn*, et al.* 2015, Sheridan*, et al.*

2016). These bacteria can thus reach high densities as they occupy a potentially abundant niche in the ecosystem. We also observed population specific hyperdominant taxa (**chapter 3**, Table 1). In the US American population, *Alistipes putredinis*, *B. ovatus*, and *B. uniformis* were identified as the most prevalent hyperdominant taxa, whereas the European populations were found to harbor members of the families Lachnospiraceae and Ruminococcaceae as hyperdominant. These differences could most likely be a consequence of differences in diet between the two populations.

Inclusive model for disentangling microbial community assembly and stability

Investigating the processes that govern assembly of microbial communities that have hyperdominance and rarity as a fundamental characteristic is crucial towards our understanding of its functioning, resilience and stability. A better knowledge of these processes will help in efficient modulation and restoration of the intestinal microbiome, including the design of minimal microbiomes to support human health. For a long time, the assembly and dynamics of communities was explained by two contrasting theoretical frameworks, namely the niche and neutral theories. The niche theory suggests a role for availability of a given niche and its size to explain assembly, composition and structure of a community. In contrast, the neutral theory suggests that assembly of communities occurs through random dispersal, speciation, and extinction events (Rosindell*, et al.* 2011, Vandermeer 1972). However, recently a more inclusive theory was proposed, the so-called "theory of community ecology" wherein processes such as selection, speciation, drift and dispersal are considered across spatial scales (Vellend 2010). This theory resonates well with the uniqueness of individual microbes, which demonstrate strain level variation in growth dynamics. Periodic selection of ecotypes (strains using the same ecological niche) and lower constraints on speciation are two examples with broader implications in understanding microbial community ecology (Cohan 2005, Cohan 2001). For understanding the assembly, composition and structure of microbial communities it is important to view it at various levels. We applied the conceptual framework for the theory of community ecology proposed by Vellend, et. al., (2010) by integrating several intestinal microbiome specific features (Figure 1). In addition to the role of neutral processes for observed hyperdominance and rarity as suggested in **chapter 3**, disentangling the

ecophysiological processes at different spatial scales may also explain these features. Considering a micro-scale sub-community as the lowest level in a given ecosystem, the presence of different bacterial species, and thus the richness of a sub-community will be defined by availability of specific niches (selection). However, which specific species (or even ecotypes) assemble into a sub-community will depend on priority effects. For example, which of the many potential primary degraders of a complex substrate such as starch occupies the abundant niche in a specific sub-community can be explained by the stochastic niche theory (see (Tilman 2004)). The dynamic changes within the sub-community will be governed by immigration and dispersal events at a higher level of a regional community. The species that have been excluded *via* competition may still persist in the regional community provided they have higher survivability, for example *via* sporulation (Kearney*, et al.* 2018). This particular characteristic is of interest for bacterial communities, as bacteria are known to persist for long periods in states of dormancy and contribute to diversity (Jones and Lennon 2010). The presence of functionally redundant species in the regional community may act as an insurance to maintain stability of sub-communities when there is loss of one or more species (Yachi and Loreau 1999). At the level of the regional community, evenness or unevenness in the number and type of niches will determine the most dominant sub-communities and subsequently the dominant species. For example, if starch is the most dominant component in an individual's diet, then a niche for starch degradation will be the most abundant, and consequently sub-communities with resulting starch-associated trophic chains will be dominant. At the highest level, *i.e.* at global community level, the influx of species through the oral route, i.e. *via* food, water, etc., providing a trickle of species from the small intestine, will result in several transient species in the regional community. Events of catastrophic destruction such as antibiotic treatment, may lead to breakdown of the sub-communities. This may give rise to successful establishment of transient species into the sub-communities causing a ripple effect ultimately destabilizing the regional community. This may also be the case for potentially pathogenic bacteria, such as *Clostridioides difficile*, which may be present as dormant spores in the regional or global community. However, during conditions favoring its growth, it can successfully establish itself in several sub-communities, thereby

destabilizing the community. When we incorporate such a multi-level approach, it seems plausible that both niche and neutral theory are playing a role in community dynamics, and also explain the observed hyperdominance and rarity. Hyperdominant are those bacteria that occupy the most abundant niche and are part of the associated trophic chain. Rare bacteria are either those that are excluded from sub-communities and are potentially dormant or those that are part of sub-communities occupying rare niches. For example, in **chapter 3**, we observed that bacteria occupying the niche for sulfate reduction are consistently rare in the human intestinal microbiome (**Chapter 3**). Recently, it was also suggested that different types of rarity can be explained across a continuum of deterministic and stochastic processes (Jia*, et al.* 2018). Testing the theory mentioned above is challenging in natural communities because of the multitude of variables and lack of understanding of metabolic traits of individual species. To address this limitation, development of minimal microbial communities has recently received considerable attention (De Roy*, et al.* 2014, Großkopf and Soyer 2014). To this end, we define a minimal microbiome as a consortium of well-studied microbes, which as a community mimics ecological properties of natural microbiomes and cover(s) one or more metabolic pathways which may not be completed by a single species. These well-defined minimal microbial communities can be useful for disentangling the role of different ecological processes on community assembly and functioning.

Ecophysiology of key human intestinal bacteria

Whereas in macro-ecosystems functional trait based ecological studies are common, such studies are lacking in microbial ecology. One of the major challenges is the strain-level variation and lack of studies focusing on characterizing physiology of already cultured bacteria. Due to the success of recent culturing efforts, it has been suggested that more than 70% of the microbial species present in the human intestinal tract have been cultured in the laboratory though most of them are yet to be characterized and validly described (Lagier*, et al.* 2016). Even more so, only few have been investigated in detail as highlighted in **chapter 2**. In **chapter 5, 6 and 7**, we focused on studying the physiology of selected key butyrate- and propionate-producing bacteria, including

Intestinimonas butyriciproducens and two *Eubacterium hallii* strains, which we recently reclassified as *Anaerobutyricum hallii and Anaerobutyricum soehngenii* (**Chapter 6**)*.*

Impact of isolation source on ecology and physiology of intestinal bacteria

In **chapter 5**, we focused on investigating the ecophysiology of *I. butyriciproducens*. Gnotobiotic mice are routinely used as *in-vivo* models for mechanistic investigation of the role played by microbes in pathophysiology of diseases and health benefits conferred by the microbe. However, the origin of bacterial strains can have a crucial effect on the applicability of such *in-vivo* studies since the overlap between human and mice microbiomes is limited (Xiao*, et al.* 2015). Therefore, it is important to investigate the genomic, physiological and ecological differences between bacterial strains originating from human and mice. A meta-analysis of the 16S rRNA gene sequences in publically available databases revealed the presence of three major species groups of *Intestinimonas* (**Chapter 5**). One of the 16S rRNA gene sequences representing *I. butyriciproducens*-like species represented 90% of the sequences originating from samples of human origin. The other major group of sequences originated mostly from murine intestinal samples (**Chapter 5**). Previously, two strains of *I. butyriciproducens*, SRB521 and AF211 were isolated from a mouse and human host, respectively (Bui*, et al.* 2016, Kläring*, et al.* 2013). We comprehensively compared both strains to better understand host associated adaptive features. Strain level comparison of bacteria originating from different hosts can provide key insights regarding the adaptive evolution and selective forces that may lead to functional diversification. The intestinal tracts of mice and humans have different anatomy and different environmental features which can influence growth. We observed that AF211 was better adapted to the presence of bile in the growth medium as compared to SRB521, and the cellular fatty acid composition was remarkably different between the two strains. Niche occupation and adaptation by bacteria is often associated with chromosomal remodeling (Eisen*, et al.* 2000). The genomic architecture of AF211 and SRB521 demonstrates symmetric chromosomal inversion, but potential implications of this genome reorganization remain unknown (**Chapter 5**, Figure 3). Since, the two strains exist in two different microbiomes, they are exposed to different microbial partners. One of the common features of bacterial diversification is *via* their ability to

horizontally acquire genes, thus creating a flexible pool of genetic information (Wiedenbeck and Cohan 2011). The majority of the horizontally transferred genes can be found in genomic islands, and we observed pronounced differences in the location and content of genomic islands between the two strains. These observations further support the role of the host-associated microbiome in driving genomic diversification between different strains of the same species. Most of the genes identified in these genomic islands were annotated as coding for hypothetical proteins, which should be targets for future in-depth functional characterization to determine their role in fitness potential of the strains to specific host microbiomes. One way of shortlisting the targets among the hundreds of hypothetical proteins encoding genes would be to study the global expressed proteome of the two strains to identify the actively expressed hypothetical proteins. These actively expressed hypothetical proteins can then be further characterized by heterologous over-expression in suitable bacterial model organisms, purification and functional assays.

Unravelling the active metabolic pathways in *Anaerobutyricum* species

Analysis of genomic data can reveal valuable information regarding species diversification, adaptation and functional potential (Eisen*, et al.* 2000, Kankainen*, et al.* 2009, Bui*, et al.* 2016, Sun*, et al.* 2015, Zhao*, et al.* 2017). Recently, genome sequence data has been harnessed to address a major challenge in prokaryotic taxonomy (Parks*, et al.* 2018). This study used phylogeny inferred from 120 ubiquitous single-copy proteins to suggest reclassification of 58% of a total of 94,759 publicly available genomes (Parks*, et al.* 2018). The information on phenotypic traits resulting in major diversification in species is likely to be stored in the sequence data other than the 16S rRNA genes that are widely used to inform phylogeny and taxonomic classification (Van Belkum*, et al.* 2001, Parks*, et al.* 2018). Average Nucleotide Identity (ANI) and Genome-to-Genome Distance Calculations (digital DNA-DNA hybridization, dDDH) are powerful approaches for differentiating closely related species (Goris*, et al.* 2007, Meier-Kolthoff*, et al.* 2013, Chun*, et al.* 2018). We used these two approaches together with a variety of functional tests to check for similarity between two strains of *Eubacterium hallii*, DSM 3353 and DSM 17630 (L2-7) and observed that the two strains are in fact different species. This led to a more detailed genome guided polyphasic taxonomic

study, which confirmed the novelty of strain L2-7 (**Chapter 6**). Investigation of the 16S rRNA gene phylogeny confirmed the previous finding of misclassification of *Eubacterium hallii* and the need to reclassify and correct its nomenclature (Wade 2006). We proposed a novel genus, *Anaerobutyricum,* to accommodate the two species *Anaerobutyricum hallii* (DSM 3353^T) and *Anaerobutyricum soehngenii* DSM 17630^T (L2-7^T). This reclassification will have a major impact on human intestinal microbiome research, which largely depends on short read based microbial community profiling. The analysis is commonly done at a coarse taxonomic resolution, where text-based merging at genus level, for example merging counts of all OTUs that are classified as genus *Eubacterium*, is done because species level classification is doubtful due to short read lengths. Since the genus *Eubacterium* is polyphyletic, this leads to incorrect observations and subsequent interpretation. However, at least for a key core species of the human intestinal tract, *Anaerobutyricum hallii*, we have now provided a clear taxonomic delineation, which is supported by both traditional and recent genome-based taxonomy (**Chapter 6**). This reclassification is congruent with the recently proposed genome-based taxonomy (Parks*, et al.* 2018). Similar studies are required for other bacterial genera within the order Clostridiales, which will lead to improvements in the accuracy of molecular microbial profiling studies.

The findings from the taxonomic analysis guided the research described in **chapter 7** of this thesis, where we observed that the ability to utilize sucrose was one of the differentiating characteristics of *A. soehngenii* strain L2-7. The other three important metabolites of interest were 1,2 propanediol (1,2 PD) which is a cross-feeding metabolite resulting from utilization of fucose and rhamnose by bacteria such as *Akkermansia mucniphila*, D- sorbitol widely used as a non-glucose sweetener and DL-lactate, which is an end product of fermentation by primary degraders such as *Bifidobacterium* (De Vuyst and Leroy 2011, Scott*, et al.* 2014) . We first started with investigating the proteomic response during three different growth stages, namely early, mid and late exponential growth phase, to characterize the global proteome expression of *A. soehngenii*. This approach allowed us to capture the abundant proteins involved in metabolic response to nutrient availability. The early exponential phase allowed us to identify proteins related to transport and regulation. For example,

ribonuclease III was a significantly abundant protein in the early growth phase (**Chapter 7**, Figure 1). This protein is involved in post-transcriptional regulation and protein modification (Deutscher 2015). We observed that the mid-exponential stage was characterized by production of high amounts of proteins involved in central carbohydrate metabolism and butyrogenesis. Finally, in the late exponential phase, *A. soehngenii* produced high amounts of proteins involved in protein degradation. Interestingly, during routine laboratory cell growth experiment and cell passages, *A. soehngenii* cells showed a high degree of autolysis. The global proteome suggests that *A. soehngenii* produces high amounts of serine-type D-Ala-D-Ala carboxypeptidase, which is a known autolysin (Johnson*, et al.* 2013). Various questions arise here, such as, why do these cells undergo autolysis, especially in competitive environments such as the intestinal tract, and what can be the advantage for these bacteria? While the importance of programmed cell death has been well studied in eukaryotic cells, this phenomenon remains understudied for prokaryotic cells (Rice and Bayles 2008). One of the hypotheses suggests that autolysis will lead to recycling of nutrients and can be advantageous for bacterial cells (Rice and Bayles 2008). Another hypothesis suggests that autolysis may play an important role in separating two bacterial cells (Forsberg and Rogers 1971). Interestingly, the serine-type D-Ala-D-Ala carboxypeptidase was produced in both mid and late exponential stage suggesting possible production of these proteins in actively growing cells of L2-7. In addition, it has been reported that autolysis can play an important role in development and dispersal of biofilm formation in host associated *Pseudoalteromonas tunicata* (Mai-Prochnow*, et al.* 2006). Autolysis in bacteria from the intestinal tract warrants attention in future studies as it may have implications for fitness of bacteria in the human intestinal tract.

Global expressed proteome data was used for reconstructing the acetyl-CoA pathway for butyrogenesis based on the actively produced proteins (**Chapter 7**). In addition, this led to identification of proteins involved in 1,2 PD utilization, microcompartment formation and vitamin B12 biosynthesis. A combination of genomic and proteomic analyses assisted in identifying a unique gene cluster that suggested adaptation of *Anaerobutyricum soehngenii* to efficiently utilize DL-lactate. DL-lactate provides a special niche in the human intestinal tract. Lactic acid has a low pKa as compared to

acetate, butyrate and propionate. Therefore, accumulation of lactic acid results in highly acidic conditions, which can have detrimental effects on the host as well as the surrounding microbial community. Lactate-utilizing bacteria play a major role in maintaining the physiological pH necessary for growth of non-acidophilic bacteria (Slyter 1976, Duncan, *et al.* 2004).

Table 1: Calculated Gibbs` free energy for conversion of glucose to pyruvate and lactate to pyruvate.

Reaction	ΔG (physiological conditions)
Equation 1: $$\text{Glucose} + 2\text{Pi} + 2\text{ADP} + 2\text{NAD}^+ \rightarrow 2\text{Pyruvate} + 2\text{ATP} + 2\text{NADH} + 2\text{H}^+ + 2\text{H}_2\text{O}$$	-88 kJ/mol
Equation 2a: $$\text{Lactate} + \text{NAD}^+ \leftrightarrow \text{Pyruvate} + \text{NADH}$$ *Equation 2b* as proposed by (Weghoff, *et al.* 2015): $$\text{Lactate} + \text{Fd}^{2-} + 2\text{NAD}^+ \rightarrow \text{Pyruvate} + \text{Fd} + 2\text{NADH}$$	+25 kJ/mol -9.5 kJ/mol

Unlike growth on glucose (equation 1), growth on lactate poses a major energetic barrier because it is an energy dependent process with low energy yield (equation 2a). The free energy resulting from the coupling with oxidation of ferredoxin is -9.5 kJ/mol, see equation 2b (Weghoff, *et al.* 2015). Only few bacteria residing in the human intestinal tract are known to utilize lactate, and even fewer have been reported to be capable of utilizing both enantiomers, D- and L-lactate, to produce butyrate (Duncan, *et al.* 2004). *Anaerostipes caccae*, *A*. *rhamnosivorans* and the two *Anaerobutyricum* species *A. soehngenii* and *A. hallii* are capable of converting DL-lactate plus acetate to butyrate. Other common bacteria such as *Faecalibacterium prausnitzii*, *Roseburia intestinalis*, *Agathobacter rectalis* are known to consume acetate (Duncan and Flint 2008, Duncan, *et al.* 2002a, Heinken, *et al.* 2014a). As a consequence, a community with co-occurrence of acetate utilizers will have high competition for acetate, which poses another challenge for the efficient utilization of lactate (along with low energy) by *Anaerobutyricum* and *Anaerostipes* spp. The initial analysis of the genome of strain

L2-7 identified a lactate dehydrogenase (EHLA_0807, EHLA_2279, EHLA_2826, EHLA_3075) and no lactate racemase. However, comparison of the proteome in cultures grown on lactate and other carbons sources, sucrose and sorbitol, revealed a cluster of highly abundant proteins, which were initially annotated as CO FAD dehydrogenase (EHLA_0974) and protein of unknown function (EHLA_0978). Further investigation of these proteins across databases and analysis of protein active domains revealed that EHLA_0974 is a unique class of lactate dehydrogenase (LDH) closely related to glyolate dehydrogenase. Downstream of the gene encoding this LDH are two genes coding for electron transport proteins, EtfB and EtfA (EHLA_0975, EHLA_0976 respectively). This unique organization of genes encoding LDH, EtfB, EtfA, along with a nickel dependent racemase (EHLA_0978) and a permease, was first described in 2015 by (Weghoff, *et al.* 2015). Extensive search of more than 50,000 genomes revealed that this gene cluster is highly conserved in *Anaerobutyricum soehngenii, Anaerobutyricum hallii* and *Anaerostipes caccae* among the butyrate producers from the human intestinal tract. This unique genomic organization suggests that both *Anaerobutyricum* species and *Anaerostipes caccae* have adapted to a lifestyle involving lactate plus acetate utilization in the human intestinal tract.

A. soehngenii also encodes an inducible transport protein for sorbitol utilization and a sorbitol 6-dehydrogenase which converts sorbitol to fructose and glucose 6 phosphate, which then enters glycolysis. Similarly, for sucrose, we identified a potentially novel transporter with the sugar PTS that is involved in the utilization of sucrose. This particular gene along with glycosyl hydrolases GH_77 family was not identified in the genome of another species from this genus, *A. hallii*, explaining the inability of this species to utilize sucrose **(Chapter 6).** Overall, using a combination of basic physiological, genomic and proteomic study we were able to reconstruct the major metabolic pathways in *A. soehngenii*. The improved understanding of metabolism of bacteria has several implications, as we are able to understand how these bacteria can grow better than others by using DL- lactate.

Future research should focus on unravelling the active metabolic pathways for other core bacteria such as *Coprococcus catus*, Lachnospiraceae bacterium 7_1_58FAA, *Subdoligranulum variabile*, *Ruminococcus* sp. 5_1_39BFAA and other species identified

in **chapter 2 and 8**. Mono-culture based investigations are a stepping stone for informed design of co-cultures or multi species communities and provide information for building predictive models of community assembly dynamics (D'hoe*, et al.* 2018, Venturelli*, et al.* 2018).

Putting it together: reconstructing functional networks of the human intestinal microbiome

Conceptual understanding for the design of minimal microbiomes

As discussed in the previous sections, complexity in the microbial world is at every level, ranging from entire ecosystems to the individual microbial cell. Complex ecological processes determine the successful assembly of microbial communities, and thermodynamic constraints, metabolic pathways and regulatory circuits play a major role in successful survival and propagation at the level of individual microbial cells (Scheffer*, et al.* 2018, Brown*, et al.* 2004). Therefore, integrating these features in top-down and bottom-up approaches for the design of minimal microbiomes is essential. The latter approach would involve understanding the metabolic roles played by each of the bacteria identified in the human intestinal microbiome. The size and complexity of a minimal microbiome can be tuned to address two main broadly defined aims, i.e. 1) unravelling metabolic interactions, 2) investigating key ecological concepts. For example, *Anaerobutyricum* and related species have the ability to utilize DL- lactate and acetate to form butyrate. Lactate is produced as a result of fermentation and breakdown of polysaccharides by bacteria such as *Eubacterium rectale* or *Bifidobacterium* spp. Thus, in order to understand trophic metabolic interactions addressing polysaccharide degradation and butyrate production, two-species systems have been used (De Vuyst and Leroy 2011, Scott*, et al.* 2014). Similarly, trophic metabolic interactions between mucus degraders and butyrate producers have been studied using two-species systems (Belzer*, et al.* 2017). Such two-species systems can be upgraded to incorporate ecosystem processes of competition, including e.g. two competing polysaccharide degraders, and two butyrate producers competing for polysaccharide breakdown products. Such 4-species cultures can be used to investigate pairwise species competition/complementarity as well as metabolic inter-

dependencies. To address specific ecological concepts, the design should aim at higher complexity to more comprehensively mimic the human intestinal microbiome. For example, to investigate the effect of functional redundancy on community assembly, selection of bacterial species that have functional overlap at different trophic levels will be crucial. Conceptual understanding of both microbial physiology and ecology applied in synergy will help in rational design of appropriate minimal microbiomes. Ecophysiology guided approaches that incorporate the knowledge of physiology, metabolic potential of each species with their ecological roles, and properties such as prevalence, dominance and rarity will be important to create microbiomes that mimic natural ecosystems.

Reconstructing a Diet-based Minimal Microbiome (Db-MM)

Diet is a major deterministic factor that governs the assembly of intestinal microbial communities, including their composition, structure and function (Gibson 1999, De Filippo, *et al.* 2010, Claesson, *et al.* 2012, David, *et al.* 2014, Salonen, *et al.* 2014). Therefore, efforts towards a mechanistic understanding of the effect of dietary components on intestinal microbiota is important and will aid in designing microbiome modulation strategies (Gibson 1999, O'Keefe, *et al.* 2015, Shetty, *et al.* 2017a). Most studies have focused on utilizing fecal inocula for investigating the effect of different dietary fibers on microbial community composition, diversity and functions (Kovatcheva-Datchary, *et al.* 2009, Scott, *et al.* 2014). A major challenge associated with such studies is the presence of several unknown bacterial species, presence of bacteriophages, and high variability across different sample inocula. Additionally, most currently employed sequencing based molecular techniques are incapable of species/strain level variation with high confidence, while design and application of qPCR primers for several species is technically challenging and expensive (Ramiro-Garcia, *et al.* 2016, Edgar 2017). Strain-level resolution can be obtained from shotgun metagenome sequencing data, albeit limited to the top 0.1% of the microbes in the total community and at higher cost (Scholz, *et al.* 2016a, Truong, *et al.* 2017, Li, *et al.* 2016). For better understanding of a complex systems such as the human intestinal microbiome, a pragmatic approach would be to study the ecosystem in parts under well-controlled conditions. Studying these defined microbial communities could form

a promising avenue where major properties such as known species composition and their genetic potential (sequenced genome), as well as known general physiological characteristics can be leveraged to better understand compositional and functional community dynamics. A proof of concept study that followed the reasoning mentioned above (and see *Conceptual understanding for design of minimal microbiomes*) is described in **chapter 8**. We incorporated existing knowledge of physiology of individual species and their ecological characteristics of dominance, prevalence and rarity to design a diet-based butyrate- and propionate- producing minimal microbiome. In contrast to a mucus based minimal microbiome, the diet-based minimal microbiome (Db-MM) was designed to reconstruct metabolic interactions starting from the breakdown of common dietary compounds leading to production of the health promoting metabolites butyrate and propionate as major end products (Figure 2).
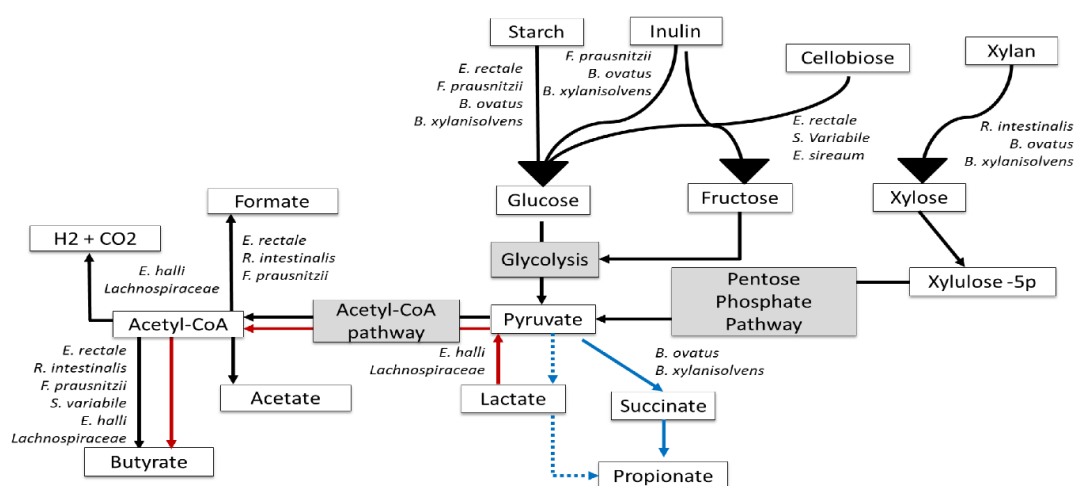

Figure 2: Major metabolic interaction routes from selected dietary carbohydrates towards production of butyrate and propionate that are incorporated in the diet based minimal microbiome.

We selected representative core bacteria with known key metabolic features to demonstrate deterministic community assembly, as well as compositional and functional dynamics. The community assembly was strongly driven by the available dietary compound. Additionally, when we analyzed the fraction of species and their contribution to relative abundance, we observed qualitatively conserved trends, where a similar number of species accounted for a similar cumulative relative abundance at specific time points (Figure 3). The co-existence of 6 out of 9 of the community

members contributing high proportions of the final community suggests that metabolic inter-dependencies and availability of diverse substrates avoids mono-dominance even in absence of external forces such as immigration. Additionally, comparison of the initial and end point relative abundances clearly showed differential responses of key polysaccharide degrading species to different substrates and consequently the responding dependent species (Figure 4).
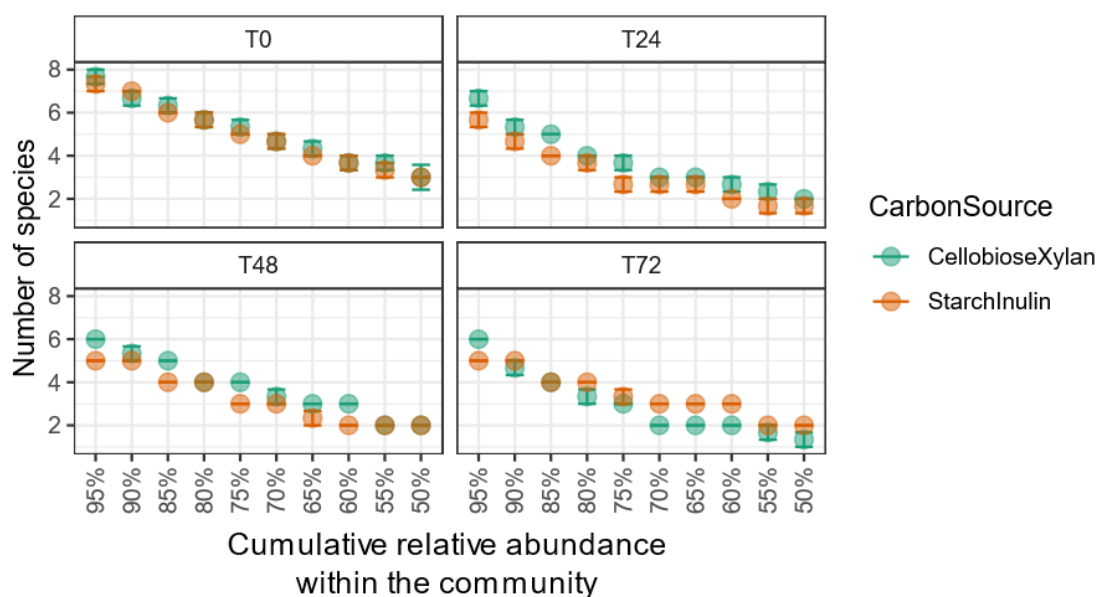

Figure 3: Number of species that account for a given cumulative relative abundance within the community. The panel heading T0, T24, T48, T72 represent sampling time points at 0 h, 24 h, 48 h, and 72 h, respectively. While initially at T0, 7-8 species (out of the total 9) accounted for 95% cumulative relative abundance, at the end of the experiment 5-6 species were co-existing to account for 95% cumulative relative abundance. The error bars represent standard error of triplicate values.

The Db-MM community consisted of key species that can use substrates that are commonly present in the human diet as well as a known prebiotic such as inulin. We used equal amounts of each of the substrates in our experiments. However, it would be interesting to investigate in future studies the effect of varying ratios of the different substrates on determining the community composition and functional output of core intestinal bacterial communities. The initial abundance of *B. xylanisolvens* and *B. ovatus* was higher than that of *E. rectale* in the presence of starch. However, *E. rectale* dominated the cultures with starch-inulin at all other time points, indicating that *E. rectale* is a specialized starch degrader, and initial abundances may not necessarily be

advantageous. Future experiments should incorporate transcriptome analyses to better understand how these species recruit their molecular machineries and regulate gene expression, as well as detailed metabolomics characterization.
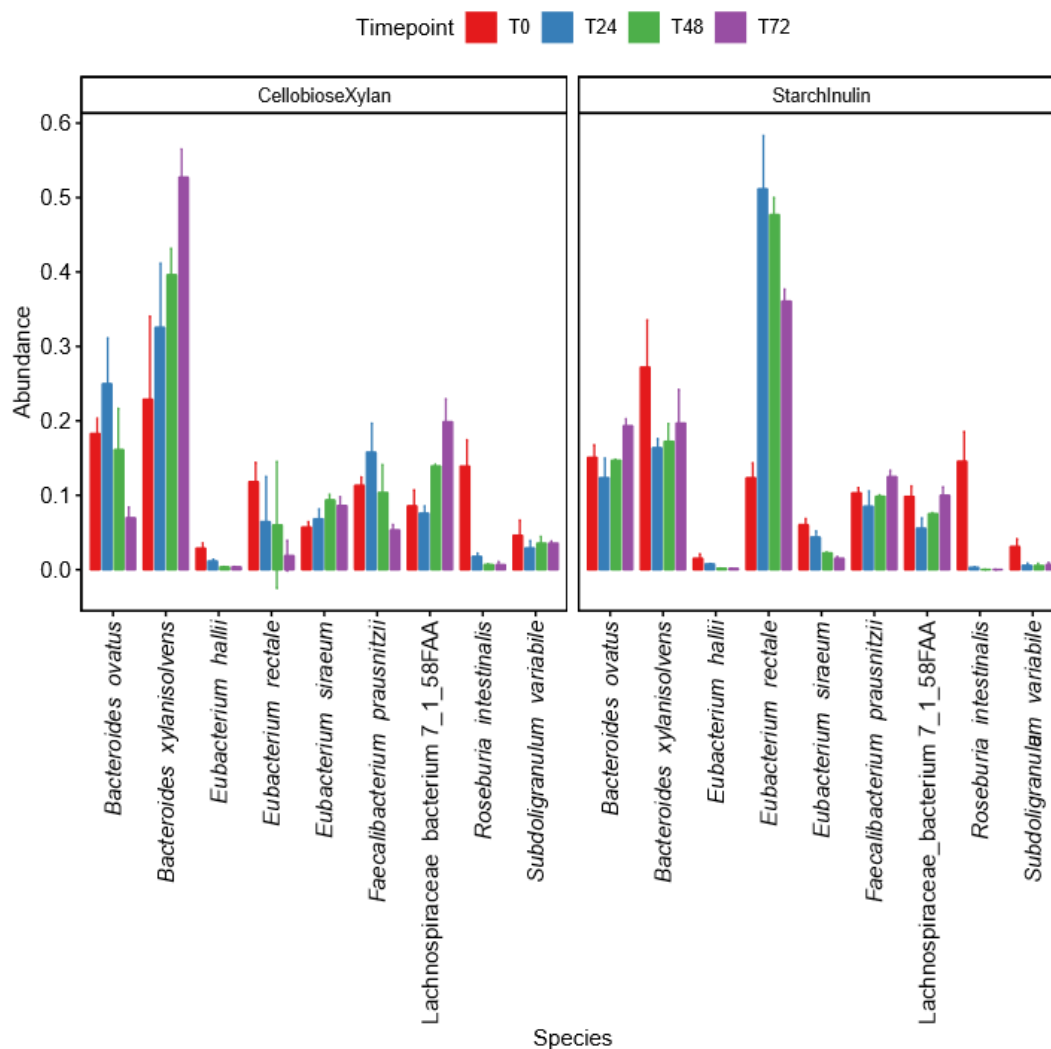

Figure 4: Variation in relative abundance of individual species when grown in consortia. The x-axis represents relative abundance (proportion). The error bars represent standard deviation of triplicate values.

Recently, in a 3-species synthetic gut community, incorporating transcriptomics allowed for identifying emergent behaviour (D'hoe*, et al.* 2018). In a 12-member synthetic gut community, analysis of metabolites assisted in identifying similar resource utilization niches between phylogenetically distinct species, and different niches being occupied by phylogenetically distinct species (Venturelli*, et al.* 2018). Currently, we are using the Db-MM consortium described in **chapter 8** to investigate longer temporal dynamics in bioreactors with controlled pH in both batch and semi-

continuous feeding mode to better understand the compositional stability and transcriptome level responses to interacting partners. We expect that such detailed multi-omics approaches will reveal interesting emergent behavior and improve the development of predictive models at least for the core intestinal microbiome.

Challenges, opportunities and future prospects

Over the course of the research described in this thesis, several challenges and knowledge gaps were identified. From isolation of highly prevalent intestinal bacterial species, characterization of prevalent cultured bacteria, improvement of bioinformatic and biostatistics tools, genome annotations to metabolic modelling to investigate species interactions.

Several bacteria remain uncharacterized

In **chapter 2**, while performing a literature survey for the core bacteria in the human intestinal tract, it was noted that several key species remain uncharacterized. There exists a major lacuna in our understanding of the metabolic roles of individual species, including several core species such as *Subdoligranulum variable*, *Coprococcus eutactus*, *Lachnospira pectinoschiza* and members of *Dialister* and *Collinsella,* to just name a few. In addition, to these, in **chapter 8**, using metagenomic data we identified *Oscillibacter*, uncharacterized Lachnospiraceae and uncharacterized Ruminococcoceae species among the core microbiota. These bacteria are consistently identified in molecular profiling studies of the microbiome, however, due to lack of metabolic characterization, their roles in the community remains elusive. In addition to the cultured bacterial species, there remain a few key bacterial groups that have not be cultured in laboratory conditions, one example being *Oscillospira* and related bacteria (Gophna*, et al.* 2017, Konikoff and Gophna 2016). While high-throughput cultivation assays, also termed culturomics, have achieved success in cultivating >70% of the human intestinal bacteria, isolation of some key species will require more targeted approaches (Lagier*, et al.* 2016). These approaches will require integrating the knowledge of their ecology and predicted nutrient requirements based on metagenome-assembled genomes.

Improving existing bioinformatics tools for microbiome research

In **chapter 3**, while analyzing large scale EMP datasets, we realized that there was a lack of a user-friendly tool that would provide a firsthand insight into the properties of sequencing data such as quality of data, distribution of taxa and other routinely used data visualizations and formatting options. To address this gap, we developed the *microbiomeutilities* and *microbiomeutilities shiny* applications based on the open software environment R and R/Shiny. The *microbiomeutilities* R package supports the parent *microbiome* R package (https://tinyurl.com/y9kzu4uc) and provides a suite of tools that further enhance the quality and reproducibility of amplicon data analysis. It can be used to identify core microbiota in microbiome datasets, calculate several alpha diversity indices and identify the number of hyperdominant taxa as demonstrated in **chapter 3**. However, several additional functionalities need to be incorporated in these tools that should address the following aspects: a) Time-series analysis is a crucial functionality as observed during the analysis of data in Chapter 8, b) Normalizing copy number variation especially when analyzing defined communities, where users can use custom values of gene copies depending on the species being used, c) Automated comparison of taxa identified in positive and negative controls and samples to identify potential contaminants.

Another challenge identified in the currently available bioinformatics tools was related to the genome and metagenome function annotations. The improvements in accuracy of genome annotation tools will be crucial for metabolic modelling approaches that are used to simulate and predict microbial interactions in defined as well as natural communities. Recently, a large number of semi-curated constraint based metabolic models of human intestinal bacteria were created (Magnúsdóttir*, et al.* 2017). These semi-curated constraint based metabolic models are now being used to investigate microbial interactions in communities as well as in pairs of microorganisms (Baldini*, et al.* 2018). A graph theory-based approach employing metabolic networks to identify species complementarity and competition is also available (Levy and Borenstein 2013). Results and observations obtained from both constraint-based metabolic models and graph theory based metabolic networks are only as good as the functional gene annotation that the current bioinformatics tools provide. A major challenge is to

annotate transporter genes, which are also key functions that influence accuracy of prediction of microbial interactions using *in-silico* metabolic models (Greenblum*, et al.* 2013, Magnúsdóttir*, et al.* 2017). The development of high-quality metabolic models will require detailed physiological analysis of individual bacterial species. By integrating multi-omics data and physiological studies, metabolic models have been developed for *Akkermansia muciniphila* (iAkkMuc_588) and *Faecalibacterium prausnitzii* (*iFpraus_v1.0*), *Bacteroides thetaiotamicron* (*iBth1201*) and *Eubacterium rectale* (*iEre400*), and the methanogen *Methanobrevibacter smithii* (*iMsi385*) (Shoaie*, et al.* 2013, Heinken*, et al.* 2014b, Ottman*, et al.* 2017). In **chapter 5, 6 and 7**, detailed physiological properties of the key intestinal bacteria *Intestinimonas butyriciproducens* and *Anaerobutyricum soehngenii* have been unravelled. The information obtained in these three chapters not only provides a better understanding of their metabolic capabilities and ecological niches, it also will aid in development of high-quality metabolic models. Focus on developing improved metabolic models for these and other core bacteria will be crucial for improving our understanding of the metabolic interaction networks and predictive modeling, involving the designing of minimal microbiomes with known ecophysiological properties.

Minimal microbiome(s) to understand the intricacies of the intestinal microbiome

Minimal microbiomes will be crucial for unravelling active metabolic networks and potential interactions which may be hidden due to the extensive technical noise and several unknowns in the studies based on natural communities (for e.g. feces). Additionally, minimal microbiomes allow for studying emergent metabolic behaviors that could explain evolution of cooperation and competition between the microbial members in the microbiome (D'hoe *et al.*, 2018, Venturelli *et al.*, 2018, Desai *et al.*, 2016). In a proof of concept study, we demonstrated how diet and metabolic specialization drive assembly of the community, evenness and species composition (**Chapter 8**). Incorporating metatranscriptomics in future follow up experiments can reveal the active metabolic responses of individual bacteria to other bacteria within the community. In fact, we are currently replicating the study with integrated meta-omics to unravel metabolic interaction networks driven by dietary components. There still remains a wide-open field for similar studies investigating several combinations of core

and non-core species to address diverse research questions. An important metabolic interaction that occurs in the human intestine is driven by host derived mucus (Belzer and De vos 2012). Trophic interactions between mucus degrading bacteria and butyrate producing bacteria have been studied (Belzer *et al.,* 2017, Chia *et al.,* 2018). However, multi-species interactions, which incorporate competition for mucus breakdown products and other nutrients, and potential emergent properties of these interactions have not been investigated. In addition, the effect of diet and mucus degrading key stone species on the overall community dynamics remains understudied. We propose that future development of minimal microbiomes should address these questions by designing specific minimal microbiomes. Similar to the Db-MM, a mucus based minimal microbiome (Mb-MM) which could include keystone species such as *Akkermansia muciniphila* and other butyrate and propionate producers, as well as sulfate reducers should be designed. Furthermore, the Db-MM can be integrated with the Mb-MM to create a minimal microbiome which can be useful for studying the overall microbiome in a well-controlled setting. The research on minimal microbiomes requires development and innovative solutions to several practical aspects, as listed below.

1) Development of high-throughput anaerobic cultivation methods. Examples include micro-titer plates, micro-fluidic devices and small volume bioreactors to name a few (Shah *et al.,* 2016).

2) Development of cheaper, faster and better ways of estimating community composition High-throughput full-length 16S rRNA gene sequencing can aid in investigating dynamics of closely related species in consortia.

3) Improved metatranscriptomics, metaproteomics and metabolomics methods for handling small sample volumes and cell densities. This would be particularly useful when using high-throughput approaches such as small volume bioreactors. There is scope for single-cell RNA-seq for in-depth assessment of ecotype responses in minimal microbiomes.

4) Development of bioinformatics tools that are open and allow collaboration between multi-disciplinary groups collaborating on minimal microbiome

research. Importantly, tools that allow seamless integration of multi-omic datasets, including growth and pH measurements are the need of the hour.

Concluding remarks

Our understanding of the human microbiome and its role in human health has grown substantially over the past few decades. The last few years has seen a rise in studies that move forward from mere associations to identifying mechanisms of how microbes influence host health. During the writing of this thesis a scientific blog post asked "Is it time to stop measuring, and put the 'ology' back into microbiome research?" (https://goo.gl/fRgrMv). We definitely should not to stop measuring, as there is a large and diverse human population for which the microbiome profiles are unknown, however, we should also not stop at just measuring. Identifying, classifying and characterizing microbes is one of the aspects of microbiology, as is understanding the physiology and their interactions. With a synergy of different approaches with hypothesis driven investigation of microbiomes, there is enormous potential for developing better biotherapeutics that ultimately will improve the quality of life.

Thesis Summary
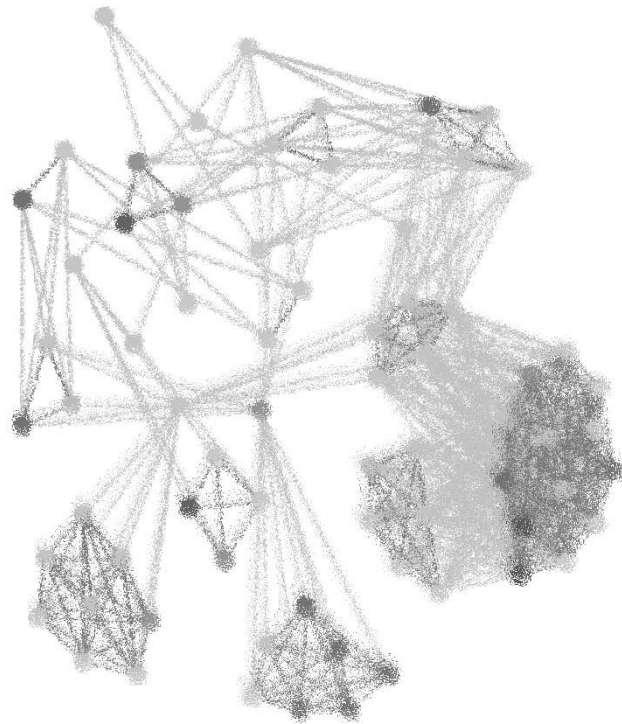
Summary

Research in this thesis aims to understand the human intestinal microbiome at three levels of complexity, ranging from ecosystems to defined mixed consortia to individual bacterial species. Understanding the role of the microbiome in defining and tipping the balance between health and disease can greatly benefit from studies that approach the intestinal microbiome from these three levels. In **Chapter 1**, a brief overview of the current knowledge of the human intestinal microbiome is provided, and the motivation for the research described in this thesis is highlighted. **Chapter 2** focuses on the ecosystem level understanding of the intestinal microbiome and discusses possible leads for the identification of targets for microbial diagnostics that could be used for predicting changes or homeostasis as well as for determining the efficacy of intervention strategies.

A key ecological property of hyperdominance is identified in diverse microbial ecosystems in **Chapter 3**. We hypothesize that niche-neutral processes likely explain the hyperdominance and dynamics of a few well-adapted species in the human intestinal microbiome. Most of the hyperdominant species are known to play key roles in degradation of complex polysaccharides as well as in the production of health promoting butyrate and/or propionate.

Analysis of molecular data obtained from high-throughput sequencing of the 16S rRNA gene is an important part of unravelling the community composition, diversity and structure. **Chapter 4** addresses key bottle necks often encountered by researchers performing microbial community surveys, *i.e.* handling, analysis and visualization of high throughput marker gene sequencing data. To this end, we developed the *microbiomeutilities* software package using the open source R software environment which complements and extends the functionality of other widely used R packages such as *phyloseq* and *microbiome*. Additionally, an open source user-friendly web-based tool that requires no programming experience was developed to provide a first-hand view of sequencing data quality, taxonomic composition, alpha diversity, ordinations and core microbiota.

Animal models are widely used in studies investigating the mechanistic role of the microbiome. However, there can be strain level variation within microbial species of interest driven by the source of isolation. As described in **Chapter 5**, we identified several strain level differences between human and mouse intestinal isolates of *Intestinimonas butyriciproducens*. The findings of this and other studies indicated the need for taking into account such host origin features of strains of the same bacterial species when employing animal models for mechanistic studies.

Microbial profiling studies rely on appropriate phylogenetic classification of bacteria, and misclassifications can impact the interpretation of the observations. *Eubacterium hallii* was identified as a core bacterium in both phylogenetic microarray (HITChip) and shotgun metagenomic profiling data. Investigation of the genome of previously classified strains *Eubacterium hallii* L2-7 (DSM 17630) and *Eubacterium hallii* DSM3353 revealed that they are likely different species. In **Chapter 6**, we resolved the misclassification of these key intestinal species using a classical polyphasic taxonomic approach. We proposed a novel genus name *Anaerobutyricum* for *Eubacterium hallii* and related species and classified *Eubacterium hallii* L2-7 (DSM 17630) as a novel species, *Anaerobutyricum soehngenii*. To better understand the ecology and metabolic characteristics of *A. soehngenii*, we carried out proteome analyses of cultures grown with different substrates as described in **Chapter 7**. We identified active metabolic pathways of butyrogenesis, vitamin B12 biosynthesis and a unique gene cluster for DL-lactate utilization. This unique DL-lactate utilization cluster likely explains the ability of *A. soehngenii* to grow and produce butyrate from this low energy substrate.

Analysis of natural ecosystems is challenging due to their high complexity, whereas the applicability of mono-culture studies is limited by their simplicity. A pragmatic approach to better understand complex microbial ecosystems is the use of defined consortia consisting of a limited number of species representing key functionalities. Design of such minimal microbiomes should incorporate both ecological and physiological knowledge of communities and species, respectively. A proof of concept for such an integrated approach is described in **Chapter 8**, where we used a combination of nine species that were selected based on the carbon sources used in this experiment, and which reproducibly assembled into defined microbiomes

producing butyrate and propionate. The key features of this diet-based minimal microbiome were high competition and low complementarity, functional redundancy, deterministic assembly and community dynamics. The consortium included some well characterized strains while others are in the process of being characterized. These features provided an opportunity to model emergent metabolic features in the presence of several interacting partners in a well-controlled environment.

Finally, in **Chapter 9**, all the research findings of this thesis are discussed in detail highlighting the importance of employing a synergistic approach to study the intestinal microbiome, which incorporates both microbial ecology and physiology. During the course of the research described in this thesis, several key areas were identified for future research including detailed characterization of several core bacteria, hardware development for high-throughput anaerobic cultivation and bioinformatics and systems biology modelling tools to handle and interpret multi-omic datasets.

Nederlandse samenvatting
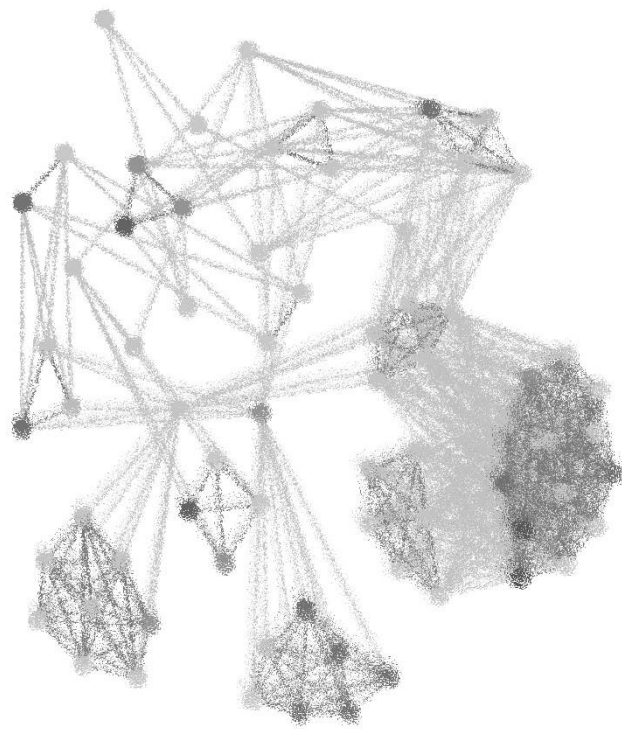
Nederlandse samenvatting

Het onderzoeksdoel van dit proefschrift was om het menselijke darm microbioom beter te leren begrijpen door het te onderzoeken op drie niveau's van complexiteit. Deze niveau's variëren van hele ecosystemen, naar gedefinieerde consortia tot individuele bacteriële soorten. Studies op deze drie niveau's van complexiteit kunnen veel bijdragen aan het onderzoek naar het microbioom en de rol ervan voor de balans tussen gezondheid en ziekte. Hoofdstuk 1 geeft een beknopt overzicht over wat er op dit moment bekend is over het menselijke darm microbioom, en de motivatie waarom er voor dit onderzoek is gekozen in dit proefschrift. Hoofdstuk 2 richt zich op het beter begrijpen van het darm microbioom op het niveau van het ecosysteem en bediscussieerd mogelijke toepassingen van gevonden targets in de microbiële diagnostiek. Microbiële biomarkers zouden kunnen worden gebruikt voor het voorspellen van afwijkingen van een "gezonde" darm (homeostase) en voor het vaststellen van de doeltreffendheid van bijvoorbeeld voedingsinterventies. Hoofdstuk 3 beschrijft het fenomeen van 'Hyperdominance', een belangrijke eigenschap van ecosystemen dat ook werd gevonden in diverse microbiële ecosystemen. In een notendop betekend 'hyperdominance' dat ook in complexe ecosystemen als het menselijke darm microbioom met honderden of duizenden verschillende soorten micro-organismen maar een zeer beperkt aantal verschillende soorten goed zijn voor het grootste gedeelte van alle cellen in het systeem. Onze hypothese is dat niche-neutrale processen waarschijnlijk de reden zijn voor 'hyperdominance' en de dynamiek van een beperkt aantal goed aangepaste soorten van het menselijke darm microbioom. Van de meeste 'hyperdominante' soorten is bekend dat ze een belangrijke rol spelen in de afbraak van complexe koolhydraten en de productie van gezondheidsbevorderende korte keten vetzuren zoals boterzuur en/of propionzuur.

Moleculaire analyse gebaseerd op high-throughput sequencing van het 16S ribosomaal RNA gen wordt hedendaags door veel laboratoria routinematig gebruikt voor het ontrafelen van de compositie, diversiteit en structuur van microbiële gemeenschappen. Hoofdstuk 4 behandelt de problemen die onderzoekers tegenkomen tijdens het onderzoeken van microbiële gemeenschappen, in bijvoorbeeld de analyse, weergave en interpretatie van soortgelijke grote marker gen sequentie datasets. Daarom hebben we het microbiomeutilities software pakket ontwikkeld dat gebruikt kan worden in de Open Source omgeving van de R software. De toepassingsmogelijkheden van dit pakket worden verder vergroot wanneer het wordt gecomplementeerd met andere R software pakketten zoals phyloseq en microbiome. Hier op volgend is er een gebruiksvriendelijke open source website ontwikkeld om mensen zonder programmeer ervaring te voorzien van een eenvoudige eerste inkijk in sequentie data kwaliteit, microbiële samenstelling, complexiteit, multivariate analyses en geconserveerde delen van de microbiota.

Proefdiermodellen worden regelmatig gebruikt in het onderzoek om de rol die bepaalde bacteriën spelen in het darm microbioom te onderzoeken. Echter de bron waaruit bacteriën worden geïsoleerd kan invloed hebben op genotypische en fenotypische variatie op stam niveau binnen een microbiële soort. Zoals beschreven in Hoofdstuk 5 vonden we verschillen tussen stammen van Intestinimonas butyriciproducens die geïsoleerd zijn uit ontlasting van mens en muis. De resultaten

hiervan en van andere studiës geven aan dat het belangrijk is om met dergelijke verschillen rekening te houden bij het uitvoeren en interpreteren van studies met proefdieren.

De moleculaire analyse van de microbiële samenstelling in bijvoorbeeld ontlasting is afhankelijk van de juiste fylogenetische classificatie van bacteriën, want misclassificatie kan een grote invloed hebben op de interpretatie van de waarnemingen. *Eubacterium halllii* was eerder in verschillende studies en datasets geïdentificeerd als één van de micro-organismen die in de meesten, zo niet alle, mensen deel uitmaakt van de darm microbiota. Genoom onderzoek van de in het verleden geclassificeerde stammen *Eubacterium hallii* L2-7 (DSM17630) en *Eubacterium hallii* DSM3353 heeft echter nu laten zien dat deze twee stammen waarschijnlijk verschillende soorten zijn. In Hoofdstuk 6 kon de misclassificatie van deze belangrijke darm bacteriën worden aangetoond met behulp van klassieke polyfasische taxonomie. We hebben daarom voorgesteld, om beide stammen als nieuwe soorten in een nieuw geslacht, *Anaerobutyricum*, onder te brengen, waarbij *Eubacterium hallii* L2-7 (DSM 17630) is geclassificeerd als een nieuwe soort, *Anaerobutyricum soehngenii*. Om de ecologie en de metabole eigenschappen van *A. soehngenii* beter te leren begrijpen, hebben we een proteoom analyse uitgevoerd op kweken van *A. soehngenii* gegroeid op verschillende substraten zoals wordt beschreven in Hoofdstuk 7. We hebben verschillende metabole routes gevonden voor de productie van boterzuur en vitamine B12 en een unieke gen cluster voor het gebruik van DL-melkzuur als koolstofbron. Het gebruik van dit DL-melkzuur gen cluster kan mogelijkerwijs verklaren waarom A. soehngenii kan groeien en boterzuur kan produceren van dit weinig energie leverende substraat.

Analyse van natuurlijke ecosystemen is zeer uitdagend vanwege de complexiteit die ze bezitten. daarentegen is het onderzoek aan reinculturen gelimiteerd vanwege zijn eenvoudigheid. Het is makkelijk voor te stellen dat een bacterie in reincultuur anders gaat gedragen dan in een complex ecosysteem zoals het menselijke darm microbioom. Een pragmatische aanpak voor het beter begrijpen van complexe ecosystemen is daarom de toepassing van gespecificeerde consortia bestaand uit een beperkt aantal soorten die een aantal essentiële functies vertegenwoordigen. Dit soort gedefinieerde microbiomen moet worden ontworpen met invulling van ecologische en fysiologische kennis over microbiële gemeenschappen en soorten. De toepassing van dit concept wordt beschreven in Hoofdstuk 8, waar we een combinatie van negen soorten hebben geselecteerd op basis van een gedefinieerd dieet, dat wil zeggen de koolstofbron die de bacteriën kunnen gebruiken zoals bijvoorbeeld zetmeel, evenals de metabolieten die worden gevormd, zoals bijvoorbeeld melkzuur, boterzuur en propionzuur. In onze experimenten vormden deze geselecteerde bacteriën herhaaldelijk eenvoudige en stabiele microbiomen die boterzuur en propionzuur produceren. De belangrijkste eigenschappen van deze op dieet gebaseerde gedefinieerde microbiomen waren een hoge mate van competitie en beperkte complementariteit, overlap in functionele bijdrages van individuele stammen, gerichte assemblage evenals dynamiek van het microbioom met betrekking tot relatieve hoeveelheden van de negen verschillende stammen. Het consortium bestond uit een aantal goed gekarakteriseerde stammen, en een aantal stammen die nog niet goed zijn beschreven. Daardoor was het mogelijk, om de metabole eigenschappen van de verschillende stammen te modelleren in de

aanwezigheid van verschillende interacterende partners in een gecontroleerde omgeving.

Afsluitend geeft Hoofdstuk 9 een algemene discussie van alle bevindingen in dit proefschrift met speciale aandacht voor het belang van een geïntegreerde aanpak voor het bestuderen van het darm microbioom, waarbij rekening wordt gehouden met zowel de microbiële ecologie als de fysiologie. Gedurende het verloop van dit onderzoek zijn er verschillende belangrijke aspecten geïdentificeerd die in toekomstige vervolgstudies kunnen worden onderzocht, zoals het karakteriseren van de essentiële kernleden van het microbioom, technologische ontwikkelingen voor een meer efficiënt kweken van tot nog toe niet of moeilijk kweekbare anaerobe bacteriën, bio-informatica en systeem-biologische modellen en praktijken voor het in kaart brengen en interpreteren van multi-omic datasets.

References
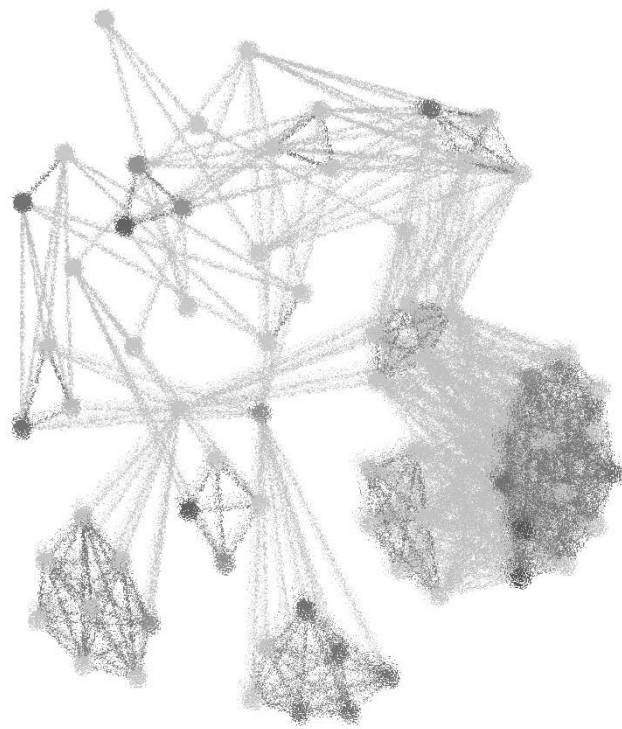
References

1.   Abdulamir AS, Hafidh RR, Bakar FA. The association of *Streptococcus bovis/gallolyticus* with colorectal tumors: the nature and the underlying mechanisms of its etiological role. *Journal of Experimental & Clinical Cancer Research* 2011;**30**: 1.

2.   Abubucker S, Segata N, Goll J *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology* 2012;**8**: e1002358.

3.   Albenberg L, Esipova TV, Judge CP *et al.* Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota in humans and mice. *Gastroenterology* 2014;**147**: 1055-63.

4.   Alkanani AK, Hara N, Gottlieb PA *et al.* Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. *Diabetes* 2015;**64**: 3510-20.

5.   Allen-Vercoe E, Daigneault M, White A *et al. Anaerostipes hadrus comb. nov.,* a dominant species within the human colonic microbiota; reclassification of *Eubacterium hadrum* Moore et al. 1976. *Anaerobe* 2012;**18**: 523-9.

6.   Allen RH, Stabler SP. Identification and quantitation of cobalamin and cobalamin analogues in human feces. *American Journal of Clinical Nutrition* 2008;**87**: 1324-35.

7.   Alneberg J, Bjarnason BS, De Bruijn I *et al.* Binning metagenomic contigs by coverage and composition. *Nature methods* 2014;**11**: 1144.

8.   Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *Journal of Molecular Biology* 1990;**215**: 403-10.

9.   Amir A, McDonald D, Navas-Molina JA *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems* 2017;**2**: e00191-16.

10.   Andersen KSS, Kirkegaard RH, Karst SM *et al.* ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* 2018: 299537.

11.   Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

12.   Arumugam M, Raes J, Pelletier E *et al.* Enterotypes of the human gut microbiome. *Nature* 2011;**473**: 174-80.

13.   Atashgahi S, Shetty SA, Smidt H *et al.* Flux, impact and fate of halogenated xenobiotic compounds in the gut. *Frontiers in physiology* 2018;**9**: 888.

14.   Aziz R, Bartels D, Best A *et al.* The RAST server: Rapid annotations using subsystems technology. *BMC Genomics* 2008;**9**: 75.

15.   Bäckhed F, Roswall J, Peng Y *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell host & microbe* 2015;**17**: 690-703.

16.   Bakir MA, Kitahara M, Sakamoto M *et al. Bacteroides intestinalis* sp. nov., isolated from human faeces. *International journal of systematic and evolutionary microbiology* 2006;**56**: 151-4.

17.   Baldassano SN, Bassett DS. Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease. *Scientific reports* 2016;**6**.

18.   Baldini F, Heinken A, Heirendt L *et al.* The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. *bioRxiv* 2018: 318485.

19.   Barcenilla A, Pryde SE, Martin JC *et al.* Phylogenetic relationships of butyrate-producing bacteria from the human gut. *Applied and environmental microbiology* 2000;**66**: 1654-61.

20.   Barr JJ, Auro R, Furlan M *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* 2013;**110**: 10771-6.

21.   Baumann-Dudenhoeffer AM, D'Souza AW, Tarr PI *et al.* Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nature medicine* 2018: 1.

22.   Baxter NT, Ruffin MT, Rogers MA *et al.* Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome medicine* 2016;**8**: 37.

23.   Begley M, Hill C, Gahan CGM. Bile salt hydrolase activity in probiotics. *Applied and Environmental Microbiology* 2006;**72**: 1729-38.

24.   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. *Applied and environmental microbiology* 2006;**72**: 3593-9.

25.   Belzer C, Chia LW, Aalvink S *et al.* Microbial metabolic networks at the mucus layer lead to diet-independent butyrate and vitamin B12 production by intestinal symbionts. *MBio* 2017;**8**: e00770-17.

26.   Belzer C, de Vos WM. Microbes inside—from diversity to function: the case of Akkermansia. *The ISME journal* 2012;**6**: 1449-58.

27.   Bezuidt O, Lima-Mendez G, Reva O. SeqWord Gene Island Sniffer: a program to study the lateral genetic exchange among bacteria. *World Academy of Science, Engineering and Technology* 2009;**58**: 1169-11274.

28.   Biagi E, Nylund L, Candela M *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PloS one* 2010;**5**: e10667.

29.   Biswas A, Gagnon JN, Brouns SJ *et al.* CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA biology* 2013;**10**: 817-27.

30.   Bocci V. The neglected organ: bacterial flora has a crucial immunostimulatory role. *Perspectives in Biology and Medicine* 1992;**35**: 251-60.

31.   Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics* 2014;**15**: 211.

32.   Boisvert S, Raymond F, Godzaridis É *et al.* Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology* 2012a;**13**: R122-R.

33.   Boisvert S, Raymond F, Godzaridis É *et al.* Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology* 2012b;**13**: R122.

34.   Bokulich NA, Chung J, Battaglia T *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* 2016;**8**: 343ra82-ra82.

35.   Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**: 2114-20.

36.   Brioukhanov AL, Netrusov AI. Catalase and superoxide dismutase: Distribution, properties, and physiological role in cells of strict anaerobes. *Biochemistry* 2004;**69**: 949-62.

37.   Brown JH, Gillooly JF, Allen AP *et al.* Toward a metabolic theory of ecology. *Ecology* 2004;**85**: 1771-89.

38.   Browne HP, Forster SC, Anonye BO *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. 2016;**533**: 543-6.

39.   Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature methods* 2014;**12**: 59.

40.   Buckel W, Janssen PH, Schuhmann A *et al. Clostridium viride* sp. nov., a strictly anaerobic bacterium using 5-aminovalerate as growth substrate, previously assigned to Clostridium aminovalericum. *Archives of microbiology* 1994;**162**: 387-94.

41.   Bui TPN. Ecophysiology of novel intestinal butyrate-producing bacteria: Wageningen University, 2016.

42.   Bui TPN, 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**.

43.   Bui TPN, Shetty SA, Lagkouvardos I *et al.* Comparative genomics and physiology of the butyrate-producing bacterium *Intestinimonas butyriciproducens*. *Environmental microbiology reports* 2016;**8**: 1024-37.

44.   Callahan BJ, McMurdie PJ, Rosen MJ *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods* 2016;**13**: 581.

45.	Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC bioinformatics* 2009;**10**: 421.

46.	Canani RB, Costanzo M, Leone L *et al.* Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. *World J Gastroenterol* 2011;**17**: 1519-28.

47.	Cao Y, Wang Y, Zheng X *et al.* RevEcoR: an R package for the reverse ecology analysis of microbiomes. *BMC bioinformatics* 2016;**17**: 294.

48.	Caporaso JG, Kuczynski J, Stombaugh J *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature methods* 2010;**7**: 335.

49.	Caporaso JG, Lauber CL, Costello EK *et al.* Moving pictures of the human microbiome. *Genome biology* 2011;**12**: 1.

50.	Carlier J-P, Bedora-Faure M, K'ouas G *et al.* Proposal to unify *Clostridium orbiscindens* Winter et al. 1991 and *Eubacterium plautii* (Séguin 1928) Hofstad and Aasjord 1982, with description of *Flavonifractor plautii* gen. nov., comb. nov., and reassignment of *Bacteroides capillosus* to *Pseudoflavonifractor capillosus* gen. nov., comb. nov. *International journal of systematic and evolutionary microbiology* 2010;**60**: 585-90.

51.	Castellarin M, Warren RL, Freeman JD *et al. Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome research* 2012;**22**: 299-306.

52.	Chang CH, Frey PA. Cloning, Sequencing, Heterologous Expression, Purification, and Characterization of Adenosylcobalamin-dependentd-Lysine 5,6-Aminomutase from *Clostridium sticklandii*. *Journal of Biological Chemistry* 2000;**275**: 106-14.

53.	Charlop-Powers Z, Brady SF. phylogeo: an R package for geographic analysis and visualization of microbiome data. *Bioinformatics* 2015;**31**: 2909-11.

54.	Chassard C, Delmas E, Lawson PA *et al. Bacteroides xylanisolvens* sp. nov., a xylan-degrading bacterium isolated from human faeces. *International journal of systematic and evolutionary microbiology* 2008;**58**: 1008-13.

55.	Chassard C, Delmas E, Robert C *et al.* The cellulose-degrading microbial community of the human gut varies according to the presence or absence of methanogens. *FEMS microbiology ecology* 2010;**74**: 205-13.

56.	Chen P, Ailion M, Bobik T *et al.* Five promoters integrate control of the cob/pdu regulon in *Salmonella typhimurium*. *Journal of bacteriology* 1995;**177**: 5401-10.

57.	Chen W, Liu F, Ling Z *et al.* Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS one* 2012;**7**: e39743.

58.	Cheng J. Shiny: Easy web applications in R*The R User Conference, useR! 2013 July 10-12 2013 University of Castilla-La Mancha, Albacete, Spain* volume 10, 2013, 93.

59.	Chia LW, Hornung BV, Aalvink S *et al.* Deciphering the trophic interaction between *Akkermansia muciniphila* and the butyrogenic gut commensal *Anaerostipes caccae* using a metatranscriptomic approach. *Antonie van Leeuwenhoek* 2018: 1-15.

60.	Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* 2012;**13**: 260.

61.	Chun J, Oren A, Ventosa A *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International journal of systematic and evolutionary microbiology* 2018;**68**: 461-6.

62.	Chung WSF, Walker AW, Louis P *et al.* Modulation of the human gut microbiota by dietary fibres occurs at the species level. *BMC biology* 2016;**14**: 3.

63.	Claesson MJ, Cusack S, 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;**108**: 4586-91.

64.	Claesson MJ, Jeffery IB, Conde S *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* 2012;**488**: 178-84.

65.   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**: e6669.

66.   Clarke J, Wu H-C, Jayasinghe L *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology* 2009;**4**: 265.

67.   Clarke SF, Murphy EF, O'Sullivan O *et al.* Exercise and associated dietary extremes impact on gut microbial diversity. *Gut* 2014: gutjnl-2013-306541.

68.   Cockburn DW, Orlovsky NI, Foley MH *et al.* Molecular details of a starch utilization pathway in the human gut symbiont *Eubacterium rectale*. *Molecular microbiology* 2015;**95**: 209-30.

69.   Cohan FM. Bacterial species and speciation. *Systematic biology* 2001;**50**: 513-24.

70.   Cohan FM. Periodic selection and ecological diversity in bacteria*Selective sweep*: Springer, 2005, 78-93.

71.   Cole JR, Wang Q, Fish JA *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 2014;**42**: D633-D42.

72.   Connell JH, Sousa WP. On the evidence needed to judge ecological stability or persistence. *American Naturalist* 1983: 789-824.

73.   Cornick NA, Jensen N, Stahl D *et al. Lachnospira pectinoschiza* sp. nov., an anaerobic pectinophile from the pig intestine. *International Journal of Systematic and Evolutionary Microbiology* 1994;**44**: 87-93.

74.   Costea PI, Munch R, Coelho LP *et al.* metaSNV: a tool for metagenomic strain level analysis. *PLoS One* 2017a;**12**: e0182392.

75.   Costea PI, Zeller G, Sunagawa S *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotech* 2017b;**35**: 1069.

76.   Costello EK, Stagaman K, Dethlefsen L *et al.* The application of ecological theory toward an understanding of the human microbiome. *Science* 2012;**336**: 1255-62.

77.   Cotillard A, Kennedy SP, Kong LC *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* 2013;**500**: 585-8.

78.   Cox LM, Yamanishi S, Sohn J *et al.* Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell* 2014;**158**: 705-21.

79.   Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: Networks, competition, and stability. *Science* 2015;**350**: 663-6.

80.   Crost EH, Tailford LE, Le Gall G *et al.* Utilisation of mucin glycans by the human gut symbiont Ruminococcus gnavus is strain-dependent. *PLoS One* 2013;**8**: e76341.

81.   D'hoe K, Vet S, Faust K *et al.* Integrated culturing, modeling and transcriptomics uncovers complex interactions and emergent behavior in a synthetic gut community. *bioRxiv* 2018: 299644.

82.   Damaskos D, Kolios G. Probiotics and prebiotics in inflammatory bowel disease: microflora 'on the scope'. *British journal of clinical pharmacology* 2008;**65**: 453-67.

83.   Dao MC, Everard A, Aron-Wisnewsky J *et al. Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* 2015: gutjnl-2014-308778.

84.   Datta MS, Sliwerska E, Gore J *et al.* Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nat Commun* 2016;**7**: 11965.

85.   David LA, Maurice CF, Carmody RN *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 2014;**505**: 559-63.

86.   Davie JR. Inhibition of histone deacetylase activity by butyrate. *The Journal of nutrition* 2003;**133**: 2485S-93S.

87.   De Filippo C, Cavalieri D, Di Paola M *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences* 2010;**107**: 14691-6.

88.   de Goffau MC, Fuentes S, van den Bogert B *et al.* Aberrant gut microbiota composition at the onset of type 1 diabetes in young children. *Diabetologia* 2014;**57**: 1569-77.

89.   De Roy K, Marzorati M, Van den Abbeele P *et al.* Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental microbiology* 2014;**16**: 1472-81.

90.   de Vos WM. Fame and future of faecal transplantations–developing next-generation therapies with synthetic microbiomes. *Microbial biotechnology* 2013;**6**: 316-25.

91.   de Vos WM, de Vos EA. Role of the intestinal microbiome in health and disease: from correlation to causation. *Nutrition reviews* 2012;**70**: S45-S56.

92.   de Vos WM, Nieuwdorp M. Genomics: A gut prediction. *Nature* 2013;**498**: 48-9.

93.   De Vuyst L, Leroy F. Cross-feeding between bifidobacteria and butyrate-producing colon bacteria explains bifidobacterial competitiveness, butyrate production, and gas production. *International journal of food microbiology* 2011;**149**: 73-80.

94.   Defnoun S, Labat M, Ambrosio M *et al. Papillibacter cinnamivorans* gen. nov., sp. nov., a cinnamate-transforming bacterium from a shea cake digester. *International journal of systematic and evolutionary microbiology* 2000;**50**: 1221-8.

95.   Degnan PH, Barry NA, Mok KC *et al.* Human gut microbes use multiple transporters to distinguish vitamin B12 analogs and compete in the gut. *Cell host & microbe* 2014a;**15**: 47-57.

96.   Degnan PH, Taga ME, Goodman AL. Vitamin B12 as a modulator of gut microbial ecology. *Cell metabolism* 2014b;**20**: 769-78.

97.   Deis RC, Kearsley MW. Sorbitol and mannitol. *Sweeteners and sugar alternatives in food technology* 2012: 331-46.

98.   Deptula P, Kylli P, Chamlagain B *et al.* BluB/CobT2 fusion enzyme activity reveals mechanisms responsible for production of active form of vitamin B_{12} by *Propionibacterium freudenreichii*. *Microbial Cell Factories* 2015;**14**: 1-12.

99.   Derrien M, Belzer C, de Vos WM. *Akkermansia muciniphila* and its role in regulating host functions. *Microbial pathogenesis* 2017;**106**: 171-81.

100.   Desguin B, Goffin P, Viaene E *et al.* Lactate racemase is a nickel-dependent enzyme activated by a widespread maturation system. *Nat Commun* 2014;**5**: 3615.

101.   Despres J, Forano E, Lepercq P *et al.* Xylan degradation by the human gut *Bacteroides xylanisolvens* XB1A T involves two distinct gene clusters that are linked at the transcriptional level. *BMC genomics* 2016;**17**: 326.

102.   Deutscher MP. How bacterial cells keep ribonucleases under control. *FEMS microbiology reviews* 2015;**39**: 350-61.

103.   Dhariwal A, Chong J, Habib S *et al.* MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic acids research* 2017: gkx295.

104.   Dhillon BK, Chiu TA, Laird MR *et al.* IslandViewer update: improved genomic island discovery and visualization. *Nucleic acids research* 2013;**41**: W129-W32.

105.   Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. *Nature* 2014;**509**: 357.

106.   Dinh DM, Volpe GE, Duffalo C *et al.* Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *The Journal of infectious diseases* 2014;**211**: 19-27.

107.   Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology* 2016;**14**: 20-32.

108.   Donia MS, Fischbach MA. Small molecules from the human microbiota. *Science* 2015;**349**: 1254766.

109.   Downes J, Munson M, Wade W. *Dialister invisus* sp. nov., isolated from the human oral cavity. *International journal of systematic and evolutionary microbiology* 2003;**53**: 1937-40.

110.   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**: e7063.

111.   Duncan SH, Flint HJ. Proposal of a neotype strain (A1-86) for *Eubacterium rectale*. Request for an opinion. *International journal of systematic and evolutionary microbiology* 2008;**58**: 1735-6.

112.   Duncan SH, Hold GL, Barcenilla A *et al. Roseburia intestinalis* sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. *International journal of systematic and evolutionary microbiology* 2002a;**52**: 1615-20.

113.   Duncan SH, Hold GL, Harmsen HJ *et al.* Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *International journal of systematic and evolutionary microbiology* 2002b;**52**: 2141-6.

114.   Duncan SH, Louis P, Flint HJ. Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Applied and Environmental Microbiology* 2004;**70**: 5810-7.

115.   Duvallet C, Gibbons SM, Gurry T *et al.* Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 2017;**8**: 1784.

116.   Dworkin M. *The prokaryotes: vol. 4: bacteria: firmicutes, cyanobacteria*: Springer Science & Business Media, 2006.

117.   Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 2013;**10**: 996-8.

118.   Edgar RC. Accuracy of microbial community diversity estimated by closed-and open-reference OTUs. *PeerJ* 2017;**5**: e3889.

119.   Eisen JA, Heidelberg JF, White O *et al.* Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome biology* 2000;**1**: research0011. 1.

120.   Elsayed S, Zhang K. Bacteremia caused by *Clostridium symbiosum*. *Journal of clinical microbiology* 2004;**42**: 4390-2.

121.   Engels C, Ruscheweyh H-J, Beerenwinkel N *et al.* The common gut microbe *Eubacterium hallii* also contributes to intestinal propionate formation. *Frontiers in Microbiology* 2016a;**7**.

122.   Engels C, Ruscheweyh H-J, Beerenwinkel N *et al.* The common gut microbe *Eubacterium hallii* also contributes to intestinal propionate formation. *Frontiers in microbiology* 2016b;**7**: 713.

123.   Epskamp S, Cramer AO, Waldorp LJ *et al.* qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software* 2012;**48**: 1-18.

124.   Erickson AR, Cantarel BL, Lamendella R *et al.* Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PloS one* 2012;**7**: e49138.

125.   Everard A, Belzer C, Geurts L *et al.* Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences* 2013;**110**: 9066-71.

126.   Ewels P, Magnusson M, Lundin S *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**: 3047-8.

127.   Faith JJ, Guruge JL, Charbonneau M *et al.* The long-term stability of the human gut microbiota. *Science* 2013;**341**: 1237439.

128.   Falony G, Joossens M, Vieira-Silva S *et al.* Population-level analysis of gut microbiome variation. *Science* 2016;**352**: 560-4.

129.   Faust K, Bauchinger F, Laroche B *et al.* Signatures of ecological processes in microbial community time series. *Microbiome* 2018;**6**: 120.

130.   Faust K, Lahti L, Gonze D *et al.* Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current opinion in microbiology* 2015;**25**: 56-66.

131.   Faust K, Sathirapongsasuti JF, Izard J *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* 2012;**8**: e1002606.

132.   Favier CF, de Vos WM, Akkermans AD. Development of bacterial and bifidobacterial communities in feces of newborn babies. *Anaerobe* 2003;**9**: 219-29.

133.   Fellows R, Denizot J, Stellato C *et al.* Microbiota derived short chain fatty acids promote histone crotonylation in the colon through histone deacetylases. *Nat Commun* 2018;**9**: 105.

134.   Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985;**39**: 783-91.

135.   Finlayson HJ. The effect of pH on the growth and metabolism of *Streptococcus bovis* in continuous culture. *Journal of applied bacteriology* 1986;**61**: 201-8.

136.   Finn RD, Bateman A, Clements J *et al.* Pfam: the protein families database. *Nucleic Acids Research* 2014;**42**: D222-D30.

137.   Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* 2014;**9**: e102451.

138.   Flint HJ, N. A. Logan HML-S, Oyston PCF. *The significance of prokaryote diversity in the human gastrointestinal tract. Prokaryotic Diversity*: Cambridge University Press, 2006.

139.   Flint HJ, Scott KP, Duncan SH *et al.* Microbial degradation of complex carbohydrates in the gut. *Gut microbes* 2012a;**3**: 289-306.

140.   Flint HJ, Scott KP, Louis P *et al.* The role of the gut microbiota in nutrition and health. *Nature Reviews Gastroenterology and Hepatology* 2012b;**9**: 577-89.

141.   Flint HJ, Scott KP, Louis P *et al.* The role of the gut microbiota in nutrition and health. *Nature Reviews Gastroenterology and Hepatology* 2012c;**9**: 577.

142.   Flores GE, Caporaso JG, Henley JB *et al.* Temporal variability is a personalized feature of the human microbiome. *Genome biology* 2014;**15**: 1.

143.   Forsberg C, Rogers HJ. Autolytic Enzymes in Growth of Bacteria. *Nature* 1971;**229**: 272.

144.   Foster KR, Schluter J, Coyte KZ *et al.* The evolution of the host microbiome as an ecosystem on a leash. *Nature* 2017a;**548**: 43.

145.   Foster ZS, Sharpton TJ, Grünwald NJ. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLoS computational biology* 2017b;**13**: e1005404.

146.   Frank DN, Pace NR. Gastrointestinal microbiology enters the metagenomics era. *Current opinion in gastroenterology* 2008;**24**: 4-10.

147.   Freilich S, Kreimer A, Meilijson I *et al.* The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic acids research* 2010;**38**: 3857-68.

148.   Fridovich I. Superoxide radical and superoxide dismutases. *Annual Review of Biochemistry* 1995;**64**: 97-112.

149.   Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012;**8**: e1002687.

150.   Fujita H, Eishi Y, Ishige I *et al.* Quantitative analysis of bacterial DNA from *Mycobacteria* spp., *Bacteroides vulgatus*, and *Escherichia coli* in tissue samples from patients with inflammatory bowel diseases. *Journal of gastroenterology* 2002;**37**: 509-16.

151.   Fukami T, Nakajima M. Community assembly: alternative stable states or alternative transient states? *Ecology letters* 2011;**14**: 973-84.

152.   Galisteo M, Duarte J, Zarzuelo A. Effects of dietary fibers on disturbances clustered in the metabolic syndrome. *The Journal of nutritional biochemistry* 2008;**19**: 71-84.

153.   Galley JD, Bailey M, Dush CK *et al.* Maternal obesity is associated with alterations in the gut microbiome in toddlers. *PloS one* 2014;**9**: e113026.

154.   Garmasheva I, Vasyliuk O, Kovalenko N *et al.* Intraspecies cellular fatty acids heterogeneity of *Lactobacillus plantarum* strains isolated from fermented foods in Ukraine. *Letters in Applied Microbiology* 2015;**61**: 283-92.

155.   Garsin DA. Ethanolamine utilization in bacterial pathogens: roles and regulation. 2010;**8**: 290-5.

156.   Gevers D, Kugathasan S, Denson LA *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe* 2014;**15**: 382-92.

157.   Gibson GR. Dietary modulation of the human gut microflora using the prebiotics oligofructose and inulin. *The Journal of nutrition* 1999;**129**: 1438S-41S.

158.   Gini C. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series* 1936;**208**: 73-9.

159.   Giongo A, Crabb DB, Davis-Richardson AG *et al.* PANGEA: pipeline for analysis of next generation amplicons. *The ISME journal* 2010;**4**: 852.

160.   Goldford JE, Lu N, Bajić D *et al.* Emergent simplicity in microbial community assembly. *Science* 2018;**361**: 469-74.

161.   Goodrich JK, Waters JL, Poole AC *et al.* Human genetics shape the gut microbiome. *Cell* 2014;**159**: 789-99.

162.   Gophna U, Konikoff T, Nielsen HB. *Oscillospira* and related bacteria–From metagenomic species to metabolic features. *Environmental microbiology* 2017;**19**: 835-41.

163.   Goris J, Konstantinidis KT, Klappenbach JA *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology* 2007;**57**: 81-91.

164.   Gorvitovskaia A, Holmes SP, Huse SM. Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle. *Microbiome* 2016;**4**: 1.

165.   Grant JR, Arantes AS, Stothard P. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC genomics* 2012;**13**: 202.

166.   Grech-Mora I, Fardeau M-L, Patel B *et al.* Isolation and characterization of *Sporobacter termitidis* gen. nov., sp. nov., from the digestive tract of the wood-feeding termite Nasutitermes lujae. *International Journal of Systematic and Evolutionary Microbiology* 1996;**46**: 512-8.

167.   Greenblum S, Chiu H-C, Levy R *et al.* Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. *Current opinion in biotechnology* 2013;**24**: 810-20.

168.   Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences* 2012;**109**: 594-9.

169.   Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic acids research* 2007;**35**: W52-W7.

170.   Grönlund M-M, Lehtonen O-P, Eerola E *et al.* Fecal microflora in healthy infants born by different methods of delivery: permanent changes in intestinal flora after cesarean delivery. *Journal of pediatric gastroenterology and nutrition* 1999;**28**: 19-25.

171.   Großkopf T, Soyer OS. Synthetic microbial communities. *Current opinion in microbiology* 2014;**18**: 72-7.

172.   Gutleben J, Chaib De Mares M, van Elsas JD *et al.* The multi-omics promise in context: from sequence to microbial isolate. *Critical reviews in microbiology* 2018;**44**: 212-29.

173.   Haarman M, Knol J. Quantitative real-time PCR analysis of fecal *Lactobacillus* species in infants receiving a prebiotic infant formula. *Applied and environmental microbiology* 2006;**72**: 2359-65.

174. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research* 2009;**19**: 1141-52.

175. Hamer HM, Jonkers D, Venema K *et al.* Review article: the role of butyrate on colonic function. *Alimentary Pharmacology & Therapeutics* 2008;**27**: 104-19.

176. Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 2011;**27**: 611-8.

177. Hardham JM, King KW, Dreier K *et al.* Transfer of *Bacteroides splanchnicus* to *Odoribacter* gen. nov. as *Odoribacter splanchnicus* comb. nov., and description of *Odoribacter denticanis* sp. nov., isolated from the crevicular spaces of canine periodontitis patients. *International journal of systematic and evolutionary microbiology* 2008;**58**: 103-9.

178. Harris K, Parsons TL, Ijaz UZ *et al.* Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process. 2014.

179. Harwood CS, Canale-Parola E. Branched-Chain Amino Acid Fermentation by a Marine Spirochete: Strategy for Starvation Survival. *Journal of Bacteriology* 1981;**148**: 109-16.

180. Hayashi H, Shibata K, Sakamoto M *et al. Prevotella copri* sp. nov. and *Prevotella stercorea* sp. nov., isolated from human faeces. *International journal of systematic and evolutionary microbiology* 2007;**57**: 941-6.

181. Hazen TC, Dubinsky EA, DeSantis TZ *et al.* Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science* 2010;**330**: 204-8.

182. Heinken A, Khan MT, Paglia G *et al.* A functional metabolic map of *Faecalibacterium prausnitzii,* a beneficial human gut microbe. *Journal of bacteriology* 2014a: JB. 01780-14.

183. Heinken A, Khan MT, Paglia G *et al.* Functional metabolic map of *Faecalibacterium prausnitzii*, a beneficial human gut microbe. *Journal of Bacteriology* 2014b;**196**: 3289-302.

184. Heintz-Buschart A, May P, Laczny CC *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology* 2017;**2**: 16180.

185. Hekstra DR, Leibler S. Contingency and statistical laws in replicate microbial closed ecosystems. *Cell* 2012;**149**: 1164-73.

186. Henderson PA, Magurran AE. Direct evidence that density-dependent regulation underpins the temporal stability of abundant species in a diverse animal community. *Proc R Soc B* 2014;**281**: 20141336.

187. Henz SR, Huson DH, Auch AF *et al.* Whole-genome prokaryotic phylogeny. *Bioinformatics* 2004;**21**: 2329-35.

188. Hildebrand F, Tadeo R, Voigt AY *et al.* LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* 2014;**2**: 30.

189. Hinnebusch BF, Meng S, Wu JT *et al.* The effects of short-chain fatty acids on human colon cancer cell phenotype are associated with histone hyperacetylation. *The Journal of Nutrition* 2002;**132**: 1012-7.

190. Holdeman LV, Moore W. New genus, *Coprococcus*, twelve new species, and emended descriptions of four previously described species of bacteria from human feces. *International Journal of Systematic and Evolutionary Microbiology* 1974;**24**: 260-77.

191. Holdeman LV, Moore W. New Genus, *Coprococcus*, Twelve New Species, and Emended Descriptions of Four Previouly Described Species of Bacteria from Human Feces. *International Journal of Systematic and Evolutionary Microbiology* 1975;**25**: 98-.

192. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one* 2012;**7**: e30126.

193. Holmstrøm K, Collins MD, Møller T *et al. Subdoligranulum variabile* gen. nov., sp. nov. from human feces. *Anaerobe* 2004;**10**: 197-203.

194. Hunter S, Jones P, Mitchell A *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* 2012;**40**: 4725.

195.   Huse SM, Ye Y, Zhou Y *et al.* A core human microbiome as viewed through 16S rRNA sequence clusters. *PloS one* 2012;**7**: e34242.

196.   Huson DH, Beier S, Flade I *et al.* MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology* 2016;**12**: e1004957.

197.   Huss VAR, Festl H, Schleifer KH. Studies on the spectrophotometric determination of DNA hybridization from renaturation rates. *Systematic and Applied Microbiology* 1983;**4**: 184-92.

198.   Huttenhower C, Gevers D, Knight R *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**: 207.

199.   Huys G, D'Haene K, Swings J. Influence of the culture medium on antibiotic susceptibility testing of food-associated lactic acid bacteria with the agar overlay disc diffusion method. *Letters in Applied Microbiology* 2002;**34**: 402-6.

200.   Ilhan ZE, Marcus AK, Kang D-W *et al.* pH-Mediated Microbial and Metabolic Interactions in Fecal Enrichment Cultures. *mSphere* 2017;**2**: e00047-17.

201.   Ioannidis JP. Why most published research findings are false. *PLoS medicine* 2005;**2**: e124.

202.   Jalanka-Tuovinen J, Salonen A, Nikkilä J *et al.* Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. *PloS one* 2011;**6**: e23035.

203.   Jarzembowska M, Sousa DZ, Beyer F *et al. Lachnotalea glycerini* gen. nov., sp. nov., an anaerobe isolated from a nanofiltration unit treating anoxic groundwater. *International journal of systematic and evolutionary microbiology* 2016;**66**: 774-9.

204.   Jeffery IB, Claesson MJ, O'Toole PW *et al.* Categorization of the gut microbiota: enterotypes or gradients? *Nature Reviews Microbiology* 2012;**10**.

205.   Jeraldo P, Hernandez A, Nielsen HB *et al.* Capturing One of the Human Gut Microbiome's Most Wanted: Reconstructing the Genome of a Novel Butyrate-Producing, Clostridial Scavenger from Metagenomic Sequence Data. *Frontiers in Microbiology* 2016;**7**: 783.

206.   Jia X, Dini-Andreote F, Salles JF. Community Assembly Processes of the Microbial Rare Biosphere. *Trends in microbiology* 2018.

207.   Jiang W, Wu N, Wang X *et al.* Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. *Scientific reports* 2015;**5**.

208.   Johansson ME, Phillipson M, Petersson J *et al.* The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the national academy of sciences* 2008;**105**: 15064-9.

209.   Johnson EL, Heaver SL, Walters WA *et al.* Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes. *Journal of Molecular Medicine* 2017;**95**: 1-8.

210.   Johnson JW, Fisher JF, Mobashery S. Bacterial cell-wall recycling. *Annals of the New York Academy of Sciences* 2013;**1277**: 54-75.

211.   Jones BV, Begley M, Hill C *et al.* Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences* 2008;**105**: 13580-5.

212.   Jones SE, Lennon JT. Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences* 2010;**107**: 5881-6.

213.   Kabeerdoss J, Sankaran V, Pugazhendhi S *et al. Clostridium leptum* group bacteria abundance and diversity in the fecal microbiota of patients with inflammatory bowel disease: a case–control study in India. *BMC gastroenterology* 2013;**13**: 1.

214.	Kageyama A, Benno Y. Emendation of genus *Collinsella* and proposal of *Collinsella stercoris* sp. nov. and *Collinsella intestinalis* sp. nov. *International journal of systematic and evolutionary microbiology* 2000;**50**: 1767-74.

215.	Kamada N, Chen GY, Inohara N *et al.* Control of pathogens and pathobionts by the gut microbiota. *Nature immunology* 2013;**14**: 685.

216.	Kämpfer P, Kroppenstedt RM. Numerical analysis of fatty acid patterns of coryneform bacteria and related taxa. *Canadian Journal of Microbiology* 1996;**42**: 989-1005.

217.	Kang D-W, Park JG, Ilhan ZE *et al.* Reduced incidence of *Prevotella* and other fermenters in intestinal microflora of autistic children. *PloS one* 2013;**8**: e68322.

218.	Kang S, Denman SE, Morrison M *et al.* Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. *Inflammatory bowel diseases* 2010;**16**: 2034-42.

219.	Kankainen M, Paulin L, Tynkkynen S *et al.* Comparative genomic analysis of *Lactobacillus rhamnosus* GG reveals pili containing a human-mucus binding protein. *Proceedings of the National Academy of Sciences* 2009;**106**: 17193-8.

220.	Kant R, Rasinkangas P, Satokari R *et al.* Genome sequence of the butyrate-producing anaerobic bacterium *Anaerostipes hadrus* PEL 85. *Genome Announcements* 2015;**3**.

221.	Karst SM, Dueholm MS, McIlroy SJ *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat Biotech* 2018;**36**: 190.

222.	Karst SM, Kirkegaard RH, Albertsen M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv* 2016: 059121.

223.	Kasai C, Sugimoto K, Moritani I *et al.* Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC gastroenterology* 2015;**15**: 1.

224.	Kassinen A, Krogius-Kurikka L, Mäkivuokko H *et al.* The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects. *Gastroenterology* 2007;**133**: 24-33.

225.	Kearney SM, Gibbons SM, Poyet M *et al.* Endospores and other lysis-resistant bacteria comprise a widely shared core community within the human microbiota. *The ISME journal* 2018: 1.

226.	Kelsen JR, Wu GD. The gut microbiota, environment and diseases of modern society. *Gut Microbes* 2012;**3**: 374-82.

227.	Kim D, Song L, Breitwieser FP *et al.* Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* 2016.

228.	Klaassens ES, De Vos WM, Vaughan EE. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Applied and environmental microbiology* 2007;**73**: 1388-92.

229.	Klampfer L, Huang J, Sasazuki T *et al.* Inhibition of Interferon γ Signaling by the Short Chain Fatty Acid Butyrate. *Molecular Cancer Research* 2003;**1**: 855-62.

230.	Kläring K, Hanske L, Bui N *et al. Intestinimonas butyriciproducens* gen. nov., sp. nov., a butyrate-producing bacterium from the mouse intestine. *International Journal of Systematic and Evolutionary Microbiology* 2013;**63**: 4606-12.

231.	Kläring K, Just S, Lagkouvardos I *et al. Murimonas intestini* gen. nov., sp. nov., an acetate-producing bacterium of the family Lachnospiraceae isolated from the mouse gut. *International Journal of Systematic and Evolutionary Microbiology* 2015;**65**: 870-8.

232.	Knights D, Ward TL, McKinlay CE *et al.* Rethinking "enterotypes". *Cell host & microbe* 2014;**16**: 433-7.

233. Kolmeder CA, De Been M, Nikkilä J *et al.* Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS one* 2012;**7**: e29913.

234. Kolmeder CA, Ritari J, Verdam FJ *et al.* Colonic metaproteomic signatures of active bacteria and the host in obesity. *Proteomics* 2015;**15**: 3544-52.

235. Konikoff T, Gophna U. Oscillospira: a central, enigmatic component of the human gut microbiota. *Trends in microbiology* 2016;**24**: 523-4.

236. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012;**28**: 3211-7.

237. Koren O, Goodrich JK, Cullender TC *et al.* Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* 2012;**150**: 470-80.

238. Koren O, Knights D, Gonzalez A *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol* 2013;**9**: e1002863.

239. Korpela K. mare: Microbiota Analysis in R Easily. R package version 1.0. 2016, DOI http://doi.org/10.5281/zenodo.50310.

240. 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**: e90702.

241. Korpela K, Salonen A, Virta LJ *et al.* Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. *Nat Commun* 2016;**7**.

242. Koskella B, Hall LJ, Metcalf CJE. The microbiome beyond the horizon of ecological and evolutionary theory. *Nature ecology & evolution* 2017;**1**: 1606.

243. 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. *Environmental microbiology* 2009;**11**: 914-26.

244. Kruse T, Levisson M, de Vos WM *et al.* vanI: a novel d-Ala-d-Lac vancomycin resistance gene cluster found in *Desulfitobacterium hafniense*. *Microbial Biotechnology* 2014;**7**: 456-66.

245. Kultima JR, Sunagawa S, Li J *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PloS one* 2012;**7**: e47656.

246. Kurtz S, Phillippy A, Delcher A *et al.* Versatile and open software for comparing large genomes. *Genome Biology* 2004;**5**: R12.

247. Kurtz ZD, Müller CL, Miraldi ER *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 2015a;**11**: e1004226.

248. Kurtz ZD, Müller CL, Miraldi ER *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 2015b;**11**: e1004226.

249. Kuykendall L, Roy M, O'neill J *et al.* Fatty acids, antibiotic resistance, and deoxyribonucleic acid homology groups of *Bradyrhizobium japonicum*. *International Journal of Systematic and Evolutionary Microbiology* 1988;**38**: 358-61.

250. Lagier J-C, Khelaifia S, Alou MT *et al.* Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature microbiology* 2016;**1**: 16203.

251. Lagkouvardos I, Fischer S, Kumar N *et al.* Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* 2017;**5**: e2836.

252. Lagkouvardos I, Weinmaier T, Lauro FM *et al.* Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. 2014;**8**: 115-25.

253. Lahti L, Salojärvi J, Salonen A *et al.* Tipping elements in the human intestinal ecosystem. *Nat Commun* 2014;**5**.

254. Lahti L, Shetty SA. Tools for microbiome analysis in R. 2018.

255. Lang JM, Eisen JA, Zivkovic AM. The microbes we eat: abundance and taxonomy of microbes consumed in a day's worth of meals for three diet types. *PeerJ* 2014;**2**: e659.

256.   Langille MG, Brinkman FS. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 2009;**25**: 664-5.

257.   Lawson PA, Finegold SM. Reclassification of *Ruminococcus obeum* as *Blautia obeum* comb. nov. *International journal of systematic and evolutionary microbiology* 2015;**65**: 789-93.

258.   Le Cao K-A, Costello M-E, Lakis VA *et al.* MixMC: a multivariate statistical framework to gain insight into microbial communities. *PloS one* 2016;**11**: e0160169.

259.   Le Chatelier E, Nielsen T, Qin J *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* 2013;**500**: 541-6.

260.   Lederberg J, McCray AT. Ome SweetOmics--A Genealogical Treasury of Words. *The Scientist* 2001;**15**: 8-.

261.   Leth ML, Ejby M, Workman C *et al.* Differential bacterial capture and transport preferences facilitate co-growth on dietary xylan in the human gut. *Nature microbiology* 2018: 1.

262.   Levine JM, D'Antonio CM. Elton revisited: a review of evidence linking diversity and invasibility. *Oikos* 1999: 15-26.

263.   Levine UY, Looft T, Allen HK *et al.* Butyrate-producing bacteria, including mucin degraders, from the swine intestinal tract. *Applied and Environmental Microbiology* 2013;**79**: 3879-81.

264.   Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences* 2013;**110**: 12804-9.

265.   Levy R, Borenstein E. Metagenomic systems biology and metabolic modeling of the human microbiome: from species composition to community assembly rules. *Gut microbes* 2014;**5**: 265-70.

266.   Ley JD, Cattoir H, Reynaerts A. The quantitative measurement of DNA hybridization from renaturation rates. *European Journal of Biochemistry* 1970;**12**: 133-42.

267.   Ley RE, Lozupone CA, Hamady M *et al.* Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* 2008;**6**: 776-88.

268.   Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 2006;**124**: 837-48.

269.   Li D, Liu C-M, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**: 1674-6.

270.   Li F, Hinderberger J, Seedorf H *et al.* Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. *Journal of Bacteriology* 2008;**190**: 843-50.

271.   Li J, Jia H, Cai X *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotech* 2014;**32**: 834-41.

272.   Li K, Bihan M, Methé BA. Analyses of the stability and core taxonomic memberships of the human microbiome. *PLoS One* 2013;**8**: e63139.

273.   Li K, Bihan M, Yooseph S *et al.* Analyses of the microbial diversity across the human microbiome. *PloS one* 2012;**7**: e32118.

274.   Li SS, Zhu A, Benes V *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* 2016;**352**: 586-9.

275.   Li XZ, Livermore DM, Nikaido H. Role of efflux pump(s) in intrinsic resistance of *Pseudomonas aeruginosa*: resistance to tetracycline, chloramphenicol, and norfloxacin. *Antimicrobial Agents and Chemotherapy* 1994;**38**: 1732-41.

276.   Liao B, Ning Z, Cheng K *et al.* iMetaLab 1.0: A web platform for metaproteomics data analysis. *Bioinformatics* 2018.

277.   Liu Y, Leal NA, Sampson EM *et al.* PduL is an evolutionarily distinct phosphotransacylase involved in B12-dependent 1, 2-propanediol degradation by *Salmonella enterica* serovar *Typhimurium* LT2. *Journal of bacteriology* 2007;**189**: 1589-96.

278.   Louis P, Flint HJ. Formation of propionate and butyrate by the human colonic microbiota. *Environmental microbiology* 2017;**19**: 29-41.

279.   Louis P, Young P, Holtrop G *et al.* Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. *Environmental Microbiology* 2010;**12**: 304-14.

280.   Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 2014;**15**: 550.

281.   Lozupone CA, Li M, Campbell TB *et al.* Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* 2013;**14**: 329-39.

282.   Lozupone CA, Stombaugh JI, Gordon JI *et al.* Diversity, stability and resilience of the human gut microbiota. *Nature* 2012;**489**: 220.

283.   Ludwig W, Schleifer K-H, Whitman WB. Revised road map to the phylum Firmicutes*Bergey's Manual® of Systematic Bacteriology*: Springer, 2009, 1-13.

284.   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**: 119-29.

285.   Lynch MD, Neufeld JD. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology* 2015;**13**: 217.

286.   Macfarlane G, Englyst H. Starch utilization by the human large intestinal microflora. *Journal of Applied Bacteriology* 1986;**60**: 195-201.

287.   MADDEN RH. Isolation and Characterization of Clostridium stercorarium sp. nov., Cellulolytic Thermophile. 1983.

288.   Magnúsdóttir S, Heinken A, Kutt L *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotech* 2017;**35**: 81.

289.   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.

290.   Mai-Prochnow A, Webb JS, Ferrari BC *et al.* Ecological advantages of autolysis during the development and dispersal of *Pseudoalteromonas tunicata* biofilms. *Applied and environmental microbiology* 2006;**72**: 5414-20.

291.   Manimaran S, Bendall M, Diaz S *et al.* PathoStat: PathoStat Statistical Microbiome Analysis Package. R package version 1.6.1, . 2018, DOI 10.18129/B9.bioc.PathoStat.

292.   Manrique P, Bolduc B, Walk ST *et al.* Healthy human gut phageome. *Proceedings of the National Academy of Sciences* 2016;**113**: 10400-5.

293.   Marcille F, Gomez A, Joubert P *et al.* Distribution of genes encoding the trypsin-dependent lantibiotic ruminococcin A among bacteria isolated from human fecal microbiota. *Applied and environmental microbiology* 2002;**68**: 3424-31.

294.   Markowitz VM, Chen I-MA, Palaniappan K *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research* 2012;**40**: D115-D22.

295.   Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* 2011;**17**: pp. 10-2.

296.   Martin ME, Solnick JV. The gastric microbial community, Helicobacter pylori colonization, and disease. *Gut microbes* 2014;**5**: 345-50.

297.   Martínez I, Lattimer JM, Hubach KL *et al.* Gut microbiome composition is linked to whole grain-induced immunological improvements. *The ISME journal* 2013;**7**: 269-80.

298.   Martinez X, Pozuelo M, Pascal V *et al.* MetaTrans: an open-source pipeline for metatranscriptomics. *Scientific reports* 2016;**6**: 26447.

299. 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**: 1591-4.

300. McCann KS. The diversity–stability debate. *Nature* 2000;**405**: 228.

301. McGill BJ, Etienne RS, Gray JS *et al.* Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters* 2007;**10**: 995-1015.

302. McMurdie PJ, S H. Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking. *Bioinformatics* 2014;**31**: 282-3.

303. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* 2013;**8**: e61217.

304. Meier-Kolthoff JP, Auch AF, Klenk H-P *et al.* Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC bioinformatics* 2013;**14**: 60.

305. Meyer F, Paarmann D, D'Souza M *et al.* The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 2008;**9**: 386.

306. Miller LT. Single derivatization method for routine analysis of bacterial whole-cell fatty acid methyl esters, including hydroxy acids. *Journal of Clinical Microbiology* 1982;**16**: 584-6.

307. Minot S, Bryson A, Chehoud C *et al.* Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences* 2013;**110**: 12450-5.

308. Minot S, Sinha R, Chen J *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research* 2011;**21**: 1616-25.

309. Miquel S, Martin R, Rossi O *et al. Faecalibacterium prausnitzii* and human intestinal health. *Current opinion in microbiology* 2013;**16**: 255-61.

310. Moeller AH, Caro-Quintero A, Mjungu D *et al.* Cospeciation of gut microbiota with hominids. *Science* 2016;**353**: 380-2.

311. Moore W, Holdeman LV. Human fecal flora: the normal flora of 20 Japanese-Hawaiians. *Applied microbiology* 1974;**27**: 961-79.

312. Moore WC, Johnson J, Holdeman L. Emendation of *Bacteroidacea*e and *Butyrivibrio* and descriptions of *Desulfomonas* gen. nov. and ten new species in the genera *Desulfomonas, Butyrivibrio, Eubacterium, Clostridium*, and *Ruminococcus*. *International Journal of Systematic and Evolutionary Microbiology* 1976;**26**: 238-52.

313. Morgan XC, Tickle TL, Sokol H *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012;**13**: R79.

314. Morin A, Urban J, Sliz P. A quick guide to software licensing for the scientist-programmer. *PLoS computational biology* 2012;**8**: e1002598.

315. Morotomi M, Nagai F, Watanabe Y. Description of *Christensenella minuta* gen. nov., sp. nov., isolated from human faeces, which forms a distinct branch in the order Clostridiales, and proposal of Christensenellaceae fam. nov. *International journal of systematic and evolutionary microbiology* 2012;**62**: 144-9.

316. Mukhopadhyay A, Redding AM, Joachimiak MP *et al.* Cell-wide responses to low-oxygen exposure in *Desulfovibrio vulgaris* Hildenborough. *Journal of Bacteriology* 2007;**189**: 5996-6010.

317. Muth T, Behne A, Heyer R *et al.* The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of proteome research* 2015;**14**: 1557-65.

318. Mutshinda CM, O'Hara RB, Woiwod IP. What drives community dynamics? *Proceedings of the Royal Society of London B: Biological Sciences* 2009;**276**: 2923-9.

319. Nakamura S, Oku T, Ichinose M. Bioavailability of cellobiose by tolerance test and breath hydrogen excretion in humans. *Nutrition* 2004;**20**: 979-83.

320.   Nam Y-D, Jung M-J, Roh SW *et al.* Comparative analysis of Korean human gut microbiota by barcoded pyrosequencing. *PLoS One* 2011;**6**: e22109.

321.   Ni Y, Li J, Panagiotou G. COMAN: a web server for comprehensive metatranscriptomics analysis. *BMC genomics* 2016;**17**: 622.

322.   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 Biotech* 2014;**32**: 822-8.

323.   Nieuwdorp M, De Vos WM. Method for preventing and/or treating insulin resistance: Google Patents, 2012.

324.   Nishikawa J, Kudo T, Sakata S *et al.* Diversity of mucosa-associated microbiota in active and inactive ulcerative colitis. *Scandinavian journal of gastroenterology* 2009;**44**: 180-6.

325.   Niu B, Paulson JN, Zheng X *et al.* Simplified and representative bacterial community of maize roots. *Proceedings of the National Academy of Sciences* 2017: 201616148.

326.   Noguera-Julian M, Rocafort M, Guillén Y *et al.* Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine* 2016;**5**: 135-46.

327.   Nurk S, Meleshko D, Korobeynikov A *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome research* 2017: gr. 213959.116.

328.   O'Keefe SJ, Li JV, Lahti L *et al.* Fat, fibre and cancer risk in African Americans and rural Africans. *Nat Commun* 2015;**6**.

329.   Oksanen J, Blanchet FG, Kindt R *et al.* Package 'vegan'. *Community ecology package, version* 2013;**2**.

330.   Okusu H, Ma D, Nikaido H. AcrAB efflux pump plays a major role in the antibiotic resistance phenotype of Escherichia coli multiple-antibiotic-resistance (Mar) mutants. *Journal of Bacteriology* 1996;**178**: 306-8.

331.   Ottman N, Davids M, Suarez-Diez M *et al.* Genome-scale model and omics analysis of metabolic capacities of Akkermansia muciniphila reveal a preferential mucin-degrading lifestyle. *Applied and environmental microbiology* 2017: AEM. 01014-17.

332.   Ottman N, Smidt H, De Vos WM *et al.* The function of our microbiota: who is out there and what do they do. *Front Cell Infect Microbiol* 2012;**2**: 3389.

333.   Ouwerkerk JP, van der Ark KC, Davids M *et al.* Adaptation of *Akkermansia muciniphila* to the oxic-anoxic interface of the mucus layer. *Applied and environmental microbiology* 2016;**82**: 6983-93.

334.   Ozer EA, Allen JP, Hauser AR. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* 2014;**15**: 737.

335.   Paine RT. Food web complexity and species diversity. *American Naturalist* 1966: 65-75.

336.   Paine RT, Tegner MJ, Johnson EA. Compounded perturbations yield ecological surprises. *Ecosystems* 1998;**1**: 535-45.

337.   Park BH, Karpinets TV, Syed MH *et al.* CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* 2010;**20**: 1574-84.

338.   Parks DH, Chuvochina M, Waite DW *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotech* 2018.

339.   Pasolli E, Schiffer L, Manghi P *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature methods* 2017;**14**: 1023.

340.   Paulson JN, Stine OC, Bravo HC *et al.* Differential abundance analysis for microbial marker-gene surveys. *Nature methods* 2013;**10**: 1200.

341.   Peng Y, Leung HC, Yiu S-M *et al.* IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012;**28**: 1420-8.

342. Pepper JW, Rosenfeld S. The emerging medical ecology of the human gut microbiome. *Trends in ecology & evolution* 2012;**27**: 381-4.

343. Pereira FC, Berry D. Microbial nutrient niches in the gut. *Environmental microbiology* 2017;**19**: 1366-78.

344. Pérez-Cobas AE, Artacho A, Ott SJ *et al.* Structural and functional changes in the gut microbiota associated to Clostridium difficile infection. *Front Microbiol* 2014;**5**: 335.

345. Pianka ER. On r-and K-selection. *The American Naturalist* 1970;**104**: 592-7.

346. Piccolo BD, Wankhade UD, Chintapalli SV *et al.* Dynamic Assessment of Microbial Ecology (DAME): A web app for interactive analysis and visualization of microbial sequencing data. *Bioinformatics* 2017;**1**: 3.

347. Plugge CM, Zoetendal EG, Stams AJM. *Caloramator coolhaasii* sp. nov., a glutamate-degrading, moderately thermophilic anaerobe. *International Journal of Systematic and Evolutionary Microbiology* 2000;**50**: 1155-62.

348. Png CW, Lindén SK, Gilshenan KS *et al.* Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. *The American journal of gastroenterology* 2010;**105**: 2420-8.

349. Poelarends GJ, Mazurkiewicz P, Putman M *et al.* An ABC-type multidrug transporter of *Lactococcus lactis* possesses an exceptionally broad substrate specificity. *Drug Resistance Updates* 2000;**3**: 330-4.

350. Price JR, Woloszynek S, Rosen GL *et al.* theseus-An R package for the analysis and visualization of microbial community data. *bioRxiv* 2018: 295675.

351. Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 2012;**28**: 1823-9.

352. Qin J, Li R, Raes J *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *nature* 2010;**464**: 59-65.

353. Qin J, Li Y, Cai Z *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**: 55-60.

354. Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 2013;**41**: D590-D6.

355. Rajilić-Stojanović M, de Vos WM. The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiology Reviews* 2014;**38**: 996-1047.

356. Rajilić-Stojanović M, Heilig HGHJ, 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. *Environmental Microbiology* 2009;**11**: 1736-51.

357. Rajilić-Stojanović 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. *Environmental microbiology* 2009;**11**: 1736-51.

358. Rajilić-Stojanović M, Heilig HG, Tims S *et al.* Long-term monitoring of the human intestinal microbiota composition. *Environmental microbiology* 2013;**15**: 1146-59.

359. Rajilić–Stojanović 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**: 1792-801.

360. Rakoff-Nahoum S, Foster KR, Comstock LE. The evolution of cooperation within the gut microbiota. *Nature* 2016;**533**: 255.

361. Ramiro-Garcia J, Hermes GD, Giatsis C *et al.* NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes. *F1000Research* 2016;**5**.

362. Raymond F, Ouameur AA, Déraspe M *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME Journal* 2015.

363. Reichardt N, Duncan SH, Young P *et al.* Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *The ISME journal* 2014;**8**: 1323.

364. Rettedal EA, Gumpert H, Sommer MOA. Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nature Communication* 2014;**5**.

365. Rey FE, Gonzalez MD, Cheng J *et al.* Metabolic niche of a prominent sulfate-reducing human gut bacterium. *Proceedings of the National Academy of Sciences* 2013;**110**: 13582-7.

366. Reyes A, Blanton LV, Cao S *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proceedings of the National Academy of Sciences* 2015;**112**: 11941-6.

367. Reyes A, Haynes M, Hanson N *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 2010;**466**: 334-8.

368. Rice KC, Bayles KW. Molecular control of bacterial death and lysis. *Microbiology and Molecular Biology Reviews* 2008;**72**: 85-109.

369. Ritari J, Salojärvi J, Lahti L *et al.* Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC genomics* 2015;**16**: 1056.

370. Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 2015;**43**: e47-e.

371. Roberts DW, Roberts MDW. Package 'labdsv'. *Ordination and Multivariate* 2016.

372. Rocha ER, Smith CJ. Regulation of *Bacteriodes fragilis* katB mRNA by oxidative stress and carbon limitation. *Journal of Bacteriology* 1997;**179**: 7033-9.

373. Rocha ER, Smith CJ. Role of the alkyl hydroperoxide reductase (ahpCF) gene in oxidative stress defense of the obligate anaerobe *Bacteroides fragilis*. *Journal of Bacteriology* 1999;**181**: 5701-10.

374. Rodionov DA, Hebbeln P, Eudes A *et al.* A novel class of modular transporters for vitamins in prokaryotes. *Journal of Bacteriology* 2009;**191**: 42-51.

375. Rodriguez-R LM, Konstantinidis KT. Bypassing cultivation to identify bacterial species. *Microbe* 2014;**9**: 111-8.

376. Roh SW, Abell GC, Kim K-H *et al.* Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in biotechnology* 2010;**28**: 291-9.

377. Rosenfeld S. Why do high-dimensional networks seem to be stable?-A reflection on stochasticity of dynamically unstable nonlinear systems*Fifth Workshop on Computation of Biochemical Pathways and Genetic Networks, University of Heidelberg, Heidelberg* volume 10112, 2008.

378. Rosindell J, Hubbell SP, Etienne RS. The unified neutral theory of biodiversity and biogeography at age ten. *Trends in ecology & evolution* 2011;**26**: 340-8.

379. Ross MC, Muzny DM, McCormick JB *et al.* 16S gut community of the Cameron County Hispanic Cohort. *Microbiome* 2015;**3**: 7.

380. Rossen NG, Fuentes S, Boonstra K *et al.* The musosa-associated microbiota of PSC patients is characterized by a low diversity and a low abundance of Uncultured Clostridiales II. *Journal of Crohn's and Colitis* 2014: jju023.

381. Rothschild D, Weissbrod O, Barkan E *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* 2018;**555**: 210.

382. Rowan F, Docherty NG, Murphy M *et al. Desulfovibrio* bacterial species are increased in ulcerative colitis. *Diseases of the Colon & Rectum* 2010;**53**: 1530-6.

383. Sa'ad H, Peppelenbosch MP, Roelofsen H *et al.* Biological effects of propionic acid in humans; metabolism, potential applications and underlying mechanisms. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* 2010;**1801**: 1175-83.

384.   Sahota SS, Bramley PM, Menzies IS. The fermentation of lactulose by colonic bacteria. *Microbiology* 1982;**128**: 319-25.

385.   Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 1987;**4**: 406-25.

386.   Sakamoto M, Benno Y. Reclassification of *Bacteroides distasonis*, *Bacteroides goldsteinii* and *Bacteroides merdae* as *Parabacteroides distasonis* gen. nov., comb. nov., *Parabacteroides goldsteinii* comb. nov. and *Parabacteroides merdae* comb. nov. *International journal of systematic and evolutionary microbiology* 2006;**56**: 1599-605.

387.   Sakamoto M, Iino T, Ohkuma M. *Faecalimonas umbilicata* gen. nov., sp. nov., isolated from human faeces, and reclassification of *Eubacterium contortum*, *Eubacterium fissicatena* and *Clostridium oroticum* as *Faecalicatena contorta* gen. nov., comb. nov., *Faecalicatena fissicatena* comb. nov. and *Faecalicatena orotica* comb. nov. *International journal of systematic and evolutionary microbiology* 2017;**67**: 1219-27.

388.   Salazar N, López P, Valdés L *et al.* Microbial targets for the development of functional foods accordingly with nutritional and immune parameters altered in the elderly. *Journal of the American College of Nutrition* 2013;**32**: 399-406.

389.   Salonen A, de Vos WM. Impact of diet on human intestinal microbiota and health. *Annual review of food science and technology* 2014;**5**: 239-62.

390.   Salonen A, Lahti L, Salojärvi J *et al.* Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *The ISME journal* 2014;**8**: 2218-30.

391.   Salonen A, Nikkilä 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. *Journal of microbiological methods* 2010;**81**: 127-34.

392.   Salonen A, Salojärvi J, Lahti L *et al.* The adult intestinal core microbiota is determined by analysis depth and health status. *Clinical Microbiology and Infection* 2012;**18**: 16-20.

393.   Salyers A, Vercellotti J, West S *et al.* Fermentation of mucin and plant polysaccharides by strains of Bacteroides from the human colon. *Applied and Environmental Microbiology* 1977;**33**: 319-22.

394.   Sankaran K, Holmes S. structSSI: simultaneous and selective inference for grouped or hierarchically structured data. *Journal of statistical software* 2014;**59**: 1.

395.   Sansonetti P. To be or not to be a pathogen: that is the mucosally relevant question. *Mucosal immunology* 2011;**4**: 8-14.

396.   Santiago A, Panda S, Mengels G *et al.* Processing faecal samples: a step forward for standards in microbial community analysis. *BMC microbiology* 2014;**14**: 1.

397.   Santos F, Spinler JK, Saulnier DM *et al.* Functional identification in *Lactobacillus reuteri* of a PocR-like transcription factor regulating glycerol utilization and vitamin B 12 synthesis. *Microbial cell factories* 2011;**10**: 55.

398.   Sayers EW, Barrett T, Benson DA *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* 2012;**40**: D13-D25.

399.   Sayin Sama I, Wahlström A, Felin J *et al.* Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring FXR antagonist. *Cell Metabolism* 2013;**17**: 225-35.

400.   Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Human molecular genetics* 2010;**19**: R227-R40.

401.   Scheffer M, Carpenter S, Foley JA *et al.* Catastrophic shifts in ecosystems. *Nature* 2001;**413**: 591-6.

402.   Scheffer M, van Nes EH. Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences* 2006;**103**: 6230-5.

403.	Scheffer M, van Nes EH, Vergnon R. Toward a unifying theory of biodiversity. *Proceedings of the National Academy of Sciences* 2018;**115**: 639-41.

404.	Scheperjans F, Aho V, Pereira PA *et al.* Gut microbiota are related to Parkinson's disease and clinical phenotype. *Movement Disorders* 2015;**30**: 350-8.

405.	Scher JU, Sczesnak A, Longman RS *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *elife* 2013;**2**.

406.	Schirmer M, Smeekens SP, Vlamakis H *et al.* Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* 2016;**167**: 1125-36. e8.

407.	Schloss PD, Westcott SL, Ryabin T *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* 2009;**75**: 7537-41.

408.	Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;**27**: 863-4.

409.	Scholz M, Ward DV, Pasolli E *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature methods* 2016a.

410.	Scholz M, Ward DV, Pasolli E *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature methods* 2016b;**13**: 435.

411.	Schubert AM, Rogers MA, Ring C *et al.* Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *MBio* 2014;**5**: e01021-14.

412.	Schumann P. Peptidoglycan structure*Methods in microbiology* volume 38: Elsevier, 2011, 101-29.

413.	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. *Systematic and applied microbiology* 2002;**25**: 46-51.

414.	Scott KP, Martin JC, Campbell G *et al.* Whole-genome transcription profiling reveals genes up-regulated by growth on fucose in the human gut bacterium "*Roseburia inulinivorans*". *Journal of Bacteriology* 2006;**188**: 4340-9.

415.	Scott KP, Martin JC, Duncan SH *et al.* Prebiotic stimulation of human colonic butyrate-producing bacteria and bifidobacteria, in vitro. *FEMS microbiology ecology* 2014;**87**: 30-40.

416.	Sekirov I, Russell SL, Antunes LCM *et al.* Gut microbiota in health and disease. *Physiological reviews* 2010;**90**: 859-904.

417.	Selvendran RR. The plant cell wall as a source of dietary fiber: chemistry and structure. *American Journal of Clinical Nutrition (USA)* 1984.

418.	Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* 2016;**164**: 337-40.

419.	Shade A, Jones SE, Caporaso JG *et al.* Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio* 2014;**5**: e01371-14.

420.	Sheng Y, Abreu IA, Cabelli DE *et al.* Superoxide dismutases and superoxide reductases. *Chemical Reviews* 2014;**114**: 3854-918.

421.	Sheridan P, Martin JC, Lawley TD *et al.* Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes. *Microbial Genomics* 2016;**2**.

422.	Shetty SA, Hugenholtz F, Lahti L *et al.* Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS microbiology reviews* 2017a;**41**: 182-99.

423.	Shetty SA, Ritari J, Paulin L *et al.* Complete Genome Sequence of *Eubacterium hallii* Strain L2-7. *Genome announcements* 2017b;**5**: e01167-17.

424.	Shetty SA, Zuffa S, Bui TPN *et al.* Reclassification of *Eubacterium hallii* as *Anaerobutyricum hallii* gen. nov., comb. nov., and description of *Anaerobutyricum soehngenii*

sp. nov., a butyrate and propionate-producing bacterium from infant faeces. *International Journal of Systematic and Evolutionary Microbiology* 2018.

425.   Shoaie S, Karlsson F, Mardinoglu A *et al.* Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific reports* 2013;**3**: 2532.

426.   Shoemaker WR, Locey KJ, Lennon JT. A macroecological theory of microbial biodiversity. *Nature ecology & evolution* 2017;**1**: 0107.

427.   Singh P, Teal TK, Marsh TL *et al.* Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome* 2015;**3**: 45.

428.   Slyter LL. Influence of acidosis on rumen function. *Journal of Animal Science* 1976;**43**: 910-29.

429.   Smits LP, Bouter KE, de Vos WM *et al.* Therapeutic potential of fecal microbiota transplantation. *Gastroenterology* 2013;**145**: 946-53.

430.   Son JS, Zheng LJ, Rowehl LM *et al.* Comparison of fecal microbiota in children with autism spectrum disorders and neurotypical siblings in the simons simplex collection. *PloS one* 2015;**10**: e0137725.

431.   Sonnenburg ED, Zheng H, Joglekar P *et al.* Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* 2010;**141**: 1241-52.

432.   Stams AJM, Van Dijk JB, Dijkema C *et al.* Growth of syntrophic propionate-oxidizing bacteria with fumarate in the absence of methanogenic bacteria. *Applied and Environmental Microbiology* 1993;**59**: 1114-9.

433.   Stein RR, Bucci V, Toussaint NC *et al.* Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol* 2013;**9**: e1003388.

434.   Stevens JR, Jones TR, Lefevre M *et al.* SigTree: a microbial community analysis tool to identify and visualize significantly responsive branches in a phylogenetic tree. *Computational and structural biotechnology journal* 2017;**15**: 372-8.

435.   Stickland LH. Studies in the metabolism of the strict anaerobes (genus *Clostridium*): The chemical reactions by which *Cl. sporogenes* obtains its energy. *Biochemical Journal* 1934;**28**: 1746-59.

436.   Sun Y, Cai Y, Liu L *et al.* ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research* 2009;**37**: e76.

437.   Sun Z, Harris HM, McCann A *et al.* Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* 2015;**6**: 8322.

438.   Sunagawa S, Mende DR, Zeller G *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods* 2013;**10**: 1196.

439.   Tahar Mechichi ML, Woo H. *Eubacterium aggregans* sp. nov., a new homoacetogenic bacterium from olive mill wastewater treatment digestor. *Anaerobe* 1998;**4**: 283-91.

440.   Tailford LE, Crost EH, Kavanaugh D *et al.* Mucin glycan foraging in the human gut microbiome. *Frontiers in genetics* 2015;**6**.

441.   Tamura K, Stecher G, Peterson D *et al.* MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* 2013;**30**: 2725-9.

442.   Tap J, Furet JP, Bensaada M *et al.* Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults. *Environmental microbiology* 2015;**17**: 4954-64.

443.   Tap J, Mondot S, Levenez F *et al.* Towards the human intestinal microbiota phylogenetic core. *Environmental microbiology* 2009;**11**: 2574-84.

444.   Taras D, Simmering R, Collins MD *et al.* Reclassification of Eubacterium formicigenerans Holdeman and Moore 1974 as Dorea formicigenerans gen. nov., comb. nov., and description of Dorea longicatena sp. nov., isolated from human faeces. *International journal of systematic and evolutionary microbiology* 2002;**52**: 423-8.
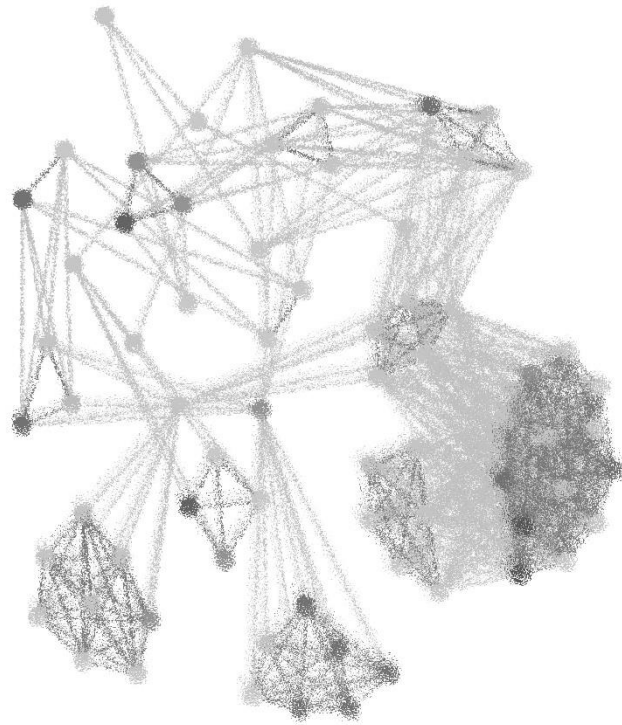
445.   ter Steege H, Pitman NCA, Sabatier D *et al.* Hyperdominance in the Amazonian Tree Flora. *Science* 2013;**342**.

446.   Thaiss CA, Zeevi D, Levy M *et al.* Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* 2014;**159**: 514-29.

447.   Thangaraju M, Cresci GA, Liu K *et al.* GPR109A is a G-protein–coupled receptor for the bacterial fermentation product butyrate and functions as a tumor suppressor in colon. *Cancer research* 2009;**69**: 2826-32.

448.   Thauer RK, Jungermann K, Decker K. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriological reviews* 1977;**41**: 100.

449.   Thompson LR, Sanders JG, McDonald D *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. 2017;**551**: 457.

450.   Tien M-T, Girardin SE, Regnault B *et al.* Anti-inflammatory effect of *Lactobacillus casei* on *Shigella*-infected human intestinal epithelial cells. *The Journal of Immunology* 2006;**176**: 1228-37.

451.   Tilman D. Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences* 2004;**101**: 10854-61.

452.   Tims S, Derom C, Jonkers DM *et al.* Microbiota conservation and BMI signatures in adult monozygotic twins. *The ISME journal* 2013;**7**: 707-17.

453.   Topping DL, Clifton PM. Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. *Physiological reviews* 2001;**81**: 1031-64.

454.   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**: e62544.

455.   Treangen TJ, Koren S, Sommer DD *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology* 2013;**14**: R2.

456.   Trosvik P, Muinck EJ. Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa. *Microbiome* 2015;**3**: 1.

457.   Truong DT, Franzosa EA, Tickle TL *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* 2015;**12**: 902.

458.   Truong DT, Tett A, Pasolli E *et al.* Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research* 2017.

459.   Turnbaugh PJ, Hamady M, Yatsunenko T *et al.* A core gut microbiome in obese and lean twins. *Nature* 2009;**457**: 480-4.

460.   Turroni F, Ribbera A, Foroni E *et al.* Human gut microbiota and bifidobacteria: from composition to functionality. *Antonie Van Leeuwenhoek* 2008;**94**: 35-50.

461.   Udayappan S, Manneras-Holm L, Chaplin-Scott A *et al.* Oral treatment with *Eubacterium hallii* improves insulin sensitivity in db/db mice. *NPJ biofilms and microbiomes* 2016a;**2**: 16009.

462.   Udayappan S, Manneras-Holm L, Chaplin-Scott A *et al.* Oral treatment with *Eubacterium hallii* improves insulin sensitivity in db/db mice. 2016b;**2**: 16009.

463.   Van Belkum A, Struelens M, de Visser A *et al.* Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clinical microbiology reviews* 2001;**14**: 547-60.

464.   Van den Abbeele P, Belzer C, Goossens M *et al.* Butyrate-producing Clostridium cluster XIVa species specifically colonize mucins in an in vitro gut model. *The ISME journal* 2013;**7**: 949-61.

465.   van Gelder AH, Aydin R, Alves MM *et al.* 1,3-Propanediol production from glycerol by a newly isolated *Trichococcus* strain. *Microbial Biotechnology* 2012;**5**: 573-8.

466.   van Nes EH, Pujoni DGF, Shetty SA *et al.* Rarity as a Sticky State. DOI Under preparation.

467. Vandermeer JH. Niche theory. *Annual review of Ecology and Systematics* 1972;**3**: 107-32.

468. Vellend M. Conceptual synthesis in community ecology. *The Quarterly review of biology* 2010;**85**: 183-206.

469. Venables WN, Smith DM. The R development core team. *An Introduction to R R Foundation for Statistical Computing, Vienna, Austria* 2006.

470. Venema K, Van den Abbeele P. Experimental models of the gut microbiome. *Best Practice & Research Clinical Gastroenterology* 2013;**27**: 115-26.

471. Venturelli OS, Carr AV, Fisher G *et al.* Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular systems biology* 2018;**14**: e8157.

472. Verberkmoes NC, Russell AL, Shah M *et al.* Shotgun metaproteomics of the human distal gut microbiota. *The ISME journal* 2009;**3**: 179-89.

473. Verdam FJ, Fuentes S, de Jonge C *et al.* Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity* 2013;**21**: E607-E15.

474. Vieira-Silva S, Falony G, Darzi Y *et al.* Species–function relationships shape ecological properties of the human gut microbiome. *Nature microbiology* 2016;**1**: 16088.

475. Vital M, Howe A, Bergeron N *et al.* Metagenomic insights into resistant starch degradation by human gut microbiota. *Appl Environ Microbiol* 2018: AEM. 01562-18.

476. Vital M, Howe AC, Tiedje JM. Revealing the bacterial butyrate synthesis pathways by analyzing (meta) genomic data. *MBio* 2014a;**5**: e00889-14.

477. Vital M, Howe AC, Tiedje JM. Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *mBio* 2014b;**5**.

478. Vital M, Karch A, Pieper DH. Colonic Butyrate-Producing Communities in Humans: an Overview Using Omics Data. *MSystems* 2017;**2**: e00130-17.

479. Vos P, Garrity G, Jones D *et al. Bergey's Manual of Systematic Bacteriology: Volume 3: The Firmicutes* volume 3: Springer Science & Business Media, 2011.

480. Wacklin P, Tuimala J, Nikkilä J *et al.* Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. *PloS one* 2014;**9**: e94863.

481. Wade WG. The genus *Eubacterium* and related genera*The prokaryotes*: Springer, 2006, 823-35.

482. Walker B, Holling CS, Carpenter SR *et al.* Resilience, adaptability and transformability in social--ecological systems. *Ecology and society* 2004;**9**: 5.

483. Waller AS, Yamada T, Kristensen DM *et al.* Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME journal* 2014;**8**: 1391-402.

484. Walther R, Hippe H, Gottschalk G. Citrate, a specific substrate for the isolation of Clostridium sphenoides. *Applied and environmental microbiology* 1977;**33**: 955-62.

485. Wang M, Molin G, Ahrné S *et al.* High proportions of proinflammatory bacteria on the colonic mucosa in a young patient with ulcerative colitis as revealed by cloning and sequencing of 16S rRNA genes. *Digestive diseases and sciences* 2007a;**52**: 620-7.

486. Wang Q, Garrity GM, Tiedje JM *et al.* Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* 2007b;**73**: 5261-7.

487. Washburne AD, Silverman JD, Leff JW *et al.* Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* 2017;**5**: e2969.

488. Weghoff MC, Bertsch J, Müller V. A novel mode of lactate metabolism in strictly anaerobic bacteria. *Environmental microbiology* 2015;**17**: 670-7.

489. Weirather JL, de Cesare M, Wang Y *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 2017;**6**.

490.   Weiss S, Van Treuren W, Lozupone C *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal* 2016.

491.   Welch JLM, Hasegawa Y, McNulty NP *et al.* Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proceedings of the National Academy of Sciences* 2017: 201711596.

492.   Whitford M, Yanke L, Forster R *et al.* Lachnobacterium bovis gen. nov., sp. nov., a novel bacterium isolated from the rumen and faeces of cattle. *International journal of systematic and evolutionary microbiology* 2001;**51**: 1977-81.

493.   Wickham H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* 2011;**3**: 180-5.

494.   Wickham H. Tidyverse: Easily install and load'tidyverse'packages. *R package version* 2017;**1**.

495.   Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS microbiology reviews* 2011;**35**: 957-76.

496.   Wolin MJ, Miller TL, Collins MD *et al.* Formate-Dependent Growth and Homoacetogenic Fermentation by a Bacterium from Human Feces: Description of Bryantella formatexigens gen. nov., sp. nov. *Applied and environmental microbiology* 2003;**69**: 6321-6.

497.   Wong VW-S, Tse C-H, Lam TT-Y *et al.* Molecular characterization of the fecal microbiota in patients with nonalcoholic steatohepatitis–a longitudinal study. *PloS one* 2013;**8**: e62885.

498.   Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 2014;**15**: R46.

499.   Wootton L. Phage biology: A new barrier at mucosal surfaces. *Nature Reviews Microbiology* 2013;**11**: 430-1.

500.   Wu C-C, Johnson J, Moore W *et al.* Emended descriptions of *Prevotella denticola*, *Prevotella loescheii*, *Prevotella veroralis*, and *Prevotella melaninogenica*. *International Journal of Systematic and Evolutionary Microbiology* 1992;**42**: 536-41.

501.   Wu GD, Chen J, Hoffmann C *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* 2011;**334**: 105-8.

502.   Xiao L, Feng Q, Liang S *et al.* A catalog of the mouse gut metagenome. *Nat Biotech* 2015;**33**: 1103-8.

503.   Xiong H, Guo B, Gan Z *et al.* Butyrate upregulates endogenous host defense peptides to enhance disease resistance in piglets via histone deacetylase inhibition. *Scientific reports* 2016;**6**: 27070.

504.   Yachi S, Loreau M. Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proceedings of the National Academy of Sciences* 1999;**96**: 1463-8.

505.   Yanling H, Youfang D, YANQUAN L. Two Cellulolytic Clostridium Species: Clostridium cellulosi sp. nov. and Clostridium cellulofermentans sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 1991;**41**: 306-9.

506.   Yatsunenko T, Rey FE, Manary MJ *et al.* Human gut microbiome viewed across age and geography. *Nature* 2012;**486**: 222-7.

507.   Youngster I, Sauk J, Pindar C *et al.* Fecal microbiota transplant for relapsing *Clostridium difficile* infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clinical Infectious Diseases* 2014;**58**: 1515-22.

508.   Zackular JP, Baxter NT, Iverson KD *et al.* The gut microbiome modulates colon tumorigenesis. *MBio* 2013;**4**: e00692-13.

509.   Zakrzewski M, Proietti C, Ellis JJ *et al.* Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions. *Bioinformatics* 2016;**33**: 782-3.

510.   Ze X, Duncan SH, Louis P *et al. Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *The ISME journal* 2012;**6**: 1535-43.

511.   Zeevi D, Korem T, Zmora N *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* 2015;**163**: 1079-94.

512.   Zeller G, Tap J, Voigt AY *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* 2014;**10**: 766.

513.   Zhang J, Guo Z, Lim AAQ *et al.* Mongolians core gut microbiota and its correlation with seasonal dietary changes. *Scientific reports* 2014;**4**.

514.   Zhang J, Guo Z, Xue Z *et al.* A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. *The ISME journal* 2015;**9**: 1979-90.

515.   Zhang X, Smits AH, van Tilburg GB *et al.* Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nature protocols* 2018;**13**: 530.

516.   Zhang Z, Zhai H, Geng J *et al.* Large-scale survey of gut microbiota associated with MHE Via 16S rRNA-based pyrosequencing: Nature Publishing Group, 2013.

517.   Zhao S, Lieberman TD, Poyet M *et al.* Adaptive evolution within the gut microbiome of individual people. *bioRxiv* 2017: 208009.

518.   Zhou Y, Liang Y, Lynch KH *et al.* PHAST: a fast phage search tool. *Nucleic acids research* 2011: gkr485.

519.   Zijlmans MA, Korpela K, Riksen-Walraven JM *et al.* Maternal prenatal stress is associated with the infant intestinal microbiota. *Psychoneuroendocrinology* 2015;**53**: 233-45.

520.   Zitomersky NL, Atkinson BJ, Franklin SW *et al.* Characterization of adherent bacteroidales from intestinal biopsies of children and young adults with inflammatory bowel disease. *PloS one* 2013;**8**: e63686.

521.   Zoetendal E, Rajilić-Stojanović M, De Vos W. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* 2008;**57**: 1605-15.

522.   Zoetendal EG, de Vos WM. Effect of diet on the intestinal microbiota and its activity. *Current opinion in gastroenterology* 2014;**30**: 189-95.

523.   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. *The ISME journal* 2012;**6**: 1415.

524.   Zoetendal EG, Vaughan EE, De Vos WM. A microbial world within us. *Molecular Microbiology* 2006;**59**: 1639-50.

525.   Zoetendal EG, von Wright A, Vilpponen-Salmela T *et al.* Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Applied and environmental microbiology* 2002;**68**: 3401-7.

526.   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**: e43052.

Appendices

+

Co-author affiliations

Willem M. de Vos[1,2,3]

Hauke Smidt[1]

Leo Lahti[4]

Floor Hugenholtz[1§]

Thi Phuong Nam Bui[1]

Steven Aalvink[1]

Simone Zuffa[1]

Luuk Mommersteeg[1]

Caroline M. Plugge[1]

Siavash Atashgahi[1]

Sjef Boeren[5]

Egbert H. van Nes[6]

Marten Scheffer[6]

Ilias Lagkouvardos[7,8]

Thomas Clavel[7,9]

Bhawani Chamlagain[10]

Francois Douillard[3]

Lars Paulin[11]

Vieno Piironen[10]

Jarmo Ritari[2]

[1]Laboratory of Microbiology, Wageningen University and Research, Stippeneng 4, Building 124, 6708 WE Wageningen, The Netherlands

[2]Department of Veterinary Biosciences, Division of Microbiology and Epidemiology, University of Helsinki, Helsinki, Finland

[3]RPU Immunobiology, Department of Bacteriology and Immunology, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[4]Department of Mathematics, University of Turku, Finland

[5]Laboratory of Biochemistry, Wageningen University, Stippeneng 4, Building 124, 6708 WE Wageningen, The Netherlands

[6]Aquatic Ecology and Water Quality Management Group, Department of Environmental Sciences, Wageningen University, Wageningen, The Netherlands

[7]Technische Universität München, ZIEL – Institute for Food & Health, Freising-Weihenstephan, Germany

[8]Technische Universität München, Junior Research Group Microbial Bioinformatics, ZIEL – Institute for Food and Health, Freising-Weihenstephan, Germany

[9]RWTH University Hospital, Institute of Medical Microbiology, Functional Microbiome Research Group, Aachen, Germany

[10]Department of Food and Environmental Sciences, University of Helsinki, Helsinki 00014, Finland.

[11]DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Acknowledgments

The journey towards achieving a doctorate in philosophy was possible only because of the support and encouragement from family, friends and colleagues. First and foremost, I dedicate my PhD to my parents and my maternal uncle. They made sure that I never faced the obstacles that had hampered their education. Being a first-generation PhD in my family fills them with pride and that is the greatest achievement for me. My brothers, Sujit and Dhiraj are the best siblings I could have asked for and they are my strength. I consider myself privileged to have a close group of friends (हम पांच ग्रुप (we five group) Prathamesh, Pramod, Pravin and Abhijeet) who have been with me for more than 18 years. The distance (6906km!!) hardly mattered in the last four and a half years. Words will never be able to express the gratitude I feel towards them. The journey for Wageningen started with me sending a short email to Prof. dr. Willem de Vos. Working with Willem has been a dream come true for me. For every meeting, I prepared myself mentally to be overwhelmed with ideas and suggestions and soon it became a habit. Now meetings without him feel extremely slow paced, vague and sometimes intellectually boring. I remember making several promises to do lots of analysis for ideas he suggested but managed to fulfill only a few of them. I know this because I went through all my previous minutes of the meetings during the last stretch of my thesis. Seriously, I don't know what I was thinking! I will always cherish all this time. Dear Willem, you have been an important part of my academic journey and I wish you all the best with the new adventures you planed for your retirement.

Dear Hauke, you have been highly supportive of my research. You brought some calmness in the avalanche of words (filled with ideas and suggestions) that were coming from Willem in our meetings. The best thing about you was that I could just enter your office to ask for anything, I could always catch you at the coffee machine and ask about random topics. The high quality of manuscripts we wrote together were possible only with your eye for details. Willem and Hauke, it's been a privilege to be mentored by you in this journey and I would not like to change a single moment of my PhD.

Dear Leo, you were the right person to compliment Willem and Hauke for supervision of my thesis. We were friends before you became my co-promotor. The numerous discussions we had ranged from the trivial of things in life to the most complicated of philosophies which were always enlightening. I finally managed to give you the real experience of "chai" with Anne in my home city. I hope we continue our friendship for many years to come.

Dear Floor Hugenholtz, thank you for helping me in my early days in Wageningen. Especially, for helping me to prepare myself for the power packed meetings with Willem. I hope we continue to collaborate in future.

Dear Alfons JM Stams, I cannot express my gratitude in words towards you for consistently pushing me to do some "actual" work in the lab, amazing scientific discussions we had over coffee, for answering my basic questions regarding physiology and any other topics. I wish that I can incorporate my career, your approach for scientific research.

The combination of molecular ecology and microbial physiology in this thesis was a result of me being embraced by colleagues and friends from both MolEco and MicFys research groups.

This thesis was part of the SIAM project. The amazing team with diverse background and incredible young researchers meant that SIAM activities were filled with fun and amazing science. It was a privilege to work with Francisca, Melvin, Yuan, Yifan, Martijn, Daan, Sigrid, Marissa, Laura, Saara, Jules, Leonor, Cornelia and Subhash.

My office mates, in the old building I shared my office with Clara, Nirunya and Leire. It was a pleasure to spend half my PhD with you next to me. The second half was with Johanna, Menia, Jie, Leire (Yes! you were my officemate in the first and the last year, happy coincidence), Dat, Atsushi, LooWee, Yangshang and Kate which included the phase of me having lot of ups and downs and plus writing of the thesis! Thank you.

Dear Johanna and Menia thanks for being there with me, I could always count on you.... Thanks for listening to my random news from India during our many breaks.

My MolEco colleagues and friends (Erwin, Detmer, Indra, Joan, Alex, Hugo, Hikma, Catarina, Carrie, Jannie, Janneke, Lennart, Sharon, Klaudyna, Carmen, Tom, Siavash, Taojun, Tikka, Emmy, Romy, Nora, Ran, Caifang, Dennis, Thomas K, Hugo, Ying, Peng, Annelies, Marjet), MicFys colleagues and friends (Caroline, Vicente, Samet, Susakul, Martijn, Daan, Yuan, Lot, Noemhi, Ivette, Nico V, Conall) and from BacGen (Ioannis M, Elleke, Nico C, Raymond, Serve, Mihris, Max, Wen, Joyshree, Prarthana, Jorrit, Jeroen, Melvin, Franklin, Despoina). Colleagues and friends at SSB (Nirupama, Erika, Jasper, Nong, Wasin, Maria, Eduardo, Dorett, Bart, Stamatis, Ruben). Special thanks to Nam for helping me in understanding the anaerobic physiology of human intestinal microbiome and being good friend.

Nachiket, Mandar, Swapnil, Shreyas, Dhiraj, Shrikant, Vikas P, Vikas G. for supporting me while my work in India.

A lot can be learned from the students, I gained experience and motivation from Simone, Bob, Ben, Luuk, and Johnathan during my thesis. Simone, Ben and Luuk, the success in our projects led to some of the interesting findings described in this thesis. Bob and Johnathan, you worked on challenging topic and the results from your work have been building blocks for our current quest to culture *Oscillospira*.

Special thanks to Wim, Ineke, Steven, Anja, Heidi, Ton, Tom, Merlijn, Iamme, Sjon, Phillippe and Wilma for all the assistance in the lab. Steven and Ineke for also being my confidants for non-work related things. Steven, thanks for helping me out with the Dutch summary of my thesis

My first exposure to diverse cultures in NL was at Plantsoen with awesome housemates (Matteo, Rachelle, Valeria, Yolanda, Mikaela, Debora, Abdulrahman, Federico, Natalia, Pau, Max, Miguel, Armand, Florence, Susie, Margaux, Matthias, Rafael, Lucile). Matteo, Federico, Abdulrahman some of my friends informed me that we looked like the mafia of Wageningen when we walked the streets in the center and when having good times at the entrance of Plantsoen and the smokey environment we created. I cherish our moments together for my life. Rachelle, Valeria, Debora, Mikaela and Yolanda (also my biking partner), you amazing girls from Plantsoen! It is a pleasure to be friends with you. Although I met quite a few people at parties in Wageningen, one person I will always remember is Camille (Mimi). We continued our friendship from one of the student parties (during which I got a free crash course in Salsa!) and had good fun times, cooking and our squash playing sessions and I wish you all the best with your social work in Mayotte and beyond. I would also thank my current housemates at

About the author

Sudarshan Anand Shetty was born on November 18th, 1988 in Pune, India. He studied Microbiology BSc (2007-2009) and MSc (2009-2011) at Department of Microbiology, Modern College of Arts, Science and Commerce, University of Pune. During his MSc study, he developed interest in human intestinal microbiome and did his thesis at the National Centre for Cell Science (NCCS), Pune, India. In 2012, he joined as a junior research fellow (JRF) at the National Centre for Microbial Resource (NCMR) at NCCS, Pune, India, to pursue his scientific interest in understanding the human intestinal microbiome. During this period, he published research on various aspects of the human intestinal microbiome in the context of Indian population.

On August 7th, 2014, Sudarshan started his PhD research at the Laboratory of Microbiology at Wageningen University and Research, under the guidance of Prof. Willem M. de Vos and Prof. Hauke Smidt in collaboration with Dr. Leo Lahti from University of Turku, Finland. The research described in this thesis is a culmination of scholarly discussions that have happened between him, his promotors, co-promotor and several colleagues from the human intestinal microbiome field. Apart from this, he has developed keen interest in open science data initiatives and aims to inculcate this in his current and future research. He has organised and assisted in scientific workshops that promote open science. He also understands the importance of science communication and is an active blog writer for the Gut Microbiota for Health web portal, created by the Gut Microbiota and Health Section of the European Society for Neurogastroenterology & Motility (ESNM), member of United European Gastroenterology (UEG). Currently, he works as a post-doctoral research at the Laboratory of Microbiology at Wageningen University and Research.

List of publications

2019

Bodkhe R*, <u>Shetty SA*</u>, Dhotre DP, Verma AK, Bhatia K, Mishra A, Kaur G, Pande P, Bangarusamy DK, Santosh BP, Perumal RC. Comparison of small gut and whole gut microbiota of first-degree relatives with adult patients with celiac disease and controls. *Frontiers in Microbiology*. 2019 10:164. doi: 10.3389/fmicb.2019.00164.

2018

<u>Shetty SA</u>#, Zuffa S, Bui TP, Aalvink S, Smidt H, De Vos WM. Reclassification of *Eubacterium hallii* as *Anaerobutyricum hallii* gen. nov., comb. nov., and description of *Anaerobutyricum soehngenii* sp. nov., a butyrate and propionate-producing bacterium from infant faeces. *International Journal of Systematic and Evolutionary Microbiology*. 2018 Oct 23. (<u>Chapter 6</u>)

Atashgahi S, <u>Shetty SA</u>, Smidt H, De Vos WM. Flux, impact and fate of halogenated xenobiotic compounds in the gut. *Frontiers in physiology*. 2018;9:888.

Ghattargi VC, Nimonkar YS, Burse SA, Davray D, Kumbhare SV, <u>Shetty SA</u>, Gaikwad MA, Suryavanshi MV, Doijad SP, Utage B, Sharma OP. Genomic and physiological analyses of an indigenous strain, *Enterococcus faecium* 17OM39. *Functional & integrative genomics*. 2018 Jul 1;18(4):385-99.

2017

<u>Shetty SA</u>, Ritari J, Paulin L, Smidt H, De Vos WM. Complete Genome Sequence of *Eubacterium hallii* Strain L2-7. *Microbiology Resource Announcements* (formerly *Genome announcements*). 2017 Oct 26;5(43):e01167-17.

<u>Shetty SA</u>#, Hugenholtz F, Lahti L, Smidt H, de Vos WM. Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiology Reviews*. 2017 Mar 1;41(2):182-199. (<u>Chapter 2</u>)

2016

Bui TP, <u>Shetty SA</u>, Lagkouvardos I, Ritari J, Chamlagain B, Douillard FP, Paulin L, Piironen V, Clavel T, Plugge CM, de Vos WM. Comparative genomics and physiology of the butyrate-producing bacterium *Intestinimonas butyriciproducens. Environmental Microbiol Reports.* 2016 Dec;8(6):1024-1037. (<u>Chapter 5</u>)

Buntin N, Hongpattarakere T, Ritari J, Douillard FP, Paulin L, Boeren S, <u>Shetty SA</u>, de Vos WM. An inducible operon is involved in inulin utilization in *Lactobacillus plantarum* strains, as revealed by comparative proteogenomics and metabolic profiling. *Applied and Environmental Microbiology*. 2016 Dec 30;83(2). pii: e02402-16.

Marathe NP*, <u>Shetty SA*</u>, Shouche YS, Larsson DG. Limited bacterial diversity within a treatment plant receiving antibiotic-containing waste from bulk drug production. *PLoS One*. 2016 Nov 3;11(11):e0165914.

Bhute S, Pande P, <u>Shetty SA</u>, Shelar R, Mane S, Kumbhare SV, Gawali A, Makhani H, Navandar M, Dhotre D, Lubree H, Agarwal D, Patil R, Ozarkar S, Ghaskadbi S, Yajnik C, Juvekar S, Makharia GK, Shouche YS. Molecular characterization and meta-analysis of gut microbial communities illustrate enrichment of *Prevotella* and *Megasphaera* in Indian subjects. *Frontiers in Microbiology*. 2016 May 9; 7:660.

Paul D, Kumbhare SV, Mhatre SS, Chowdhury SP, <u>Shetty SA</u>, Marathe NP, Bhute S, Shouche YS. Exploration of microbial diversity and community structure of lonar lake:

the only hypersaline meteorite crater lake within basalt rock. *Frontiers in microbiology*. 2016 Jan 22;6:1553.

2015

Pandey PK, Verma P, Marathe N, Shetty S, Bavdekar A, Patole MS, Stensvold CR, Shouche YS. Prevalence and subtype analysis of *Blastocystis* in healthy Indian individuals. *Infection Genetics and Evolution*. 2015 Apr; 31:296-9.

2014

Kulkarni GJ, Shetty S, Dharne MS, Shouche YS. Genome sequencing analysis reveals virulence-related gene content of *Ochrobactrum intermedium* strain 229E, a urease-positive strain isolated from the human gastric niche. *FEMS Microbiology Letters*. 2014 Oct;359(1):12-5.

Marathe NP*, Shetty SA*, Lanjekar VB, Rasane MH, Ranade DR, Shouche YS. Genome sequencing of multidrug resistant novel *Clostridium* sp. BL8 reveals its potential for pathogenicity. *Gut pathogens*. 2014 Jul;6(1):30.

Mahale KN, Paranjape PS, Marathe NP, Dhotre DP, Chowdhury S, Shetty SA, Sharma A, Sharma K, Tuteja U, Batra HV, Shouche YS. Draft genome sequences of *Yersinia pestis* strains from the 1994 plague epidemic of Surat and 2002 Shimla outbreak in India. *Indian journal of microbiology*. 2014 Dec 1;54(4):480-2.

2013

Shetty SA*, Marathe NP*, Shouche YS. Opportunities and challenges for gut microbiome studies in the Indian population. *Microbiome*. 2013 Sep 17;1(1):24.

Shetty SA*, Marathe NP*, Lanjekar V, Ranade D, Shouche YS. Comparative genome analysis of *Megasphaera* sp. reveals niche specialization and its potential role in the human gut. *PLoS One*. 2013 Nov 18; 8(11):e79353.

Kulkarni G, Dhotre D, Dharne M, Shetty S, Chowdhury S, Misra V, Misra S, Patole M, Shouche Y. Draft genome of *Ochrobactrum intermedium* strain M86 isolated from non-ulcer dyspeptic individual from India. *Gut pathogens*. 2013 Apr;5(1):7.

2012

Marathe N, Shetty S, Lanjekar V, Ranade D, Shouche Y. Changes in human gut flora with age: an Indian familial study. *BMC Microbiology*. 2012 Sep 26; 12:222.

Pawar SP, Dhotre DP, Shetty SA, Chowdhury SP, Chaudhari BL, Shouche YS. Genome sequence of *Janibacter hoylei* MTCC8307, isolated from the stratospheric air. *Journal of bacteriology*. 2012 Dec 1;194(23):6629-30.

Manuscripts in review/ preparation /pre-print

Shetty SA, Lahti L. Microbiome data science (*under review*).

Shetty SA, Smidt H, de Vos WM. Reconstructing functional networks in the human intestinal tract using synthetic microbiomes (*under review*)

van Nes EH, Pujoni DGF, Shetty SA, Straatsma G, de Vos WM, Scheffer M. Rarity as a Sticky State. (*under review*).

Shetty SA, Lahti L, van Nes EH, de Vos WM, Scheffer M, Smidt H. Hyperdominance in microbial communities. (*manuscript in preparation*).

Lahti L, Shetty SA *et al*. Tools for microbiome analysis in R. 2018-19. Version 1.5.23. URL: http://microbiome.github.com/microbiome. (*manuscript in preparation*).

<u>Shetty SA</u>, Boeren S, Smidt H, de Vos WM. Ecophysiology of *Anaerobutyricum* in the human intestinal tract. (*manuscript in preparation*).

<u>Shetty SA</u>, Mommersteeg L, Lahti L, Atashgahi S, Smidt H, de Vos WM. Design of a diet-based minimal microbiome of the human intestinal tract. (*manuscript in preparation*).

Gutleben J, Loureiro C, Ramírez Romero L, <u>Shetty SA</u>, Wijffels RH, Smidt H, Sipkema D. Cultivation of bacteria from *Aplysina Aerophoba*: effects of oxygen and nutrient gradients. (*manuscript in preparation*).

Edwards JE, van den Berg P, <u>Shetty SA</u>, Burden F, van Doorn DA, Pellikaan WF, Dijkstra J, Smidt H. The Equine Hindgut Microbiome: A Broader Perspective. (*manuscript in preparation*).

<u>Shetty SA</u>[#]. Gut microbiota features of the geographically diverse Indian population. *bioRxiv*. 2018 Jan 1:478586. (*pre-print*)

*Equal contribution; [#]Corresponding author

Overview of completed training activities

Name of the course/meeting	Organizing institute (s) and Country	Year
Discipline specific activities		
The Intestinal Microbiome and Diet in Human and Animal Health	VLAG, Netherlands	2014
Gut Day Amsterdam	Gut flora Foundation, Netherlands	2014
The Future of Pre- and Probiotics	VLAG, Netherlands	2015
KNVM/NVMM meeting	KNVM, Netherlands	2016
INRA-Rowett	INRA-France and Rowett institute Aberdeen, France	2016
SIAM Summer School	SIAM, Netherlands	2016
Fisabio Summer School in Biomedical Research and Public Health: Microbial Genomics & Metagenomics Workshop	FISABIO and JGI, Spain	2016
Genome binning-Nijmegen	SIAM, Netherlands	2016
The Intestinal Microbiome and Diet in Human and Animal Health	VLAG, Netherlands	2017
FEMS conference	FEMS, Spain	2017
Ecology workshop	KU Leuven, Belgium	2017
General courses		
VLAG PhD week	VLAG, Netherlands	2014
Competence Assessment (CA)	WGS, Netherlands	2016
WGS: PhD Workshop Carousel	VLAG, Netherlands	2015
PhD Peer Consultation - a powerful tool to tackle PhD challenges (PC)	VLAG, Netherlands	2016
SIAM writing week	SIAM, Netherlands	2017
SIAM entrepreneur pitch day	SIAM, Netherlands	2017
Mobilising your Scientific Network (MSN)	WGS, Netherlands	2018
Optional		
PhD research proposal		2014
Laboratory of Microbiology PhD study tour		2017
Weekly group meetings		2014-2018
SIAM biweekly meetings		2016-2018
PhD meetings (Biweekly)		2014-2018