Interactive Functional Networks in Microbiota

Bastian V.H. Hornung

Propositions

1. Metatranscriptomics sheds light on microbial processes contributing to the greenhouse effect and helps to decrease it.
(this Thesis)

2. Current developments in microbiome research will lead to a healthier society.
(this Thesis)

3. Besides statistics, scientists should also get education in psychology, to avoid the most common misconceptions and logical biases.

4. Overstating results will confuse scientists and will serve as reference points for future bad science.

5. Many scientific findings, including own findings, cannot be trusted, due to lack of reproducibility.

6. Current societal developments will impact the freedom of speech and thinking before they will impact the freedom of science.

7. Miscommunication is never the fault of a single person.

8. People who complain that it is difficult to manage one’s social life while managing their PhD are doing neither right.

Propositions belonging to the thesis, entitled

Interactive Functional Networks in Microbiota

Bastian Hornung

Wageningen, 1 November 2018

Interactive Functional Networks in Microbiota

Bastian V.H. Hornung

Thesis committee

Promotor

Prof. Dr H. Smidt
Personal chair at the Laboratory of Microbiology
Wageningen University & Research
Prof. Dr V.A.P. Martins dos Santos
Professor of Systems and Synthetic Biology
Wageningen University & Research

Co-promotor

Dr P.J. Schaap
Associate professor, Laboratory of Systems and Synthetic Biology
Wageningen University & Research

Other members

Prof. Dr D. de Ridder, Wageningen University & Research
Dr A.L. Zomer, Utrecht University
Dr A. Bossers, Wageningen University & Research
Dr J.P. Hays, Erasmus Medical Center, Rotterdam

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE).

Interactive Functional Networks in Microbiota

Bastian V.H. Hornung

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 Thursday 1 November 2018
at 11 a.m. in the Aula.

Bastian V.H. Hornung
Interactive Functional Networks in Microbiota,
176 pages.
PhD thesis, Wageningen University, Wageningen, NL (2018)
With references, with summary in English
ISBN: 978-94-6343-486-7
DOI: https://doi.org/10.18174/456782

Table of contents

Chapter 1: General introduction

How biology changes over time

The field of systems biology is a research area that has emerged in over the last two decades. While biology (especially molecular biology) was initially dominated by a reductionist approach, with investigating and understanding the make-up and functioning of simpler, isolated parts of bigger systems, systems biology aims to provide a holistic view on all the interconnected parts of the system and its emerging properties [1, 2].

Nevertheless, systems biology cannot stand on its own. Most often it is the case that a system cannot be understood if one does not understand at least part of the basic functioning of its components [3]. While in computer science a "black box" approach can be utilized in many cases to replicate behaviours, this is not necessarily the case in biology [3]. It is not possible to form hypotheses and to test them, if there is not at least a minimum of understanding of a system (or a related hypothesis).

But is this true if we look back into the history of science?

If we want to e.g. understand human physiology, we need to understand the functions of organs, for which we need to understand the functions of cells, the functions of enzymes, etc. The level of depth, granularity and extend of the necessary understanding depends on the complexity of the investigated system. Early progresses on human physiology were not made by observing humans as a system, but scientists like Leonardo da Vinci investigated the functions of organs and tissues. Da Vinci realized that it is not possible to understand how the motion of arms and legs is realized, if the function of muscles and nerves is not understood [4]. But in the meantime, with all gained knowledge, medicine has over the centuries been able to get a holistic view on humans, and diseases are cured with having the whole human as a system in mind. This happens despite the fact that we do not understand humans as a system fully and even new organs are being discovered [5, 6], but our level of understanding has advanced enough to generate and apply concepts at the systems level.

Systems biology and ecology

While many disciplines started with the reductionist approach, one biological discipline, ecology, is based on understanding systems. Since its emergence in the 18th century [7], this field has dealt with the interconnections between individual parts within its systems. Also here it is still true that the identity and functioning of components need to be understood first. Animal ecology cannot be studied if nothing is known about the animals themselves. A top down approach can only work, if in the course of the research knowledge about the parts of the system under investigation is gained [8]. This normally involves an iterative cycle, where knowledge is gained (potentially via simulations), confirmed in the laboratory, and used to re-evaluate the simulations. As an example, abnormalities observed in a predator-prey system were modelled with different hypotheses, and the only matching simulation was further successfully evaluated in the laboratory [9].

But is systems biology then equivalent to ecology? Many methodologies and thoughts are the same, as already outlined. Some of the early ecologists like e.g. Alexander von Humboldt could probably also be described as early systems biologists, since he showed his research in a holistic way, and not with a reductionist approach. The difference, if we actually would like to make it, seems to be rather an excluding one, not an including one (i.e. they do overlap in many areas). If definitions for both ecology [10] and systems biology [1] are considered, it is hard to delineate them if a common area (e.g. a microbial ecosystem) is viewed. Non-common areas will allow the distinction more easily. For example, studying the metabolism of a single microbe with tools like metabolic simulations or genetic engineering can be categorized as systems biology, but not as ecology. On the other hand, cataloguing the plants and animals of a newly discovered ecosystem is clearly ecology, but cannot be considered systems biology. The last important difference is that in contrast to ecology, the field of systems biology itself only emerged recently [2], and has become necessary mostly due to the increase of complexity in the studied fields, which began to rise in the omics era.

The matter of classifying research can be complicated, due to the fact that many research fields overlap, and it is probably also not possible to uniquely classify this thesis, in which microbial ecosystems were studied with systems biology approaches.

Systems biology and microbiology

The field of microbiology also started with a reductionist approach. Microbiology started when Antoni van Leeuwenhoek discovered microbes with his tiny microscopes, and at this time microbes also still needed to be understood themselves. It was necessary to isolate these microbes in order to allow testing hypotheses related to their physiology and behaviour. Since the field of microbiology is ever expanding - with only an estimated 0.0001% of bacteria being discovered [11] and only a fraction of microorganisms being amenable to currently available cultivation methods [12] - this problem of culturing is also still a challenge today [13]. Despite this lack of knowledge of many details, microbiological knowledge and technologies have advanced enough to make systems understanding at least partially possible. The recently developed field of microbiome research is one of the currently most obvious illustrations. With the wealth of already present knowledge, it is possible to test hypotheses at the level of microbial ecosystems. The earliest research has already been performed in the 19th century, with pioneers like Sergei Winogradsky [14] and Martinus Beijerink [15] investigating microbial communities and basic microbial processes. Despite its long history, the field of microbial ecology has taken a big leap in terms of information, understanding and concepts with the rapid development of sequencing technology over the past two decades. The decrease in price and increase in quality and output made it possible to sequence more, and more complicated nucleic acid compositions. The sequencing of community metagenomes is now one of the standards in microbiome research and contributed to re-inventing the entire field. This allows researchers to come closer to true systems thinking, although many details are still missing.

Systems biology and systems thinking versus specialization

As already pointed out, systems biology is about studying systems. These systems vary in complexity, with some being increasingly more complex. The ultimate goal would be to have an approximation of everything, to integrate models of all different kinds of systems and scales (as e.g. mentioned in [16]). But can this even be done? Nobody has all the necessary knowledge to fully understand these models, or the underlying data. While in the middle ages every scholar was a polymath, proficient in many different disciplines, this is not the case today. Also in later times, some outstanding scientists had wider knowledge. Da Vinci worked in many different fields from human biology to engineering, Leibniz in mathematics and history, Goethe in poetry and engineering, Darwin in geology and biology. Having such a diverse knowledge would enable those researchers to place findings in the right context. In contrast, today it is very difficult to keep up with new developments in a specific research field, sometimes not even with a subfield. As mentioned in chapter 2 of this thesis, even in the sub-sub-field concerning this thesis (microbial ecology as a part of microbiology or ecology as both part of biology, or of systems biology) there are more than a thousand publications per year, making it impossible to keep up with all details. Today, true systems thinking can rarely be done by individual researchers. With the increase of "big science" [17], it is rarely feasible that researchers are running a project single-handedly. The overall development of projects like space flight or the CERN would not be accomplishable by single individuals, and collaborations are necessary, sometimes reaching into hundreds of people being involved in the research alone. This also reaches to the field of biology, with big efforts like the human genome [18], or in connection to this thesis, the human microbiome [19]. A wide range of expertise is necessary to generate this data, and to interpret it. With collaborations, by recruiting the experts of different fields, true transdisciplinary systems biology at a higher level is possible. The more complex systems we investigate, the more effort becomes necessary to understand them.

Common to most of these big scale projects in the area of biology is that they often rely on high throughput data such as those generated by today's range of next generation sequencing (NGS) technologies.

The rise of nucleic acid sequencing

At the very beginning of the nucleic acid sequencing era in the 1960s and 70s [20], it was barely possible to sequence whole genes. Sequencing was labour intensive with the low throughput of the first generation sequencing technologies. The sequencing of the first genomes of viruses and bacteria were major endeavours and required a lot of time and money, whereas this is not anymore the case today. While high quality genomes are still a challenge with currently employed standard short read sequencing technologies (i.e. Illumina sequencing) [21-23], and plant and animal genomes are still a challenge due to size and structure, the generation of bacterial draft genomes has become a standard element of routine characterization of microbial isolates. The development of NGS strategies has enabled to produce millions of short nucleotide sequences ("reads"), in contrast to the 10.000s that were achievable with Sanger sequencing. Since the human microbiome can contain as many cells as the human body [24], a microbiome sequencing approach would not have been feasible with a low throughput technology. The average output of e.g. the Illumina HiSeq platform with 360 Million reads allows to have sensible coverage of some microbial communities without omitting critical parts. Within the last 50 years sequencing has evolved from being able to decipher short single nucleotide ranges up to 20 nucleotides in the very early stages, to generating millions of short reads with an increasing amount of nucleotides, or being able to sequence thousands of long reads with 50000 nucleotides and more.

What is being sequenced and which questions can this answer?

The field of microbiome research has diverged into many different directions. The first publications in the field were all exploratory. Craig Venter sequenced the metagenomes of ocean water samples [25], and metagenomes associated with farm soil and whale fall [26] were also elucidated, in order to explore phylogenetic and functional diversity of the microbiota that is present in these either exotic or important habitats. Within less than 15 years after these initial hallmark papers, many important habitats were investigated, providing important findings.

The most publicly known investigations are probably those that target the human microbiome. The Human Microbiome Project [19] and MetaHit [27] were successful in unravelling the composition and genetic blueprint of the microbiota associated with human gut, skin, vagina and other body parts. Pioneering work by Turnbaugh et *al.* [28] showed that the microbiome is significantly different between lean and obese individuals, and with crossover experiments (i.e. exchange of one microbiota with another) it could also be shown that this is not purely correlation, but that the microbiota has a causative role in this process. Other investigations showed that in chronic gut diseases (Inflammatory Bowel Disease, Ulcerative Colitis, Crohn's disease), and also in some more acute diseases like *Clostridium difficile* infections, the microbiota is significantly perturbed as compared to that of healthy individuals [29, 30]. This has led to efforts to mitigate the effects of these diseases with microbial interventions. The modern development of faecal microbiota transplant therapy [31] was only possible because of the direct indications of microbial involvement in these diseases, and the progress of the therapy could be studied with microarray- and sequencing technologies. Nevertheless, although a microbiome-inspired cure seems at reach for specific diseases [32], full understanding of the underlying mechanisms has not been achieved yet.

Humans are not the only hosts for which microbiomes are under investigation. The microbiota of animals has been investigated for various reasons. Some of these investigations were also exploratory, e.g. showing that pets have a microbiota more closely related to that of their owners than to random humans, suggesting interchange of their microbiomes [33]. Other animal species under investigation are more relevant from industrial, agricultural and/or environmental perspectives, such as ruminant farm animals with respect to their methane emission and the resulting environmental footprint.

The efficiency of feed for various animals is under investigation. It is not only investigated directly, by how feed can be utilized more efficiently [34], but also indirectly, in how feed-derived short chain fatty acids contribute to the caloric need of the animals [35]. The efficiency of these processes also directly relates to environmentally important research. Common waste products during digestion include various gasses. These can, like e.g. methane, contribute significantly to the greenhouse effect. A broad range of research on cows (e.g. [36] and chapter 5 of this thesis), sheep [37] and reindeer [38] has been performed to discover the underlying processes and to develop potential strategies to reduce the gas output by these animals.

The role of the gut microbiome in animal health is also being investigated. As an example, losses in pig production due to diarrhoea, and particularly post-weaning diarrhoea, is a major concern, and pre- or probiotic supplements might help to ameliorate this problem [39]. Biotechnological applications exist as well. The intestinal microbiota of various animals, which can degrade complicated carbohydrates, is under investigation as a resource for the discovery of novel enzymes for e.g. more efficient breakdown of organic waste streams and non-food biomass for the production of biofuel [40, 41]. The intestines of elephants [42], koalas [43], and termites [44] harbour microbes with a totally different capacity for polysaccharide and lignin breakdown, which could be utilized in such processes. The microbial fuel cell is also of industrial relevance, but not yet fully understood [45].

Besides health, environment and biotechnology (and various other areas within the fields of biology, medicine, agriculture and environmental sciences), also completely unrelated fields like history and space research are involved into microbiota research (due to similar reasons though). Investigations of the microbiota of ancient animals and humans [46, 47] showed that in general the gut microbiomes do not differ from current ones, if rural African populations are considered. In contrast, gut microbiomes resulting from a modern western lifestyle show a distinct pattern. This is also in agreement with investigations on the oral microbiome, which showed that it changed significantly over humans' history [48], in connection with presumed changes in eating habits.

In the rather unrelated area of space research, the humanities' outermost outpost, the International Space Station (ISS), also has been sampled [49], together with potential Mars habitats [50]. In these cases it was mainly shown that human presence beyond planet Earth definitely leads to a spread of microbiota in the environment, and that even the possibly cleanest environments, like cleanrooms on earth [49], do harbour their own unique microbiota. Another experiment has shown that some microbes, like e.g. *Deinococcus radiodurans*, can potentially survive extreme conditions in space [51], and for animals (the tardigrade) this has already been proven [52]. While someone might wonder, in how far this is relevant, the space agencies are undertaking great precautions not to contaminate other worlds already right now. The NASA steered their latest mission around Saturn, Cassini, into Saturn itself, so that it will not be possible for the space probe to crash onto one of the moons and to introduce potentially surviving microbes on its surface [53]. The contamination with earth microbiota might endanger the ecosystem on other worlds, in the same way that e.g. the introduction of rabbits, rats and other animals endangered the natural fauna in Australia.

Besides these examples, numerous other habitats of varying importance are sampled, including, e.g. glaciers [54], ant fungal gardens [55], extreme environments like soda lakes [56], the Tinto river [57], contaminated environments [58], kitchen sponges [59] or washing machines (own unpublished work), to just name a few. The multitude of environments, going from the deep sea [26] to outer space [49], tells us about the importance of the microbiota in all fields of life. On this planet, everything is interconnected. We are living with our microbes, the microbes are living with us, and on most places without us. But we cannot survive without them, and we will benefit from understanding their inner and outer workings.

Aim and outline of this thesis

The goal of this thesis is to explore the realms of microbial ecology with systems biology tools. Most of this work is done with next generation sequencing as the main way of data generation. An overview of the experimental chapters (chapters 3, 4 and 5), including information on the respective biomes under investigation and applied methods, is given in figure 1.

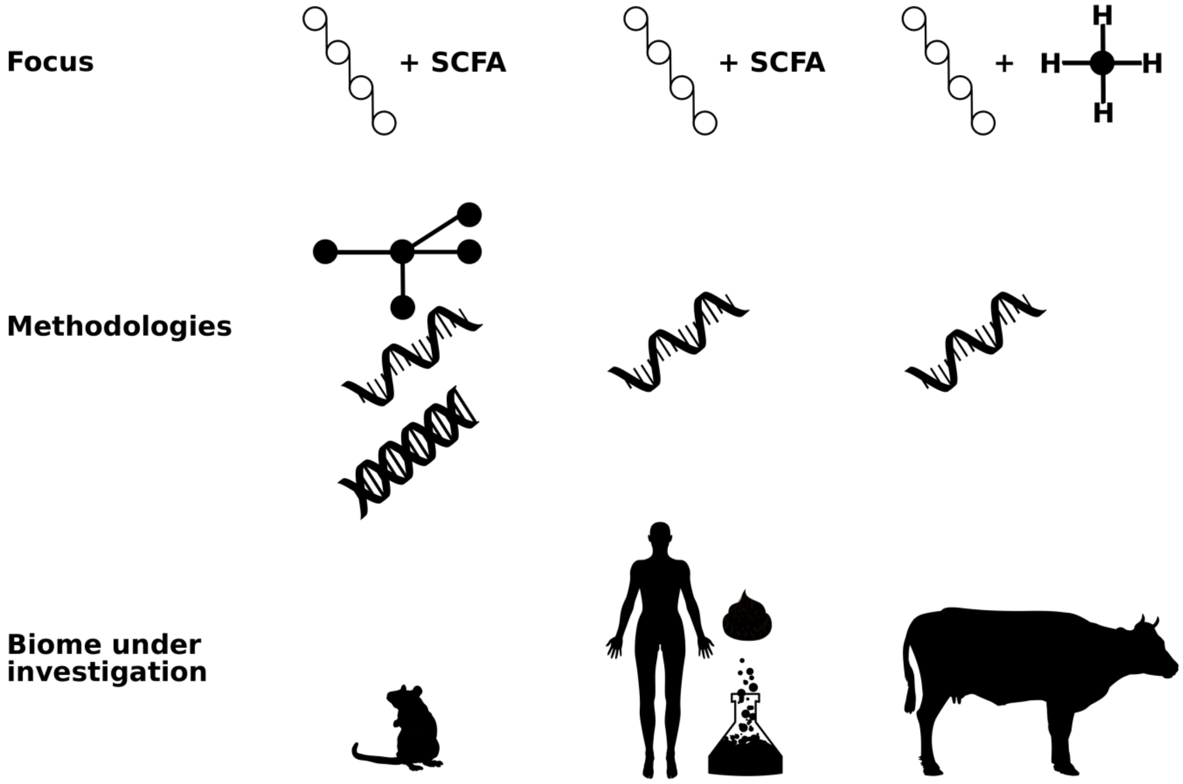


Figure 1: An overview of the biomes, methodologies and the focus in the three experimental chapters. In chapter 3 the bacterium *Romboutsia ilealis* CRIB^{T}, derived from a rat gut, was investigated with genomics, transcriptomics and metabolic modelling, with a focus on the carbohydrate degradation and SCFA production capabilities. In chapter 4, an *in vitro* fermentation system, inoculated with human faecal material, was investigated with metatranscriptomics with a similar focus. In chapter 5, the metatranscriptome of the cow rumen microbiota was sequenced, with a focus on carbohydrate degradation and methane production.

In **chapter 2**, we review how this type of work and data is beneficial for the modern human. It focuses on obesity and the metabolic syndrome, which are unarguably widespread epidemics in today's western society. Much of the ground breaking work in the microbiome field has occurred within the gut ecosystem, like demonstrating that the gut microbiota does not only change as an effect of diet, but that it can also change the metabolic state of its host and lead to obesity [28]. We summarize in this chapter how such results can be achieved by investigating the underlying technologies, starting from the different sequencing approaches, over the advantages and disadvantages of genomics, transcriptomics and other –omics, and the different methodologies in statistics, computer science and machine learning. This chapter lays the fundament for the understanding of the different approaches applied in the following research chapters.

In **chapter 3**, we dive deeper into this ecosystem, by investigating the capabilities of the ileum bacterium *Romboutsia ilealis*. This bacterium was found to be prevalent in rats, possibly related to health outcomes after gut inflammation. We in-depth characterized its metabolic capabilities, based on genomics and transcriptomics, and concluded that it is well adapted to its environment. *R. ilealis* mainly relies on the degradation of simple carbohydrates, which are abundant in the ileum, and furthermore also consumes the abundantly available vitamins and amino acids, without having the abilities to synthesize most of them.

In **chapter 4**, we change the focus to a more complex setup. An *in-vitro* human gut community, derived from human fecal samples, is studied, not relying on simple carbohydrates, but on prebiotics with more complex structures. In this setup, it was possible to show with a metatranscriptomic approach that the prebiotics not only maintain normal gut functioning, but also promote the growth of bacteria, which are in general considered to be beneficial, like *Bifidobacterium* or *Lactobacillus*. It was also possible to detect how the measured outcomes of short chain fatty acid production were achieved by the community, pinpointing the key players, and the differences in their metabolism. We hypothesize that this outcome could not have been achieved with simple carbohydrates, as it has been mentioned in an unrelated context already in earlier literature [60, 61]. As the last step in this investigation, we were able to detect how the prebiotics were degraded by the community, and that both the abundant *Bacteroides,* as well as the mentioned *Bifidobacterium* and *Lactobacillus* were playing a part in this process, and were even cooperating in some steps, which neither of them would have been capable of performing in isolation.

Chapter 5 describes an *in-vivo* study of cow rumen microbiota structure and function, being the most complicated research target tackled in this thesis. This is not only more complicated due to the fact that the animal is a less controlled environment than the *in vitro* fermentation setup employed in chapter 4, but also because the cow rumen is less well studied than the human gut. This lead to much unclassifiable data, however, we were still able to investigate the processes we were interested in. The setup had been to feed four groups of cows with diets, which differed in their maize and grass silage content, and thus in the proportions of readily digestible carbohydrates and fibres. In previous work performed by van Gastelen *et al*.[62], it was shown that the maize starch based diet lead to a reduced methane output of the rumen microbiota. This outcome is important, given that methane output of ruminants contributes considerably to the greenhouse effect (up to 35% of the anthropogenic methane production [63, 64]). Hence we were interested to elucidate the underlying biology. Using again a metatranscriptomic approach we were able to conclude that the reduced methane metabolism was an indirect effect of a metabolic change in the community. The methanogenic *Archaea* were not directly affected by the diet, however, for methanogenesis they rely on metabolites produced by other community members (presumably a member of the Clostridiales in this case), which were decreased due to the change in diet. The most beautiful part of this work was, besides the fact that it potentially could contribute to the development of novel dietary strategies towards reduced methane emissions from cows, that the data was obvious. Despite the fact that a complex and not well characterized community was studied, the data clearly displayed how methanogenesis as a whole was affected by the experimental setup. On a personal note, I was surprised by this, and hope that much of my future work will be as nice, without the need to hunt for spurious associations in weak data.

In **chapter 6**, the general discussion, I will discuss the strengths and weaknesses of this overall approach. I will also discuss the underlying problems, which resulted in the publications included in this thesis.

Chapter 2: Studying microbial functionality within the gut ecosystem by systems biology

This chapter is adapted from:

Bastian Hornung, Vitor A.P. Martins dos Santos, Hauke Smidt, Peter J. Schaap, 2018. “Studying microbial functionality within the gut ecosystem by systems biology”. In BMC Genes & Nutrition. https://doi.org/10.1186/s12263-018-0594-6

Abstract

Humans are not autonomous entities. We are all living in a complex environment, interacting not only with our peers, but as true holobionts we are also very much in interaction with our coexisting microbial ecosystems living on and especially within us, in the intestine. Intestinal microorganisms, often collectively referred to as intestinal microbiota, contribute significantly to our daily energy uptake by breaking down complex carbohydrates into simple sugars, which are fermented to short-chain fatty acids and subsequently absorbed by human cells. They also have an impact on our immune system, by suppressing or enhancing the growth of malevolent and beneficial microbes. Our lifestyle can have a large influence on this ecosystem. What and how much we consume can tip the ecological balance in the intestine. A "western diet" containing mainly processed food will have a different effect on our health than a balanced diet fortified with pre- and probiotics.

In recent years, new technologies have emerged, which made a more detailed understanding of microbial communities and ecosystems feasible. This includes progress in the sequencing of PCR-amplified phylogenetic marker genes as well as the collective microbial metagenome and metatranscriptome, allowing us to determine with an increasing level of detail, which microbial species are in the microbiota, understand what these microorganisms do and how they respond to changes in lifestyle and diet. These new technologies also include the use of synthetic and *in vitro* systems, which allow us to study the impact of substrates and addition of specific microbes to microbial communities at a high level of detail, and enable us to gather quantitative data for modelling purposes.

Here we will review the current state of microbiome research, summarizing the computational methodologies in this area, and highlighting possible outcomes for personalized nutrition and medicine.

Background

The gut is an essential part of the human body. It has so much influence on our wellbeing that it even has been dubbed a "second brain" by the media [65, 66], and in recent years this "superorgan" inhabited by trillions of microorganisms has triggered a large amount of scientific interest.

The microbial communities residing in the different parts of the gut are among the main contributors to its functioning, and therefore also directly influence health. The recent availability of high throughput methods (metagenomics, and other omics), have improved our insights into these ecosystems dramatically. Figure 1 summarizes the current state of meta-omics (all nucleotide sequencing approaches, as well as metaproteomics and meta-metabolomics) research with an intestinal focus (for details regarding the literature search methodology see additional file 1). Not surprisingly, the largest body of research has been focused on humans (Figure 1D), but other (model) organisms including pigs, rodents (mice, rats) and fishes (mainly zebrafish) have also been investigated. Non-model organisms are also under investigation, but for different purposes such as the potential biotechnological applicability of lignin degradation by termite gut microbial species [67].

Over the trajectory of the human gut, the microbiome has a varying degree of complexity [68, 69] (Figure 1 B). In general, microbial density increases from the duodenum until it reaches its maximum in the colon and faeces. At the same time these two parts are also the most studied parts (Figure 1 A). While the high complexity of the community at these specific sites makes them interesting research sites, other parts of the (healthy) human gut remain grossly under-sampled, which is mainly due to inaccessibility. Along the trajectory of the human gut, the focus of microbial metabolic activities changes profoundly, with the small intestine having a higher capacity to degrade simpler carbohydrates [70], whereas in the colon mostly complex carbohydrates are degraded [71].

Most human omics studies are observational, aimed at studying microbial diversity and function as well as host-microbe interactions, however a number of studies directly aim at improving gut health (and in proxy, individual health, Figure 1 E). These interventional studies can be broadly classified into two categories: pre-clinical and clinical interventions. Pre-clinical interventions focus mostly on improving gut health via changes in nutrition. In this field, the concept of probiotics (administering of beneficial bacteria [72]) is probably the most widely known, also in the eye of the general public, due to a wide array of commercially available products. Most interventional studies have focused on these probiotics, with a smaller part investigating the benefits of prebiotics (substrates enhancing the growth of beneficial bacteria in the gut, review see [71]). Clinical interventions in response to conditions associated with a chronic disruption of intestinal homeostasis such as ulcerative colitis, and IBS with for example faecal transplants and bariatric surgery, have only been reported in a few publications [73, 74].

With all these studies, many important factors have been discovered regarding the ecology of the human microbiome.

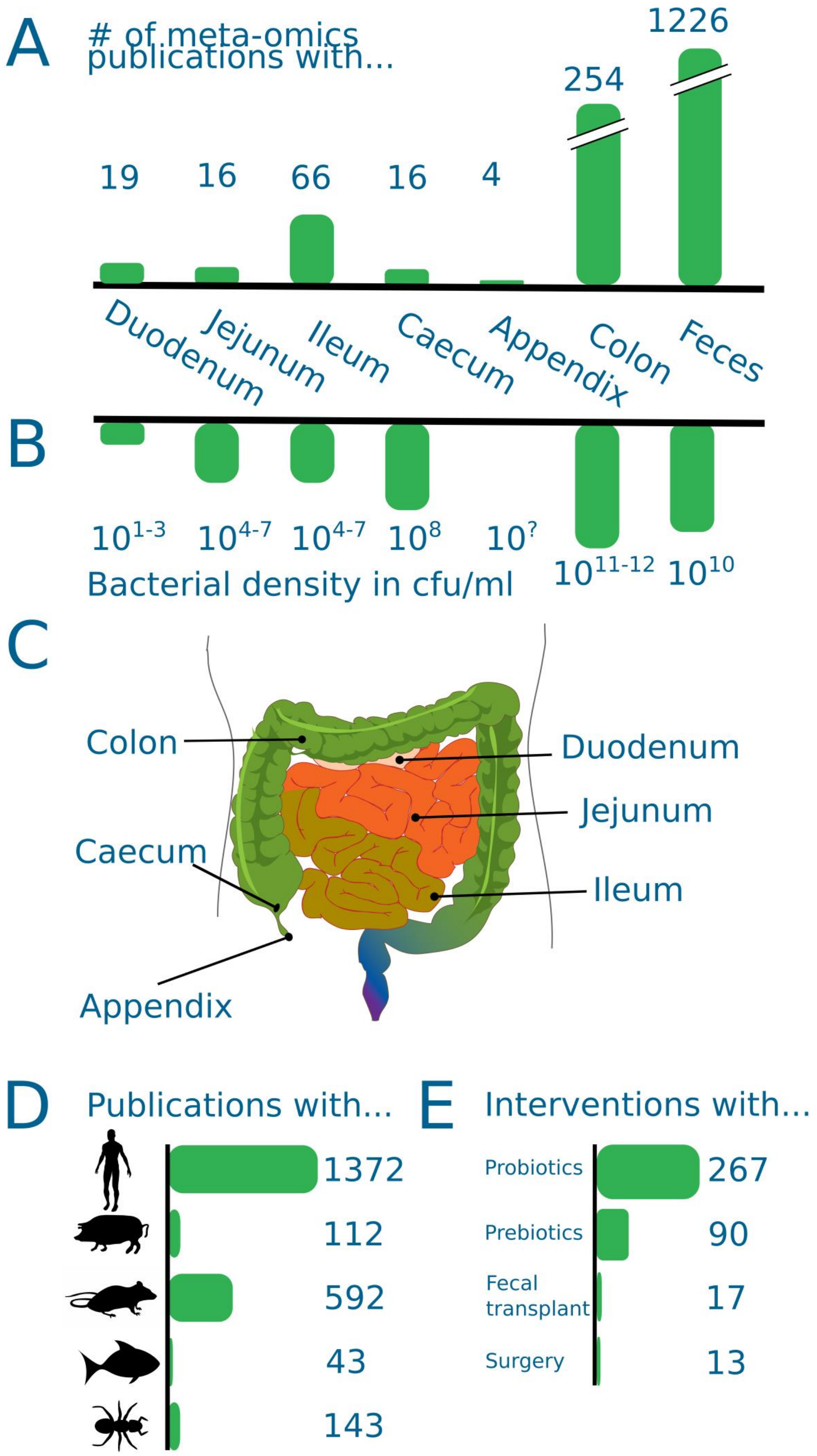


Figure 1: The gut in the focus of meta-omics science. An overview of main sampling sites and microbial complexity is given. Number of the studied hosts and methods to improve gut health are indicated. All data was retrieved via pubmed searches for the corresponding terms. For the exact search terms, please see additional file 1.

The human microbiota: Symbiosis, competition and other relationships

Our microbiota is an important part of our personal ecosystem, which is assumed to be composed of more than a trillion microbial cells [27], approximately equalling the amount of human cells in our body [24]. Whereas the microbial ecosystems associated with some niches of the human body like e.g. the vagina [75] have a low complexity with only a few different inhabitants, most body sites contain hundreds of different microbes [27]. Like in macro-ecology, they perform different roles and thus can have different relationships with each other and with the host. In the microbiota, a broad range of different interactions exist, ranging from mutualistic and commensal to predatory relationships, and competition for the same niche exists. The nature of these relationships has an impact on the habitat itself, and imbalances with respect to the abundance and function of specific members can lead to an imbalance of the whole ecosystem. Many bacteria like e.g. *Akkermansia muciniphila* [76] have a good symbiotic relationship with their host. They degrade the carbohydrates supplied by the host, and other bacteria benefit from the breakdown products of this degradation process. This leads to the production of host beneficial compounds like short chain fatty acids (SCFA; mainly acetate, propionate, butyrate) [35], which can be e.g. used by human colonocytes as energy source [77] or directly be incorporated into the human metabolism as additional carbon sources [78]. In other cases, this symbiosis applies to nutrition-derived carbohydrates that are not (fully) digested by host-derived enzymes in the small intestine such as resistant starch and other complex carbohydrates [71]. These might only be broken down by specific combinations of microorganisms for further catabolization. This can be exemplified by consortia of Bifidobacteria [79], which lead to the liberation of otherwise inaccessible substrates from e.g. indigestible plant biomass like cellulose components. In both scenarios, the liberated substrates can be further metabolized by other bacteria (e.g. [80]) to host beneficial compounds. Parasitic relationships also exist, like e.g. between *Actinomyces odontolyticus* and TM7 [81], where the parasitizing TM7 might eventually kill its microbial host. There are also predatory relationships, e.g. bacteria of the genus *Bdellovibrio* prey on other bacteria as source of energy and therefore help to regulate the diversity and balances of bacterial populations [82, 83]. Imbalances in the ecosystem might lead to bacterial overgrowth, which makes the ecosystem in general less resilient to perturbations [84]. Blooms of bacteria, e.g. *Clostridium difficile*, which infects more than half a million individuals per year and leads to 29.000 deaths in the US alone [85], will have a directly noticeable impact. The produced toxins in such an outbreak will not only affect the microbiota [86], but will also lead to a direct disease state of the host [87]. Therefore, understanding of internal and external factors that affect composition and functioning of this ecosystem, such as e.g. nutrition intake, antibiotic intake, symbiotic or predatory relationships, are essential for being able to characterize and predict the state and functioning of this ecosystem. All of these challenge the intrinsic emergent community properties such as resilience, stability and its efficiency to provide nutrients for the host.

Metabolic syndrome and the microbiome

The metabolic syndrome is a complex disorder with high associated cost, and is mainly characterized by four sub-pathologies: Obesity, elevated blood sugar/insulin resistance/diabetes type II, elevated blood pressure, and dyslipidemia [88, 89]. Although genetics [90] and lifestyle [91] play major roles, the microbiome also contributes to all of these main sub-pathologies.

Obesity might provide the most direct link. It has been shown that gut microbiota composition in obese and lean individuals is significantly different [92]. The microbiome is an important factor in carbohydrate degradation and uptake. Microbial metabolism on average contributes to up to 10% of the daily calorie intake [93], and potentially in obese subjects this contribution could be increased [28]. This is mainly due to degradation of carbohydrates, which due to the lack of necessary catabolic enzymes, are not directly accessible for the human host. These carbohydrates are converted by the microbiota into SCFA, thereby directly contributing to the energy intake of the host [94]. Since not all microorganisms are capable of such conversions, species diversity and abundance will directly influence the types of carbohydrates that can be converted into SCFA and therefore how much of the non-digestible carbohydrates will be utilized by the host-microbe holobiont. While some bacteria are specialized in carbohydrate breakdown, like e.g. *Bacteroides thetaiotaomicron* [95], others mainly rely on their peers to scavenge nutrients [96]. A microbial community consisting mainly of carbohydrate degraders will therefore be more beneficial for the host providing valuable nutrients. It is tempting to speculate that in case of obesity this beneficial trait has turned disadvantageous, and might contribute to an increased risk towards metabolic syndrome-associated pathologies.

Such differences in microbial composition have also been causally linked to obesity. It has been shown that transplantation of an “obese microbiome” into germ free animals causes an increase in body fat as compared to control animals inoculated with a “lean microbiome” [28, 97, 98], indicating that the increased capacity to harvest energy is transferred with the microbiome.

The involvement of the gut microbiome in the second most prevalent pathology, elevated blood sugar/insulin resistance leading to diabetes type II, can be explained via an indirect route, starting from inflammation. Even without an obvious disease phenotype, low grade inflammation might be present [99], caused by yet unidentified bacteria. This inflammation is hypothesized to be one of the causes of the metabolic syndrome [99, 100], and to be an early stage of Inflammatory Bowel Disease, including Ulcerative Colitis and Crohn’s Disease [101]. An invasion of bacteria into the intestinal tissue causes the presence of endotoxins (LPS, flagellin) in the blood stream, leading to chronic inflammation in the intestinal tissue. It has been suggested that as a physiological response to inflammation the blood glucose level is increased to serve as additional energy source for the various immune cells [102]. Since the inflammation is chronic, so will be the elevated glucose levels. In the long term this might lead to insulin resistance and Type II diabetes [103].

The connection between the composition of the human gut microbiota and the third and fourth pathology, elevated blood pressure and dyslipidemia, is less well characterized [104]. It has been demonstrated with cross-over experiments that gut microbiota from rats with elevated blood pressure will transfer this physiological trait to receiving rats [105]. It has also been shown that inflammatory processes [106] and effects on the nervous system [107] will affect blood pressure, but a full understanding of these relationships is still missing. For dyslipidemia, the relationship is also rather unclear, due to its strong association with obesity [108]. The clearest mode of action until now are effects of the microbiota on bile acid metabolism, which is critical for the absorption of lipids [109], but the observed associations are currently not linked to known mechanisms [110, 111].

Top down: How to investigate the microbiome

In contrast to macro-ecology, in microbial ecology it is possible to capture nearly the whole biodiversity of a habitat by sequencing its associated total DNA and/or specific phylogenetic marker genes. Different omics techniques can give the researcher information about species diversity and abundance, about their metabolic capabilities and associated symbiosis or pathogenicity factors. Technically there are different ways of obtaining this information but the ultimate goal of omics approaches is to answer the following set of questions: Who is there, what can they do, what are they actually doing?

While in macro-ecology specimen can normally be collected and studied in captivity, this is usually not the case for microbial ecosystems. It is assumed that we can only cultivate less than 1% of the bacterial diversity [12]. The rest, the so called "dark matter" cannot be readily captured by cultivation [112], although much progress has been made in recent years with high throughput culturing, the so called "culturomics" [113]. While bacteria make up most of the diversity of the human microbiota, archaea are also present in humans [114], as well as a high diversity of phages [115]. Fungi and protozoa also exist in this ecosystem, but are less well studied [116]. Why the majority of this biodiversity cannot be cultured is not clear, but different hypotheses exist. One of these hypotheses is that these organisms cannot survive on their own because of community dependencies. They are for instance microorganisms that live in a strict syntrophic relationship and are sharing nutrients and metabolites [117]. Syntrophic relationships might be due to excretion and uptake of common metabolites, but also more intricate cross-feeding networks have been reported to exist [70, 118, 119]. Other types of non-metabolic interactions also exist but are less easily quantifiable. Biofilms, which occur frequently in human associated microbiomes [120], are often not the product of a single species, but of a community [121]. They are not controlled by direct metabolic dependencies but by other mechanisms like quorum sensing [122].

Omics approaches towards understanding of the who and what of microbial communities

To answer the "who", the "what can they do", the "what are they actually doing" and "how do they respond to a diet or otherwise environmental change", different approaches can be used. To answer the "who", low cost amplicon sequencing of 16S ribosomal RNA (16S rRNA) encoding genes can be utilized. The 16S rRNA gene is present in all prokaryotes, and slowly mutating due to structural and catalytic constraints. Some of the secondary structure elements, called regions V for variable 1 to 9, are less constrained and therefore over time accumulate mutations more rapidly than other more conserved regions. Together, sequence variation within conserved and variable regions can be transformed into an evolutionary distance, allowing interference of the phylogeny of all members within a microbial community. As knowing the community composition in most studies is a prerequisite, next generation sequencing (NGS) of PCR amplicons targeting a selection of these variable regions is the most widely used approach. Despite the fact that no genomes are sequenced this is often falsely referred to as "metagenomics". This should be avoided and proper terminology should be used [123]. Nevertheless, making use of the currently available information from genomes and metagenomes, species identification in part also allows for predictions of functional capabilities [124, 125], albeit with inherent limitations with respect to their accuracy especially for understudied environments that are less well represented in currently available (meta)genome databases [126]. To more comprehensively answer the question "what can they do", metagenomics can be used. Metagenomics significantly increases both the amount and the complexity of the data. Besides the "who", and the "what can they do", community responses to diets or otherwise environmental changes can be studied by metatranscriptomics to answer the question "what are they doing". Sequencing the full transcriptome of the community provides by proxy insights in which pathways/processes are actually active. The logical progression of technology also leads to metaproteomics, which due to lack of precisely matching reference genomes [127] is still not very widely used and despite interesting results [128, 129] still remains to represent a niche discipline [130]. Meta-metabolomics (also called metabonomics [123], although this term has been used for a different purpose [131]), is currently an even less used technique.

A large body of research applying above-mentioned omics approaches is published in well-known journals. Figure 2A provides data up and until 2016. PubMed lists after the initial publications starting in the early 2000s an increasing amount of publications per year, reaching to more than a 1000 per year at the moment (Fig. 2B). The focus of most of these publications is on DNA-based approaches, including 16S rRNA gene sequencing and true metagenomics. This trend is followed distantly by metatranscriptomics, metaproteomics and meta-metabolomics. Since by far the majority of these publications are within the scope of some form of high-throughput nucleotide sequencing (16S rRNA gene, metagenomics, metatranscriptomics), in the following paragraphs we will focus on these omics approaches.

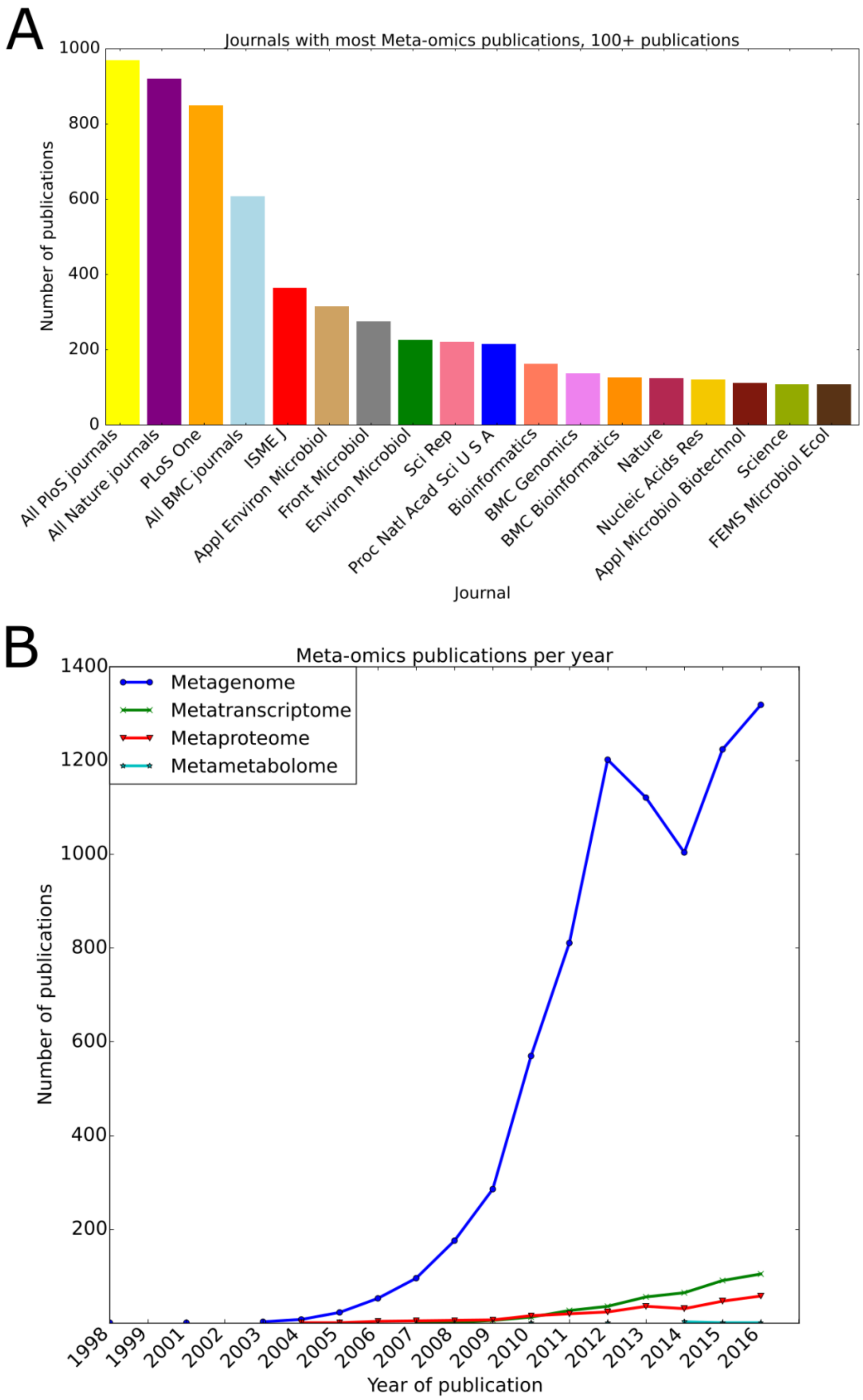


Figure 2: A) Journals with the most gut-related meta-omics publications. **b** Overview of gut-related omics publications per year. 16S rRNA gene sequencing and metagenomics are combined, since these cannot be easily distinguished via title/abstract searches due to the erroneous labelling of amplicon sequencing approaches as metagenomics by many researchers. All data was retrieved via PubMed searches for the corresponding terms. For the exact search terms, please see Additional file 1.

Differences within the omics technologies

The methods used for amplicon sequencing, metagenomics and metatranscriptomics are summarized under the term NGS technologies (also called 2nd generation technologies; for a review see [132]), including highly automated technologies represented by Illumina sequencing machines like HiSeq or MiSeq, the Roche 454, Ion Torrent and SOLiD technologies. These technologies are a follow up of Sanger sequencing, which still has the highest level of accuracy but has a rather low throughput due to limited parallelisation possibilities. NGS technologies allow millions of fragments to be sequenced in a single run. The DNA is randomly sheared, and all resulting fragments are sequenced with fluorescent nucleotides, which emit at incorporation in the new formed DNA strand certain light wavelengths. These can automatically be recorded by current systems and allow high throughput sequencing information by generating millions of short reads. One lane on a typical Illumina HiSeq machine can generate up to 360 Million reads, currently with lengths up to 350 bases. The limitation in this approach is mainly the used DNA polymerase for the extension of the newly formed DNA fragments, which tends to lose precision with increasing read length, making longer reads more error prone. Especially in metagenomics obtaining longer read lengths is important. Besides providing more information per single read, which is in general desirable in many cases, specifically for metagenomics it will i) lead to a higher chance of uniquely assigning reads to a single microbial taxon leading to a better resolution in strain and species separation, ii) make it easier to capture gene functionality, and iii) allow for a higher confidence during the assembly of the data, especially in those cases when the community harbours phylogenetically close species.

The new sequencing technologies (3rd generation sequencing) from Pacific Biosciences (PacBio) and Oxford Nanopore are ameliorating this problem. Both technologies can produce very long reads, up to 60000 bases (PacBio) and more (Nanopore). PacBio circumvents the loss of precision of the polymerase by repeatedly sequencing the same DNA fragment [133]. Oxford Nanopore channels single-stranded DNA through a pore which carries an electric current, and measures the change in current as the DNA passes by, with each of the bases causing a different change. This technology does not lose precision with increased length, but generating longer fragments and stably channelling them is the limitation [134]. Current drawbacks of both technologies as compared to the 2nd generation technologies are a higher error rate, requirement of a significantly larger amount of template DNA and higher sequencing costs. PacBio [135-140] and Oxford Nanopore [141] have already been used in microbiota sequencing and their use will most likely increase when the technologies further mature.

Extraction of information from 16S rRNA amplicon sequencing data

The 16S rRNA molecule shows a high degree of structural and sequence conservation in all prokaryotic organisms. Being part of the ribosome, it is a crucial part of the translation machinery. Because the specific secondary structure and function constraints evolutionary drift, it is, albeit with some limitations [142], possible to work with “universal” or species independent primers and therefore amplicon sequence analysis remains the standard approach to investigate microbial diversity. If two or multiple complete rRNA gene sequences have more than 97% identity, they belong to the same species. The 97% identity threshold is due to historical reasons because this value was found to be in agreement with DNA-DNA hybridization results, but otherwise no coherent species definition exists [143, 144]. In order to make clear that the actual species/genotype is often not known and might actually differ, 97% identity clusters of rRNA sequences are also referred to as "Operational Taxonomic Units" (OTU).

The 16S rRNA gene is approximately 1500 nucleotides in size and for the highest confidence the complete sequence is required. Due to the read length limitations of 2nd generation technologies researchers have therefore investigated, which sequence range of the rRNA showed the highest degree of variability and will therefore result in the best resolution [143, 145]. Using 2nd generation sequencing techniques, these regions (variable regions V1-V9) are therefore preferentially sequenced (for a review see [146]). Here, region-primer combinations need to be carefully matched as these choices can have a high impact on the results [147].

In eukaryotes, like e.g. fungi, the situation is more complicated. Sequencing 18S rRNA genes does not provide the required resolution, and often internal transcribed spacers (ITS) are sequenced instead [148].

After the amplicon sequencing data has been generated, the next step is to derive corresponding information regarding community composition. In general, since sequencing of single phylogenetic marker genes (fragments) requires less throughput than whole genomes, also the costs per sample are considerably lower, providing the necessary statistical power for a more detailed analysis [149].

Using 2nd generation sequencing techniques, there are multiple considerations involved, e.g. how similar the sequences are expected to be in the variable regions of choice, which reference database to use (SILVA [150], RDP [151] or Greengenes [152]), the significance of base-calling error rates intrinsic to high-throughput sequences data [153] and how erroneous sequences can be detected. Due to these challenges, sophisticated pipelines for taxonomic assignment have been developed, like e.g. Qiime [154], Mothur [155], Phyloseq [156], MICCA [157] and NG-Tax [158], the latter of which has been developed in our laboratories and provides computationally efficient and accurate taxonomic assignments and quantification of OTUs per sample with improved robustness against choice of region and other technical biases associated with 16S rRNA gene amplicon sequencing studies.

A range of different methods coming from macro-ecology is used to investigate a habitat's diversity. The species richness or mean species diversity of a sample is often referred to as alpha-diversity and the amount of variation in species composition among the samples (beta-diversity) can also be investigated. A range of different alpha-diversity measures is being used, including those that account for species richness (defined as the absolute count of individual populations per habitat), phylogenetically weighted richness (Faith's Phylogenetic Diversity [159]), and species diversity, including Shannon index [160] and Simpson index [161] (for a review see [162]). Diversity indices also try to incorporate the evenness of the species distribution [163], because different conclusions need to be drawn if an ecosystem is dominated by a single species with a plethora of other rare species, or if the distribution is rather even. Another important aspect is under-sampling. To estimate if the true richness of species has been captured, different methods like rarefaction analysis, Chao1 [164] or ACE [165] estimators can be used (review see [166]).

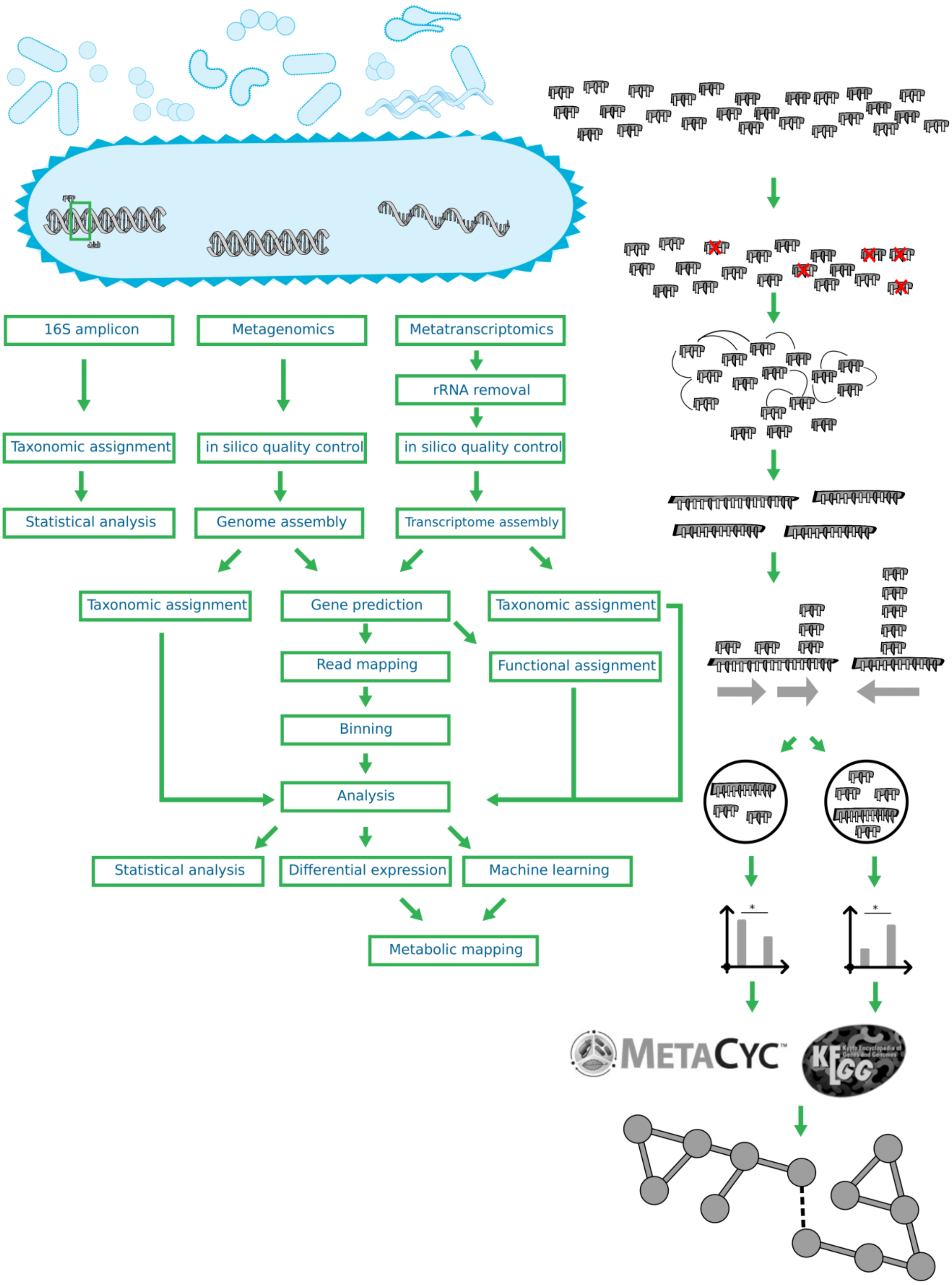
Analyses of beta-diversity make use of a number of different measures of pairwise community similarity, including e.g. Jaccard index [167], Bray Curtis dissimilarity [168] and UniFrac distance [169], the latter of which is phylogenetically weighted.

In most cases, a first look at the data is done with unconstrained multivariate statistical approaches such as Principle Component and Principle Coordinate Analysis (PC(o)A). These two methods try to fit highly dimensional data (e.g. a high amount of samples and different species in them) into a plot with two (or three) dimensions, trying to display as much of the variation in the data as possible. Factors that are potentially related to the observed variation, including e.g. environmental conditions, time points or the objective of the research, can be projected a posteriori, and their significance can be tested post-hoc.

Several of these statistical tools are standardly embedded in sequence analysis pipelines like Mothur [155], Qiime [154] or Phyloseq [156] and allow to capture measures of alpha- and beta-diversity. Choices can be made between default analysis routines and more customized procedures where users can adjust specific settings.

With these methods it has been found that e.g. the alpha-diversity in the microbiota of obese subjects is significantly reduced in contrast to the alpha-diversity in lean subjects [170]. Other successful studies in this field have already revealed that gut microbiota is transmitted vertically and that obese mice have a considerably less diverse microbiota than their lean counterparts [171]. Furthermore it has been shown that the gut microbiota changes during human development starting at birth and is different depending on geographic location [172], during long term dietary interventions [173] or when consuming specific diets even during a single day [174].

Figure 3: Overview of the different steps in the meta-omics analysis workflow. The different workflows are depicted, from left to right for 16s amplicon data, metagenomics data and metatranscriptomic data. The main steps for 16s amplicon data is the definition of OTUs together with taxonomic assignment, followed by statistical analysis. For metagenome data, first steps involve quality control steps, followed by a metagenome assembly. The workflow splits afterwards into two directions, one being the taxonomic assignment, the other one the definition of metagenomic bins and the functional annotation. Genes can be predicted from the genome assembly, which can be functionally profiled. With the coverage information of the genes, it is also possible to define genome bins. After this step is done, the same statistics as for 16s amplicon data can be performed, as well as differential expression/abundance analysis together with pattern detection through machine learning, and finally analysis of the metabolism. The workflow for metatranscriptomic data is in general the same, except that rRNA, which does not provide any information in this setting, needs to be removed before most of the steps, and that no binning is possible with transcriptome data


16S amplicon
Metagenomics
Metatranscriptomics
rRNA removal
Taxonomic assignment
in silico quality control
in silico quality control
Statistical analysis
Genome assembly
Transcriptome assembly
Taxonomic assignment
Gene prediction
Taxonomic assignment
Read mapping
Functional assignment
Binning
Analysis
Statistical analysis
Differential expression
Machine learning
Metabolic mapping
METACYC
KEGG

Extraction of functional information from metagenome data

In principle, full genomic information can be captured with metagenomics. Seminal projects in this area like MetaHit [27] and the human microbiome project [19] made great efforts to sequence the metagenomes of diverse cohorts with many subjects to investigate the full functional capacity of the different microbiomes. The amount of data required makes deeper sequencing necessary, which complicates the workflow to extract information from metagenomics data (Figure 3).

High throughput sequencing data is noisy, and quality control is a critical first step (review see [175]). One crucial step for which settings have not yet been universally agreed upon is the quality trimming [176], and no consensus advice can be given.

For simple read mapping there are a number of strategies that can be applied. BLAST [177] or Diamond [178] can be used to match reads directly to KEGG, to quantify the functions based on the number of matching reads (e.g. applied in [98]). A higher resolution is obtained when reads are mapped to a set of reference genomes [170, 179], which also allows for a taxonomic classification of observed functions [180]. If the phylogenetic distance between the reference set and the sample is small this has the advantage of speeding up the analysis. Furthermore, associated functional annotations can be directly utilized, making a separate annotation step unnecessary. A major drawback for this type of workflow is that only known species can be analysed, whereas new strains with novel functions, horizontal gene transfer and other evolutionary events will not be captured, and micro-diversity will be lost.

An alternative approach therefore is to assemble reads into larger contigs and extract genomes directly from metagenome data [181] (Figure 3). Today obtaining a high quality single genome can still be a challenge [21], and with a community genome assembly approach these challenges can multiply. Examples are chimeric assemblies between genomes due to presence of multiple strains of the same species (although miss-assemblies should not occur very often [182]), and a low coverage of low abundant species. At this point, it is also important to consider the mapping rate after the assembly. While we expect for a single organism that after the genome assembly most of the reads will map to the assembly, this can deviate for metagenomics. This is mainly due to the species richness and species evenness of the community under investigation. A complex species-rich sample of high evenness (i.e. similar abundance of many community members) will require more data to assemble the top-ranking species than a sample where a few high-ranking species have much higher abundances. Therefore species richness and species evenness need to be taken into account to evaluate if the mapping rate is appropriate for further analysis.

Some of these challenges have been tackled with specific metagenome assemblers like MetaVelvet [183], which take different properties of the sequencing data into account like e.g. the different abundances of the potentially present species. Currently a community derived assembly will also not lead to closed genomes. The next challenge is therefore to determine which of the assembled contigs/scaffolds belong to a single species. This process has been termed binning, and several tools such as MaxBin [184] or MetaCluster [185] have been developed to determine the amount of bins required and to assign contigs to bins. To do so these tools take different types of information into account, such as k-mers frequency in the data or contig read coverage. The quality control of this step is critical, since this process is also error prone, especially when phylogenetically close organisms of similar abundance occur in a community.

The most widely used method to test for correctness of binning is based on single copy marker genes, like in e.g. CheckM [186]. Based on the presence of these necessary genes, both the coverage of a genome in a bin as well as the amount of contamination from other genomes can be determined. A problem with this approach is that it is limited to contigs/scaffolds containing these core functions.

Next the taxonomic origin of the various bins can be determined (Figure 3). All programs and workflows which can perform this are reference based, but work with different mechanisms. One approach is to use BLAST [177] to compare all the metagenomic contigs against a database, like the NCBI NT database, or specialized databases like e.g. the human microbiome project [27]. The accuracy of the taxonomic assignments is proportional to the similarity score of the alignments. One of the first programs to deal with this problem is MEGAN [187], which also gives the user a graphical interface for direct analysis. The biggest drawbacks of this method are that i) it can be computationally prohibitive to use a large database, and ii) that closely related species cannot be differentiated from each other. A computationally more efficient alignment free method for the taxonomy determination is to compare the k-mer profiles of the metagenomics contigs with k-mer profiles obtained from a reference database. This has been implemented in tools like Kraken [188] or PhyloPythia [189] (for a review of programs see [190]),

To understand the underlying causes of a community change and potential effect, functional profiling needs to be performed (Figure 3). This part of the analysis is for a metagenome mainly different to a single genome in regards to the quantity, but the basic processes are the same. First gene prediction needs to be performed with gene callers like e.g. prodigal [191], which have special settings for this kind of data. A low level profiling can be obtained with a COG analysis [192]. The COG ontology consists of limited number of broad categories, which allow the detection of extensive changes. When more data is available a higher resolution can be obtained. These can be e.g. i) EC number prediction, which can be obtained via PRIAM [193] and can be linked to metabolic pathways using databases like KEGG or Metacyc [194], ii) lists of carbohydrate active enzymes [195] can e.g. be obtained via dbCAN [196], and iii) full domain profiles including GO terms [197] via e.g. InterproScan [198] or via second generation annotation tools [199]. With these so called full functional profiles, it is possible to reconstruct the metabolism of the bin [200-202], and e.g. bin-specific auxotrophies or special metabolic capabilities can be investigated. If someone wants to draw statistical conclusions for the difference in the metabolism by e.g. investigating for overrepresented functions (e.g. GO enrichment [203]), it should not be forgotten that, even for genomic information, replication is necessary [204]. If it is not possible to obtain all this data, due to lacking computational resources, also web services like IMG/M [205] or EBI metagenomics [206] can be used, which normally also have a user friendly interface, but only offer a limited depth of analysis.

Extraction of functional information from metatranscriptome data

The transcriptome approach will allow the investigator to focus on functions that are actually expressed in a given sample. A highly abundant species may show a low expression of functions of interest and vice versa (e.g. [207]). In fact, since DNA is also highly stable, the metagenomics approach might also take non-viable cell populations into account, which could falsify the conclusions, but also separate measures, like removal of non-viable cells, can be taken to prevent this [208]. Thus the metatranscriptome provides a more accurate account of actual functionality.

Most relevant steps, including QC, are the same as for single organism transcriptomics (for a review see [175], workflow see Figure 3). Not mentioned in [175], but necessary for metatranscriptome data is the *in silico* removal of spurious rRNA reads [209] as *in vitro* removal of rRNA prior to sequencing will most likely not remove all of it.

Like in metagenomics mRNA reads can either be mapped or *de novo* assembled. Mapping can be done if a set of reference genomes is available. If binning has been performed before, then the transcriptome should not be mapped to the different bins separately. If bins were separated before mapping, then the assignment of reads would be skewed if phylogenetically related bins are present (incorrect multiple assignment of reads). If no reference metagenome is available, it can be attempted to map the RNAseq data to related datasets. In this case again the absolute mapping rate of the data needs to be cautiously taken into account, because an unsuitable reference (due to large phylogenetic distance or missing species) will exhibit low mapping rate and will prevent a full analysis of the data. Alternatively, a de novo transcriptome assembly can be performed. Specific metatranscriptome assemblers have been developed to deal with the complexity of such data (for a review see [210]). Subsequent mapping of the same mRNA reads onto the de novo assembly allows for differential expression analysis, which can be performed with known tools like e.g. edgeR [211] or DESeq2 [212].

In many regards metatranscriptome analysis can function as a substitute for a metagenomics analysis while adding an additional layer of information. For instance, metatranscriptome analysis has already revealed that activity of carbohydrate degrading enzymes can be underestimated if only genomic information is considered, or how the activity of the gut microbiome responds to different diets [213, 214]. In principle similar conclusions could also be obtained from a combined metagenomics/metaproteomics approach [215] albeit at lower resolution.

A pure transcriptome assembly has the drawback that binning is not possible, since many of the binning approaches rely on the fact that in a metagenome all contigs from one species will exhibit similar coverage, which is not the case for a transcriptome. It will also not be possible to assemble very long contigs, because many intergenic regions will not be transcribed. Important changes at the ecosystem level can be assessed by analysing the expression levels of the microbiota in the community provided that species abundances are also taken into account; a 50% increase in abundance might appear as a 50% higher gene expression, but in this case does not reflect a transcriptional response on a per-microbe basis, but rather a compositional response at community level.

From information to understanding

As exemplified above many computational tools and pipelines exist that are able to extract biological information from high throughput data. Understanding the unique chemical and functional capabilities of the human microbiome and deciphering the biological roles of individual species is much more difficult. Linking microbial activities with gene expression and enzyme functionalities is just the first step. In early years of genomic research, "hairball" graphs had their appearance in many publications, showing connectivity within the available pile of data, rather than focusing on the biologically informative parts. With the increasing number of samples being analysed e.g. from patients, from replicates, from different conditions, different types of sequencing data combined with different types of computationally derived data such as EC number and domain predictions, which methods can be used to gain useful information?

The most obvious approach, especially with pure abundance data, is looking for correlations (also possible via regression [216]). It can be assumed that correlating species/OTUs have a symbiotic relationship with each other and/or with a third OTU, whereas anti-correlation can (but does not have to) indicate antagonistic behaviour. There are, however, several pitfalls. For example, OTUs, which are present only in very few samples, will be highly correlated due to the common absence in multiple samples. While this general conclusion can be true, it needs to be considered that absence in sequencing data does not have to mean absence of the organism. It can also indicate abundance below the detection threshold, or simply a failure in detecting the organism with the current pipelines.

The same methods described above for the analysis of 16S rRNA gene amplicon sequence data can also be utilized for metagenomics data. Multivariate visualization tools such as PCA can be used to see if specific sample groups, e.g. defined by specific interventions or states of health, cluster together, or if other factors are more prevalent in explaining the observed variation in the data. Nevertheless, for the in-depth analysis, more sophisticated methods should be used such as e.g. pattern recognition, which enables the researcher to find useful information in big data. This field is broadly classified into two approaches, i.e. supervised and unsupervised learning. In supervised learning, the researcher tries to classify unknown samples into categories for which already known samples exist. If, for example, samples from lean and obese subjects have been obtained, an algorithm can be trained to determine if samples of unknown origin were obtained from a lean or obese person. While supervised learning has been already used in microbiome research with great success, e.g. [29, 217] (for reviews of the methodologies see [218] [219]), and is currently researched for the application in many different fields and termed "life changing" for the general public (e.g. deep learning [220]), this approach is often hampered by the fact that samples from different studies are not comparable due to different methodological approaches with respect to e.g. DNA extraction or sequencing method and depth.

Unsupervised learning, also called clustering, does not rely on prior information. Clustering algorithms, including e.g. hierarchical clustering, k-means and dbscan, try to find unknown patterns in the given data, e.g. different patterns of gene expression over multiple conditions. This approach has also been used e.g. to determine the enterotypes [221], but also suffers from a wide array of challenges. The choice of clustering algorithm is not trivial and depends on the structure of the data, which can often not be determined in an easy way [222]. Furthermore these algorithms often rely on user-defined parameters such as the amount of clusters to find. Determining the best parameter set is its own research field, given that more than 30 different algorithms for this purpose exist [223], and not all are applicable to all clustering algorithms [222]. If at the end wrong parameters are chosen, it might lead to erroneous conclusions, like e.g. if not the optimal amount of clusters (in this case enterotypes [224]) is selected. Otherwise, a cluster might be split into multiple, or multiple distinct clusters might be treated as one.

Having said that, many of these algorithms have been implemented in different programs like ELKI [225] or WEKA [226], and can also be utilized by inexperienced users, although the final evaluation still often requires expert knowledge.

If useful patterns have been obtained after the machine learning, the last level is the biological understanding and interpretation. Simple approaches include just mapping extracted functional information such as EC numbers and KO numbers to pathway databases like KEGG [227]. More sophisticated solutions try to automatically extract the useful information from these networks, e.g. MetaModules [228]. If also other non-metabolic functions should be investigated, then a broader type of classification can be used. The most common analysis is the GO enrichment analysis, which aims to identify overrepresented functions in the dataset [203].

It also needs to be considered that the microbiome data does not have to stand on its own. If clinical or nutritional data is available, these can be used as well. Correlating such metadata with microbiome data has shown that e.g. factors like age or stool consistency are highly related to microbiome composition [229], as well as the hosts genetics [230]. Furthermore it is also possible to revert this, and use microbiome data together with clinical data to predict a persons' glycemic response to food intake [231].

Since this type of data can be highly connected, visualization of this connectivity might be necessary for a better understanding. While some visualization forms are standard, e.g. depicting the distribution of species/OTUs per sample in a bar chart, and metabolic networks as networks, sometimes more sophisticated methods are necessary. For analysis purposes, the Krona library [232] can be a useful visualization tool to explore quantitative hierarchical relationships between taxonomical groups. In many cases, there are no standard recipes for the analysis workflow, and custom solutions have to be developed. For these cases it is necessary to consider what type of data should be shown, and with which method they are obtained. Several visualization methods are available [233, 234], but standard packages for many of these are not necessarily developed yet or easily accessible.

Bottom up: Mechanistic insights into the microbiome

The next step after collecting data and investigating the communities is building models and testing hypotheses. While with single species this is very well doable, microbial communities pose more challenges to the researcher. For a single culturable species, it will be possible to collect the necessary data. It is possible to reconstruct the full metabolism (according to current knowledge), manually curate it, and measure a vast array of metabolites. In contrast, all these factors pose challenges in a community like the intestinal microbiota.

The sum is more than its parts

A community is more than an accumulation of multiple single organisms. The different microbes interact within a dynamic environment, they will behave differently, depending on who is in the surrounding, and what they are doing. Even for a single species, species abundance can lead to emergent properties such e.g. via quorum sensing, which can alter the behaviour of individual cells and the entire population dramatically [235]. In biofilms, for example, the formation itself is an emergent property, which would not be possible to observe if only single cells are considered. It also leads to the change in behaviour of the different cells, as some will get advantages in this environment (protection), whereas the cells on the surface are less protected, but also have more access to nutrients. Other forms of symbiotic relationships can also lead to emergent properties where e.g. some species in the community provide the means to overcome amino acid auxotrophies or vitamin deficiencies of others or of the host [236-238]. Another unrelated example from the oceanic microbiome is the detoxification the environment [239]. This case is commensalistic, since a big part of the microbial community benefits from the ability of one member to detoxify oxygen radicals, giving the other members a benefit, which lead in this case to genome streamlining by loss of genes related to oxidative stress. The authors even expanded their observation into the "Black Queen Hypothesis", stating that this streamlining together with a dependency on a helper organisms with leaky beneficial functions might be an universal concept. This is only possible to observe at the community level, and the investigation of a single species would not lead to such conclusions.

Numerous additional examples exist, also in the gut environment (for a more complete review see [240]).

How to predict the sum from its parts

How should the behaviour of such a community be predicted? The apparent approach is to model the metabolism of the whole community as a single entity or "supra-organism", neglecting species boundaries [241]. While this can give an idea about the metabolic capabilities, it is an oversimplification and will miss critical steps like metabolite exchanges and interdependencies between organisms. The extension of this approach would be to model single organisms, and connect these models to one community model.

Producing a good model of a single organism is the first step in this process. There exist high-throughput methods, like ModelSEED [242], Pathway Tools [243] or KBase [244], which can automatically construct a genome scale metabolic model (GSMM) from the given genomic information. Although such reconstructions can be of high quality, it is still likely that the model will contain errors or gaps, which need to be solved by manual curation [245].

If different models for the relevant organisms can be obtained, the next challenge is combining them. If the models are based on different databases/coming from different sources, then this could result in incongruences in the final model. While this should in general be avoided, it is sometimes necessary, because high quality models of different organisms exist (e.g. *Homo sapiens* [246], *Escherichia coli* [247]), and it is not feasible to integrate this work into the high-throughput frameworks. For such cases an integration of different model sources needs to be performed. The challenge is to match all the metabolites that need to be shared between all relevant models. Due to different problems, like the lack of unique identifiers, matching these names is not a trivial task, can be very error prone and requires the application of specialized tools (e.g. [248]).

Different hypotheses can be tested after a multi-organism model has been finally generated (e.g. [249, 250]). One of the first approaches should be to investigate ecological compatibility. This can be done e.g. via reverse ecology [251], by matching the metabolites in the different organisms to each other to see possible interconnections and metabolic dependencies. More advanced challenges are to actually simulate this metabolism. Finding the target, the objective function of a model, will depend on the underlying biology. Maximization of biomass is often used in single-organism models [252] (among others), and has also been used in multi-organism models (e.g. [249, 253]). This is not applicable in all cases, because e.g. competition or parasitic relationships can exist in an ecosystem and often the objective is not to maximize the biomass of the competitors in the surrounding. Therefore more sophisticated methods like D-OptCom [254] have been developed, which break the community optimization problem into multiple single problems. These consist of smaller optimization problems for each community member, and the main problem is to optimize the community. Others have extended this to even include spatial structures [255]. This allows the simulation of each bacterium's growth independently, giving a more realistic result than simulating community growth.

Metabolic models are not the only models which can be employed, metabolism is also not the only type of process which can be simulated, and the bacterial level is not the only scale which can be considered. Different kinds of kinetic models of the metabolism have been developed, some especially for the gut [256, 257], and also for related ecosystems [258], but this field is still in its infancy. The mentioned models also simulate metabolism, predicting the flow of carbohydrates into acids or extracellular polysaccharides, including different non-metabolic parameters like e.g. peristaltic movement of the gut. Also non-metabolic models exist, with the focus on antibiotic resistance in the gut [259] or the succession of organisms in the gut [260]. As it can be seen, the field is still far away from a comprehensive virtual gut model. In fact, already the whole cell model [261] is extremely complex, and contains e.g. different scales which might be lacking full integration into the model. With all the different factors to consider, integrating more data into the models with proper feedback systems, until up to the ecosystem level, will probably be a research objective for many years to come [16].

How to change the sum, and its parts

Modelling cannot be only done *in silico*. With synthetic biology, artificial model systems of the gut environment have been created [262]. These models vary in their complexity and capabilities to simulate the environment. It is important to differentiate which part of the gut is modelled, if there need to be multiple compartments, and if e.g. each of them needs to be pH controlled. These systems were shown to simulate parts of the gut appropriately [263], and e.g. [264] showed the contributions of intestinal movement to the development of inflammation in the gut.

But since these systems do not (yet) perfectly model the gut, final proof has often to be provided from animal models. Gnotobiotic animals [265] offer the possibility for controlled interventions. In contrast to the *in vitro* systems, the *in vivo* system will be able to incorporate all the necessary factors to evaluate gut functioning. Inoculation of the sterile animals with a defined microbiota ("synthetic ecology") allows studying the niches of specific bacteria [266, 267], the development of the microbiota over time [260], during development [268] and the interactions between different bacteria [117, 119, 269]. Gnotobiotic animal models have also been used, as mentioned earlier, to show that the microbiota does not only change with obesity, but that it also contributes to it [28, 97, 98, 270].

At the end, it still needs to be taken into account that animal models do not represent humans, and ways to influence our gut microbiota in a rational way are only partially understood. One of these rational methods is the gastric bypass. It is one of the last resorts for morbidly obese patients to lose weight, will have a significant effect on a subjects carbohydrate consumption and will alter the gut microbiota in different ways [271-274] (mainly an increase in Gammaproteobacteria), due to different changing factors like e.g. the distribution of bile acids. This is the most drastic method for a targeted microbiota change besides antibiotics and faecal transplantation. The latter has been used to treat severe diseases like *Clostridium difficile* infection (e.g. [31, 275]) or Ulcerative Colitis [276]. Faecal transplantation replaces a patient's gut microbiome with that of healthy donors, however, mechanisms underlying success or failure of the treatment have not yet been fully understood in all cases. The main factors do not only include the gut microbiota itself or the host genetics [230], but potentially also other factors like excreted metabolites [277, 278]. Due to the difficulties of understanding the mechanisms, it has not yet been possible to rationally design a medicine from this therapy, which would simplify the production and legal issues [32, 279], but progress is likely to be made within the coming years [240, 280].

Microbiome changes do not only have clinical impact. Pre-clinical applications are also possible. Nutritional methods can be rationally employed, without having dramatic impact on the everyday life and include mainly pre- and probiotics. The substances and microorganisms consumed are not new, and have been already consumed for millennia, e.g. as fermented milk products. But also their mode of action is not fully understood, and in some cases their usefulness is even debated [281]. Probiotics like *Lactobacillus* and *Bifidobacterium* (e.g. [282, 283]) might act in different ways. Tested hypotheses are that they might change the gut environment to make it inhospitable for pathogens [284, 285], produce antimicrobial compounds like SCFAs [286-288], alter the composition by releasing compounds from otherwise indigestible substrates (e.g. prebiotics) [282, 289] or reverse/prevent dietary effects [290, 291]. But even in such controlled setups it is too simple to attribute changes to single organisms, since the breakdown of e.g. prebiotics (leading to "postbiotics", which might be the actual bioactive compound) can involve multiple organisms (see e.g. the summary about quercetin in [292]).

Conclusions

The currently available body of research has shown that it is important to take the ecosystem as a whole into account to understand its health implications. Recently this trend is increasingly being picked up. After the first human genomes were sequenced, it was believed that it would change how medicine works. It was thought that every aspect of a human would be understood and that all treatments would be personalized [293, 294]. Although personal genome sequencing is still on the rise [295], this prediction has not turned out to be fully true [296], although it should be noted that there have also been significant successes (see e.g. table 1 in [297]) . While we for sure do not yet fully understand the human genome [298], we need to be aware now that it is not the only factor. The personal wellbeing is not only influenced by our genetic traits. Our complete ecosystem, the whole holobiont, needs to be taken into account. It is already clear that we cannot understand obesity if we do not understand our microbiome, and if we do not understand its connections to the host. With discoveries like the enterotypes [221] (caution for the results [224], as they have been discussed widely, with the notion that gradients are more likely than separate clusters), the next step after the personal genome might even be the personalized metagenome (and the first companies are even trying to market it). If people have different microbiomes, they might need to be treated differently to combat e.g. obesity. With enough data, and the understanding of its meaning, it might also be possible to prevent this lifestyle epidemic, in combination with personalized nutrition, as it is even already becoming potentially feasible [231]. We might also be able to go further, and even prevent diseases. The preventive measures are normally not part of the regular mainstream medicine, but ideas exist how incorporate preventive measures, pioneered as "4P medicine" (predictive, preventive, personalized, participatory) [299, 300]. If we know a person's microbiome, we will be able to predict if they are e.g. more prone to obesity or other risk factors (which is for some disease states already possible [29, 217]). If we understand the functionality, we will be able to take countermeasures with dietary interventions like pre- and probiotics. Since all these ecosystems are different, this approach will need to be personalized. Not only to take the personal genome and the personal microbiome into account, but also the compatibility with lifestyle, because even the best treatment might not suffice if a subject consumes by default a high fat "western diet" without any exercise. And this is all not possible, if the population does not participate. This approach will rely on everyone's personal data, which needs to be acquired. And it will only work, if the results are communicated clearly.

All of these points are future challenges. We do not yet fully understand the microbiome. With diet we are taking counter measures, but not always in rational ways. Medicine is already personalized, but not all treatments have the necessary data to be personalized. And while communication can already work (e.g. the whole "quantified self" movement is relying on achievements being communicated back), it is not always the case, and wrong communication, resulting in wrong expectations, will even discourage the users (e.g.

[301]). The researchers in the microbiome field need to be aware that this hype can also happen to the microbiome [302, 303].

Current microbiome research aims to overcome some of these challenges. Obesity research is likely to contribute in the close future to a better understanding of the underlying mechanisms, and the 4P medicine might partially become achievable in not too distant future, leading to better health and combating epidemics like obesity.

Acknowledgements

The authors want to thank Ruben van Heck, Maria Suarez-Diez (Wageningen University and Research, Laboratory of Systems and Synthetic Biology), Joan Edwards and Gerben Hermes (Wageningen University and Research, Laboratory of Microbiology) for helpful discussions.

Funding

B. Hornung is supported by Wageningen University and the Wageningen Institute for Environment and Climate Research (WIMEK) through the IP/OP program Systems Biology (project KB-17-003.02-023).

Supplementary information

Supplementary information can be found online at:

https://doi.org/10.1186/s12263-018-0594-6

Chapter 3: Genomic and functional analysis of *Romboutsia ilealis* CRIB^T reveals adaptation to the small intestine.

This chapter is adapted from:

Jacoline Gerritsen*, Bastian Hornung*, Bernadette Renckens, Sacha A.F.T. van Hijum, Vitor A.P. Martins dos Santos, Ger T. Rijkers, Peter J. Schaap, Willem M. de Vos, Hauke Smidt, 2017. "Genomic and functional analysis of *Romboutsia ilealis* CRIB^T reveals adaptation to the small intestine". In PeerJ. https://doi.org/10.7717/peerj.3698

*Contributed equally

Abstract

Background. The microbiota in the small intestine relies on their capacity to rapidly import and ferment available carbohydrates to survive in a complex and highly competitive ecosystem. Understanding how these communities function requires elucidating the role of its key players, the interactions among them and with their environment/host.

Methods. The genome of the gut bacterium *Romboutsia ilealis* CRIB^T was sequenced with multiple technologies (Illumina paired-end, mate-pair and PacBio). The transcriptome was sequenced (Illumina HiSeq) after growth on three different carbohydrate sources, and short chain fatty acids were measured via HPLC.

Results. We present the complete genome of *Romboutsia ilealis* CRIB^T, a natural inhabitant and key player of the small intestine of rats. *R. ilealis* CRIB^T possesses a circular chromosome of 2,581,778 bp and a plasmid of 6,145 bp, carrying 2,351 and eight predicted protein coding sequences, respectively. Analysis of the genome revealed limited capacity to synthesize amino acids and vitamins, whereas multiple and partially redundant pathways for the utilization of different relatively simple carbohydrates are present. Transcriptome analysis allowed identification of the key components in the degradation of glucose, L-fucose and fructo-oligosaccharides.

Discussion. This revealed that *R. ilealis* CRIB^T is adapted to a nutrient-rich environment where carbohydrates, amino acids and vitamins are abundantly available.

Introduction

Intestinal microbes live in a complex and dynamic ecosystem, and to survive in this highly competitive environment, they have developed close (symbiotic) associations with a diverse array of other intestinal microbes and with their host. This has led to a complex network of host-microbe and microbe-microbe interactions in which the intestinal microbes and the host co-metabolise many substrates [304, 305]. In addition to competition for readily available carbohydrates in the diet, intestinal microbes are able to extract energy from dietary polysaccharides that are indigestible by the host [71]. Furthermore, intestinal microbes can utilize host-derived secretions (e.g. mucus) as substrates for metabolic processes [306]. In turn, the metabolic activities of the intestinal microbes result in the production of a wide array of compounds, of which some are important nutrients for the host. For example, short chain fatty acids (SCFA), the main end-products of bacterial fermentation in the gut, can be readily absorbed by the host and further metabolized as energy sources [307, 308]. All together, the metabolic activity of the intestinal microbiota has a major impact on the health of the host, and recent studies have indicated an important role for microbial activity in diseases such as inflammatory bowel disease, irritable bowel syndrome and obesity [309, 310].

We only have a limited understanding of the heterogeneity in microbial community composition and activity in different niches along the length of the intestinal tract. To unravel the functional contribution of specific intestinal microbes to host physiology and pathology, we have to understand their metabolic capabilities at a higher resolution. It is still difficult, however, to associate a functionality in this ecosystem to specific sets of genes and in turn to individual microbial species, and vice versa. To this end, the combination of genome mining and functional analyses with single microbes or with simple and defined communities can provide an overall insight in the genetic and functional potential of specific members of the intestinal microbial community [95, 249, 311].

As mentioned above, intestinal microbes have adapted or even specialized in foraging certain niche-specific substrates. However, little is known about the adaption of intestinal microbes to the conditions in the small intestine [312-314]. Community composition and activity in the small intestine is largely determined by the host digestive fluids such as gastric acid, bile and pancreatic secretions. The small intestine is a nutrient-rich environment, and previous studies have shown that the microbial communities in the (human) small intestine are driven by the rapid uptake and conversion of simple carbohydrates [70, 179]. Genomic studies of small intestinal isolates have indicated environment-specific adaptations to the small intestine with respect to their carbohydrate utilization capacities, which was evidenced by the presence of a wide array of genes involved in nutrient transport and metabolism of, mainly simple, carbohydrates [315].

Here we describe a model driven genomic analysis of the small intestinal inhabitant *Romboutsia ilealis* CRIB^T [316]. *R. ilealis* CRIB^T is currently still the only isolate of the recently descibed species *R. ilealis*, a species that belongs to the family *Peptostreptococcaceae,* of which many members are common intestinal microbes including the well-known species *Clostridioides difficile* (previously known as *Clostridium difficile*) and *Intestinibacter bartlettii* (previously known as *Clostridum bartlettii*) [317]. An overview of the metabolic capabilities and nutritional potential of the type strain of *R.*

ilealis CRIB^T is provided here to identify potential mechanisms that enable this organism to survive in the competitive small intestinal environment.

Materials and methods

Genome sequencing, assembly and annotation

R. ilealis CRIB^T (DSM 25109) was routinely cultured in CRIB medium at 37 °C as previously described [316]. Genomic DNA extraction was performed as previously described [315]. Genome sequencing was done using 454 Titanium pyrosequencing technology (Roche 454 GS FLX), as well as Illumina (Genome Analyzer II and HiSeq2000) and PacBio sequencing (PacBio RS). Mate-pair data was generated by BaseClear (Leiden, the Netherlands). All other data was generated by GATC Biotech (Konstanz, Germany). The genome was assembled in a hybrid approach with multiple assemblers. In short, after estimation of the genome size, assembly of the genome was performed with two different assemblers in parallel using the different sequence datasets. After merging the two assemblies three rounds of scaffolding were performed, once with paired-end data and twice with mate-pair data. Gap-filling was performed after each scaffolding step.

Genome annotation was carried out with an in-house pipeline. Prodigal v2.5 was used for prediction of protein coding DNA sequences (CDS) [191], InterProScan 5RC7 for protein annotation [198], tRNAscan-SE v1.3.1 for prediction of tRNAs [318] and RNAmmer v1.2 for the prediction of rRNAs [319]. Additional protein function predictions were derived via BLAST identifications against the UniRef50 [320] and Swissprot [321] databases (download August 2013). Afterwards the annotation was further enhanced by adding EC numbers via PRIAM version March 06, 2013 [193]. Non-coding RNAs were identified using rfam_scan.pl v1.04, on release 11.0 of the RFAM database [322]. CRISPRs were annotated using CRISPR Recognition Tool v1.1 [323].

Qualitative metabolic modelling has been performed with Pathway tools v18.0 [324]. A generic default medium consisting out of ammonia/urea, sulfite, hydrogen sulfide and phosphate was assumed, and the qualitative possibility to produce all necessary biomass metabolites was tested with the supply of different carbohydrates, which had been tested before *in vitro*.

See the Supplemental Methods in Text S1 for details on the genomic DNA extraction, genome sequencing, assembly, annotation, and metabolic modelling.

Whole-genome transcriptome analysis

R. ilealis CRIB^T was grown in a basal bicarbonate-buffered medium [325] supplemented with 16 g/L yeast extract (BD, Breda, The Netherlands) and an amino acids solution as used for the growth of *C. difficile* [326]. In addition, the medium was supplemented with either 0.5 % (w/v) D-glucose (Fisher Scientific Inc., Waltham, MA USA), L-fucose (Sigma-Aldrich, St. Louis, MO, USA) or fructo-oligosaccharide (FOS) P06 (DP 2-4; Winclove Probiotics, Amsterdam, The Netherlands). The final pH of the medium was adjusted to 7.0. For each condition, triplicate cultures were set up. For RNA-seq analysis, the cells were harvested in mid-exponential phase (OD_{600nm} = 0.25-0.55, ~8-10h incubation) (Table S1).

Total RNA was purified using the RNeasy Mini Kit (QIAGEN GmbH, Hilden, Germany). Depletion of rRNA was performed using the Rib-Zero^{TM} Kit for bacteria (Epicentre Biotechnologies, Madison, WI, USA). The ScriptSeq^{TM} v2 RNA-seq Library Preparation Kit in combination with ScriptSeq^{TM} Index PCR primers (Epicentre Biotechnologies) was used for library construction for whole-transcriptome sequencing (RNA-seq). The barcoded cDNA libraries were pooled and sent to GATC Biotech (Konstanz, Germany) where 150 bp sequencing was performed on one single lane using the Illumina HiSeq2500 platform in combination with the TruSeq Rapid SBS (200 cycles) and TruSeq Rapid SR Cluster Kits (Illumina Inc., San Diego, CA, USA). Reads were mapped to the genome with Bowtie2 v2.0.6 [327] using default settings, after quality control (rRNA removal, adapter trimming, and quality trimming) had been performed. Details on the RNA-seq raw data analysis can be found in Table S2 and Supplemental Methods in Text S1.

Gene expression abundance estimates and differential expression analysis was performed using Cuffdiff v2.1.1 [328] with default settings. Differentially expressed genes were determined by pairwise comparison of a given condition to the other three conditions for a total of six pairwise comparisons. Genes were considered significantly differentially expressed when they showed a ≥1.5 log2(fold change) in any of the conditions with a false discovery rate (FDR)-corrected P value (q value) ≤ 0.05 (Tables S3-S6). Principal component analysis was performed with Canoco 5.0 [329] on log-transformed gene transcript abundances using Hellinger standardization. Gene expression heatmaps were generated based on gene transcript abundances using R v3.1.0 and R-packages svDialogs and gplots.

See the Supplemental Methods in Text S1 for details on growth on different carbohydrate media and whole genome transcriptome analysis.

Metagenomic investigations

The datasets PRJNA237362 [330] and PRJNA298762 [331] were analysed as relevant representative publicly available 16s rRNA gene amplicon datasets for the presence of 16S rRNA gene sequences closely related to that of *R. ilealis* with NG-Tax version 0.3 [158] with the –classifyRatio argument set to 0.9.

Nucleotide sequence accession number

All related data have been deposited in the European Nucleotide Archive. The raw reads for the genome of *R. ilealis* CRIB^{T} can be accessed via the accession numbers ERR366773, ERX397233, ERX397242 and ERX339449. The assembly can be accessed under LN555523-LN555524. The RNAseq data have been deposited under the numbers ERS533849- ERS533861.

Results

Genome analysis

Global genome features

R. ilealis CRIB^T contains a single, circular chromosome of 2,581,778 bp and a plasmid of 6,145 bp (Table 1 and Fig. 1). The chromosome contains 2,351 predicted protein CDS, of which 321 were annotated as hypothetical and for 91, only a domain of unknown function could be assigned. The plasmid carries eight predicted protein CDS, of which none was recognized for having a metabolic or replicative function. Furthermore, it appears to be a non-mobilizable plasmid, given that it lacks any known mobilization-associated genes. The overall G+C content of the genome is 27.9 %, which is in good agreement with a G+C content of 28.1 mol% previously determined for *R. ilealis* CRIB^T by HPLC methods [316].

Table 1: General features of the *Romboutsia ilealis* CRIB^T genome

	Chromosome	Plasmid
Size (bp)	2,581,778	6,145
G+C content (%)	27.9	29.3
Protein CDS	2,351	8
Pseudogenes	12	0
Coding density	1.10	1.02
Average gene size (bp)	899	531
rRNA genes		
16S rRNA genes	14	0
23S rRNA genes	14*	0
5S rRNA genes	14	0
tRNAs	109	0
ncRNAs	28	0
CRISPR repeats	1*71	0

* An additional 23S rRNA gene is expected in one of the gaps.

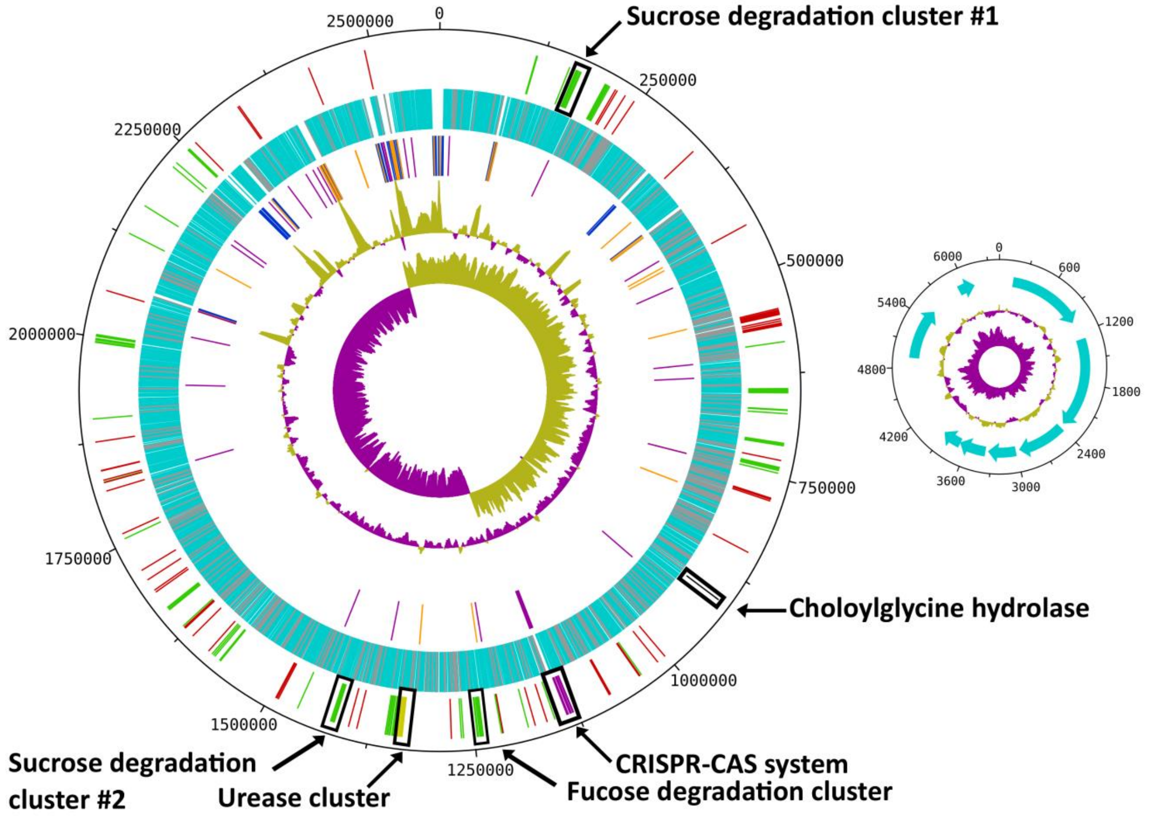


Figure 1: Circular map of the *R. ilealis* CRIB^T genome.
Both chromosome and non-mobilizable plasmid are shown. For the chromosome tracks from inside to outside are as follows: 1, GC skew; 2, G + C content; 3, RNAs [rRNAs (blue), tRNAs (orange) and ncRNAs (purple)]; 4, all predicted protein CDS [with predicted function (light-blue), hypothetical proteins and proteins to which only a domain of unknown function could be assigned (grey)]; 5, genes or gene clusters of interest [(mobile genetic elements (red), Cas proteins (pink), urease gene cluster (yellow), choloylglycine hydrolase (black), gene clusters involved in carbohydrate utilization (green)]. For the plasmid tracks from inside to outside are as follows: 1, GC skew; 2, G + C content; 3, all predicted CDS.

With a total of 14 copies of the 16S ribosomal RNA (rRNA) gene, *R. ilealis* CRIB^T is among the species with the highest number of 16S rRNA gene copies reported up to this date [332]. High numbers of rRNA operons have been proposed to be indicative for fast growth and to allow microbes to respond quickly to changes in available resources [333]. In addition, a high copy number of the rRNA operon has been suggested to be essential for successful sporulation and germination [334]. This is also reflected in the observation that in general the species that contain the highest number of reported rRNA operons, including *R. ilealis* CRIB^T, belong to the spore-forming bacterial orders *Bacillales* and *Clostridiales*. Not all of the 16S rRNA gene copies in *R. ilealis* CRIB^T are embedded in the conserved 16S-23S-5S rRNA operon structure. Of the fifteen locations containing rRNA genes, ten are in the classical order 16S-23S-5S. The other five operons are characterized by duplicated or missing rRNA genes, or a different order of the genes. It should be noted that the current assembly contains three gaps, all of which are located within rRNA operons. Diverging rRNA operon structures have been reported for other genomes containing multiple rRNA operons, as a result of duplications [335, 336].

A cluster of orthologous genes (COG) category [337] could be assigned to 1,647 of the predicted proteins (70 %) including 372 proteins (16 %) assigned to the categories R (general function prediction only) and S (function unknown) (Fig. S1). With InterProScan a predicted function could be assigned to 82 % of the predicted proteins. Based on the InterPro and PRIAM classifications [193], an enzymatic function could be predicted for more than 500 proteins.

General metabolic pathways

Analysis of the CDS predicted from the *R. ilealis* CRIB^T genome revealed the presence of a complete set of enzymes for the glycolytic pathway. In line with the anaerobic lifestyle of the organism, enzymes for the oxidative phase of the pentose phosphate pathway could not be detected. Additionally, the genes that encode enzymes involved in the tricarboxylic acid cycle were lacking. Subsequently a metabolic model was constructed with Pathway tools v18.0. A flux balance analysis with the model was performed, suggesting that *R. ilealis* CRIB^T is a mixed acid fermenter as previously reported [316]. Predicted end products of fermentation are a mixture of acetate, formate, lactate and ethanol, with the possibility of gas formation (CO_2 and H_2). In addition to ethanol, which can be produced during mixed acid fermentation, 1,2-propanediol was predicted to be formed via the L-fucose degradation pathway. The fermentation end products formate, acetate and lactate are predicted to be produced from pyruvate. No other solvents were predicted to be produced by the metabolic model. The only metabolite produced by *R. ilealis* CRIB^T that was not accounted for by the metabolic model, was propionate. None of the three established pathways for propionate production in the intestinal tract, i.e. the succinate, acrylate or the propanediol pathway [338], could be identified at the genetic level in the genome of *R. ilealis* CRIB^T. Although propionate is only produced in low amounts (max. 3 mM in 24 h) it is noteworthy because propionate production was observed repeatedly during *in vitro* growth in this study (see Table 2) and as previously reported [316].

The analysis of the genome and the prediction by the model indicated that fermentation is probably the main process for energy conservation in *R. ilealis.* However, the presence of a sulfite reductase gene cluster (CRIB_1284-CRIB_1286) of the dissimilatory *asrC*-type [339] points at possible anaerobic respiration. Similar siroheme-dependent sulfite reductases are found in many close-relatives of *R. ilealis* such as *I. bartlettii*, *Clostridium sordellii* and *C. difficile* [340]. Sulfite reduction by *R. ilealis* CRIB^T, and close relatives, has been previously demonstrated *in vitro* [316], and increased growth yield and metabolite production was observed in the presence of sulfite for *R. ilealis* CRIB^T (Table S7). In the intestinal tract, sulfite is derived from food sources that contain sulfite as a preservative, and it has been shown that neutrophils release sulfite as a part of the host defence against microbes [341].

Metabolism of growth factors and cofactors

Complete pathways are present for the biosynthesis of the amino acids aspartate, asparagine, glutamate, glutamine and cysteine, using carbon skeletons available from central metabolites or via conversion of other amino acids. However, many genes encoding enzymes required for biosynthesis of other amino acids appeared to be absent in *R. ilealis* CRIB^{T}. As most missing genes are part of well-studied pathways, it is unlikely these functionalities are encoded by unknown genes and likely represent true auxotrophies. The absence of genes to produce branched-chain amino acids (leucine, isoleucine and valine) was also reflected in the absence of branched chain fatty acids in the cell membrane of *R. ilealis*, which is characteristic for the genus *Romboutsia* [316]. From these observations it can be concluded that *R. ilealis* depends on a number of exogenous amino acids, peptides and/or proteins to fuel protein synthesis. The dependency on an exogenous source of amino acids is reflected by the identification of multiple amino acid transporters, including an arginine/ornithine antiporter, multiple serine/threonine exchangers, a transporter for branched amino acids, and several amino acid symporters and permeases without a predicted specificity. Furthermore, numerous genes were annotated as protease or peptidase, including several with a signal peptide.

R. ilealis CRIB^{T} appears to contain all genes for *de novo* purine and pyrimidine synthesis, as well as for the production of the coenzymes NAD and FAD via salvage pathways from niacin and riboflavin, respectively. While some organic cofactors can be produced by *R. ilealis* CRIB^{T}, it mainly relies on salvage pathways (e.g. for lipoic acid) or exogenous sources for the supply of precursors, mainly in the form of vitamins (e.g. thiamin, riboflavin, niacin, pantothenate, pyridoxine, biotin, vitamin B12).

Carbohydrate transport and metabolism

As previously reported, *R. ilealis* CRIB^{T} is able to utilize a wide variety of carbohydrates [316]. Previously, good growth of *R. ilealis* on L-fucose, glucose, raffinose and sucrose was described, in addition to moderate growth on D-arabinose and D-galactose and weak growth on D-fructose, inulin, lactose, maltose and melibiose. Growth on L-fucose, fructose, galactose, glucose, lactose, maltose, melibiose, raffinose and sucrose was predicted from the genome-scale metabolic model as well. For these different carbohydrates, the genes encoding the specific carbohydrate degradation enzymes were found distributed throughout the genome in gene clusters together with their respective transporters and transcriptional regulator. The only carbohydrate utilized by *R. ilealis* CRIB^{T} that was not predicted based on the metabolic model, was D-arabinose. Although a separate arabinose transporter, similar to the maltose and sucrose transporters, could be identified in the genome *R. ilealis* CRIB^{T}, no separate pathway for the use of D-arabinose could be predicted, However, it is likely that the L-fucose degradation pathway (encoded by genes CRIB_1294-CRIB_1298) is also used for D-arabinose utilization as is also observed in other intestinal species [342]. In addition to the carbohydrates for which growth was studied, a gene cluster involved in the degradation of the host-derived carbohydrate sialic acid could be predicted (CRIB_613-CRIB_619) [343]. The structure of this gene cluster is similar to the one identified in *C. difficile* [344]. The ability to degrade the predominantly host-derived carbohydrates, L-fucose and sialic acid, suggest a role in the utilization of mucin, an abundant host-derived glycoprotein in the intestinal tract [306, 345]. However, no growth on mucin was observed (Table S7), which is in line with the lack of a predicted extracellular fucosidase and/or sialidase.

Other genes encoding niche-specific functionalities

A gene cluster encoding a urease, consisting of three subunits (*ureABC*), and a number of urease accessory genes was identified (CRIB_1381-CRIB_1388). The gene cluster identified in *R. ilealis* CRIB^T is very similar to the urease gene cluster in the genome of *C. sordellii* (Fig. S3), a species in which the urease activity is used to phenotypically distinguish *C. sordellii* strains from *C. bifermentans* strains [346]. Furthermore, a possible ammonium transporter (CRIB_1389) was identified in the genome of *R. ilealis* CRIB^T next to the urease gene cluster. Ureases are nickel-containing metalloenzymes that catalyse the hydrolysis of urea to ammonia and carbon dioxide, and thereby these enzymes allow microbes to use urea as nitrogen source by assimilation via glutamate. They are ubiquitous proteins occurring in diverse organisms [347]. In the intestinal environment, where urea is abundantly present [348](Fuller & Reeds 1998), some bacteria use ureases to survive the acidic conditions in the upper part of the intestinal tract as urea hydrolysis leads to a local increase in pH [349].

Another gene encoding a niche-specific functionality is the predicted choloylglycine hydrolase. Proteins within the choloylglycine hydrolase family are bile salt hydrolases (BSHs), also known as conjugated bile acid hydrolases (CBAHs), that are widespread among intestinal microbes [350]. They are involved in the hydrolysis of the amide linkage in conjugated bile salts, releasing primary bile acids. There is a large heterogeneity among BSHs, for example with respect to their substrate specificity. The BSH of *R. ilealis* CRIB^T was found to be the most similar to the one found in *Clostridium butyricum*. Although the physiological advantages of BSHs for the microbes are not completely understood, it has been hypothesized that they constitute a mechanism to detoxify bile salts and thereby enhance bacterial colonization [340].

Metabolite and transcriptome analysis

Metabolite and transcriptome analysis of R. ilealis CRIB^T during growth on different carbohydrates

To study key pathways predicted to be involved in carbohydrate utilization and their regulation in more detail, a genome-wide transcriptome analysis was performed, focussing on four experimental conditions. Firstly, growth on glucose, a preferred substrate for many microbes present in the intestinal tract, was studied. Secondly the growth on fructans, oligo-and polysaccharides present in many food items was examined. Previously weak growth on inulin, a polysaccharide consisting of long chains of ß1→2 linked fructose units, was observed [316]. For this study a shorter fructan (FOS P06, DP2-4) was chosen, because growth on shorter fructans is likely more relevant for microbes living in the small intestine [70]. Thirdly, growth on L-fucose was examined, as growth on this substrate was found to be unique for *R. ilealis* CRIB^T compared to other related microbes. Finally, *R. ilealis* CRIB^T was also grown in the basal medium in the absence of an additional carbon source for comparison (control condition).

Based on measurements of optical density and pH during growth (growth characteristics of individual cultures can be found in Table S1), samples were drawn in the mid-exponential phase (~8-10 h incubation; used for transcriptome analysis) and in stationary phase (24 h incubation), and sugar utilization and fermentation products were measured with HPLC (Table 2). In neither of the experimental conditions the supplied carbohydrates were depleted, and metabolites were still produced at the time of sampling at ~8-10 h and 24 h, which further confirmed that samples obtained for transcriptome analysis at ~8-10 h were taken during exponential growth. In the FOS cultures, an accumulation of extra-cellular fructose was observed. As predicted from the metabolic model, growth on glucose resulted in the production of formate, acetate and lactate (Table 2).

Growth on FOS was marginally lower than that on glucose, however, after 24 h of growth, the same fermentation products were observed in similar amounts (Table 2). Growth on L-fucose showed production of 1,2-propanediol instead of lactate. The fact that 1,2-propanediol was observed in one of the control cultures could be explained by the fact that an L-fucose grown culture was used as inoculum for this culture, leading to carry-over of minor amounts of metabolites.

For the genome-wide transcriptome analysis of triplicate cultures grown in the four different conditions (i.e. a total of 12 cultures), a total of 159,250,634 150bp-reads were generated by RNA-seq (overview in Table S2). Principal component analysis of the transcriptomes of the individual cultures showed that the cultures clustered by condition (Fig. 2).

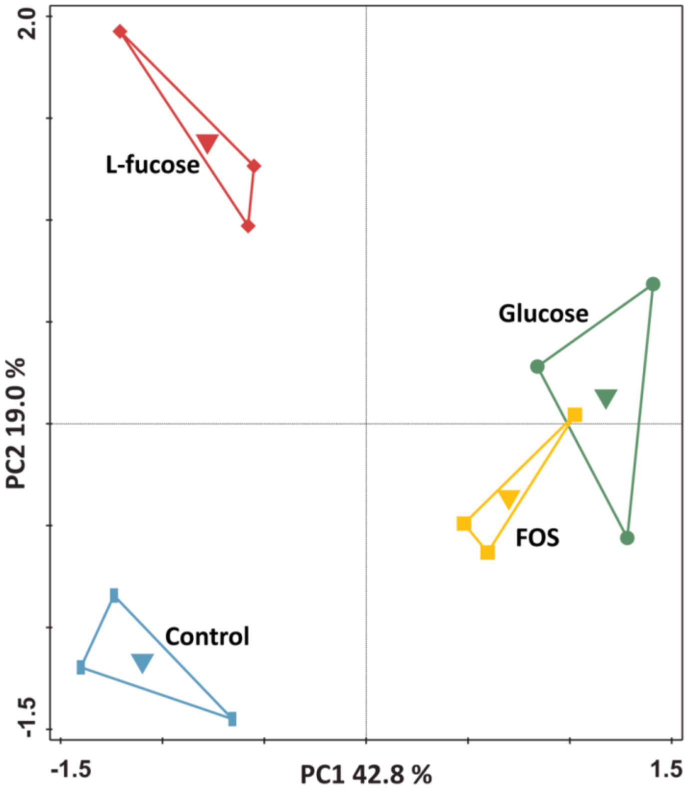


Figure 2: Principal component analysis of the transcriptomes of *R. ilealis* CRIB^T grown on different carbohydrates (glucose, FOS and L-fucose) or in the absence of an additional carbon source (control).
First and second ordination axes are plotted, explaining 42.8% and 19.0% of the variability in the data set, respectively. Individual transcriptomes are symbol-coded by experimental condition: glucose (circles), FOS (squares), L-fucose (diamonds) and control (rectangles). The experimental conditions were used as supplementary variables as well and could explain 62.9% of the variation.

Table 2: Fermentation end products of *R. ilealis* CRIB^{T} produced during growth on different carbohydrates (glucose, FOS or L-fucose) or in basal medium in the absence of a carbon source (control condition).
Samples were obtained during mid-exponential phase (~8–10 h incubation; used for transcriptome analysis) and in stationary phase (24 h incubation). For the control cultures, fermentation products are shown for the individual cultures separating the carbohydrates used for preconditioning of the inoculum. For the three other conditions, values represent means of triplicate cultures with standard deviations. N.D.: Not detected

	Formate (mM)		**Acetate (mM)**		**Propionate (mM)**		**Lactate (mM)**		**1,2-propane-diol (mM)**	
	8-10 h	**24 h**	**8-10 h**	**24 h**	**8-10 h**	**24 h**	**8-10 h**	**24 h**	**8-10 h**	**24 h**
Control:										
basal medium	3.2	7.7	2.0	6.2	2.0	2.2	N.D.	N.D.	N.D.	N.D.
(glucose inoc.)	4.5	9.2	2.4	7.4	2.4	2.9	N.D.	N.D.	N.D.	N.D.
(FOS inoc.)	4.8	10.8	2.3	9.8	2.3	3.0	N.D.	N.D.	1.0	1.0
(L-fucose inoc)										
Basal medium + glucose (5 % w/v)	4.4±1.2	28.2±4.3	1.0±0.9	16.3±2.2	1.0±0.9	1.3±0.1	N.D.	3.0±0.7	N.D.	N.D.
Basal medium + FOS (5 % w/v)	4.7±0.6	27.3±2.5	1.4±0.0	17.7±1.4	1.4±0.0	1.6±0.1	N.D	2.5±0.3	N.D.	N.D.
Basal medium + L-fucose (5 % w/v)	6.7±0.1	19.5±3.6	2.8±0.1	16.3±2.9	2.8±0.1	2.8±0.4	N.D.	N.D.	1.3±0.1	7.7±1.4

Differential expression of genes involved in carbohydrate degradation and fermentation in *R. ilealis* CRIB^{T}

To identify differentially regulated genes, pairwise comparisons were done with cuffdiff [328] using a cut off of ≥1.5 log2 (fold-change) and q-value ≤0.05. Figure 3 shows a heat map of all differentially regulated genes, and exact numbers can be found in Tables S3-6.

The gene cluster involved in glycolysis (CRIB_186-CRIB_191) was most abundantly expressed in the conditions that support the highest growth rates determined by the highest cell density reached in the time period that was measured (glucose, followed by FOS; Fig. 3). This was also reflected in the fact that expression of genes encoding proteins involved in replication such as ribosomal proteins, proteins involved in cell wall biosynthesis and general cell division processes were most strongly expressed during growth in the presence of glucose and to a lesser extent FOS. Other genes involved in the central sugar metabolic pathways (e.g. CRIB_1849, CRIB_140, CRIB_2223, and CRIB_105) were upregulated in these conditions, albeit not significantly differentially regulated. This suggests that these genes are less tightly regulated at the transcriptional level, probably because they are also involved in other processes than sugar degradation [351]. The metabolic model suggests that this is indeed the case, as some of the enzymes produce intermediates which can be consumed by fatty acid biosynthesis and amino acid biosynthesis processes.

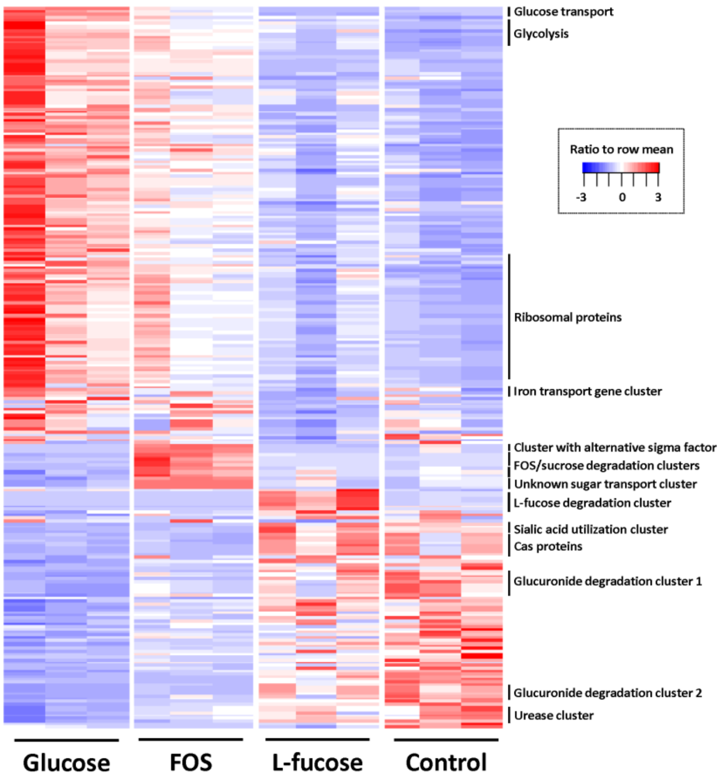
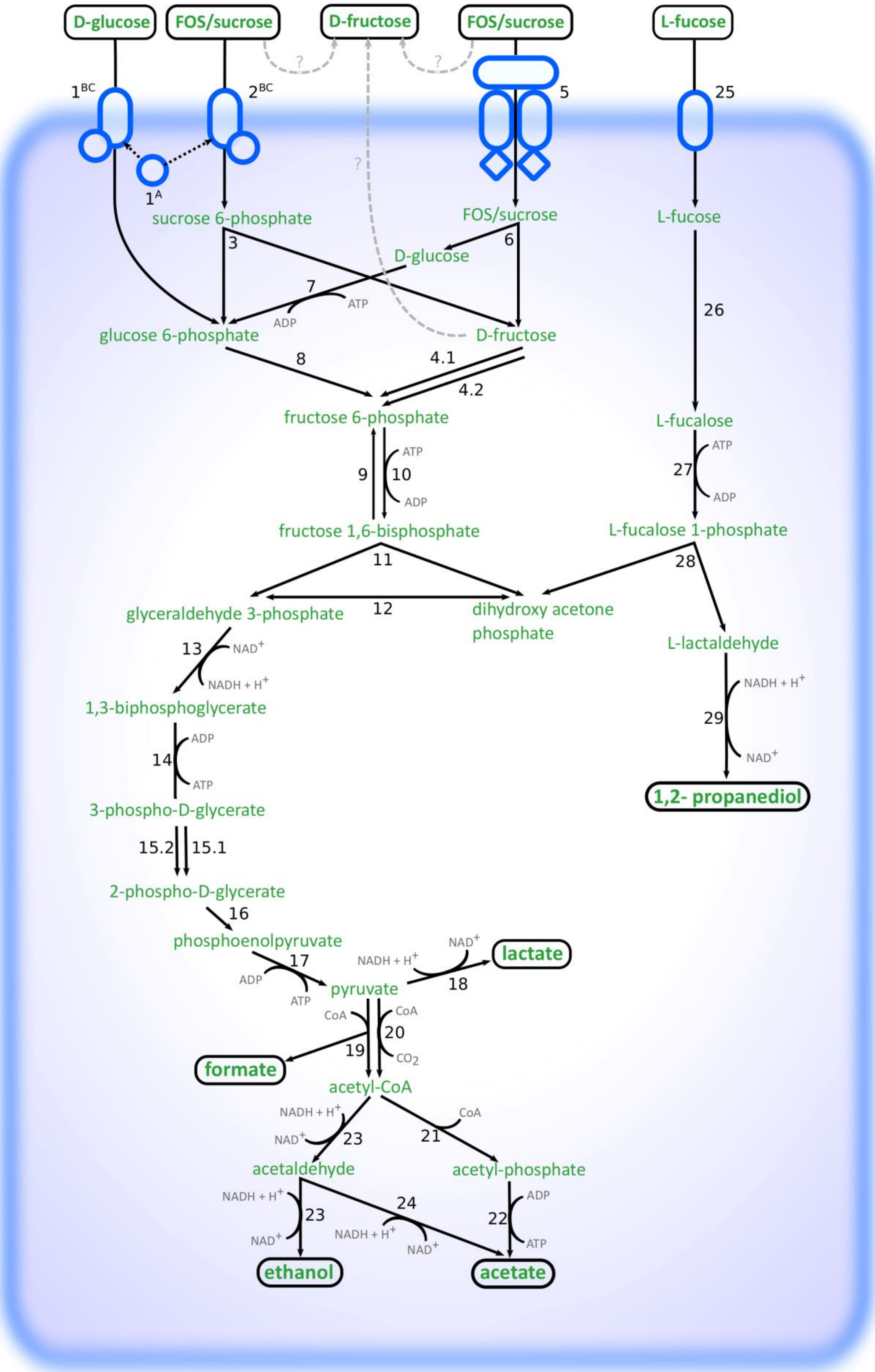


Figure 3: Heatmap of genes differentially expressed in at least one of the four conditions (≥1.5 log2 (fold change) and *q* value ≤ 0.05).
Colour coding by ratio to row mean. Key gene clusters are indicated



D-glucose
FOS/sucrose
D-fructose
FOS/sucrose
L-fucose
?
?
?
1BC
2BC
1A
5
25
sucrose 6-phosphate
FOS/sucrose
L-fucose
3
6
D-glucose
7
ATP
ADP
26
glucose 6-phosphate
D-fructose
8
4.1
4.2
fructose 6-phosphate
L-fucalose
ATP
9
10
ADP
27
ATP
ADP
fructose 1,6-bisphosphate
L-fucalose 1-phosphate
11
28
glyceraldehyde 3-phosphate
12
dihydroxy acetone phosphate
L-lactaldehyde
13
NAD+
NADH + H+
1,3-biphosphoglycerate
NADH + H+
29
NAD+
14
ADP
ATP
1,2- propanediol
3-phospho-D-glycerate
15.2
15.1
2-phospho-D-glycerate
16
phosphoenolpyruvate
17
ADP
ATP
NADH + H+
NAD+
lactate
pyruvate
18
CoA
CoA
20
19
CO2
formate
acetyl-CoA
NADH + H+
NAD+
23
21
CoA
acetaldehyde
acetyl-phosphate
NADH + H+
23
NAD+
24
NADH + H+
NAD+
22
ADP
ATP
ethanol
acetate

Figure 4: Schematic overview of the pathways involved in degradation of glucose, FOS and L-fucose in *R. ilealis* CRIB^T.
1^A; PTS system glucose-specific EIIA component (CRIB_2018); 1^{BC}, PTS system glucose-specific EIIBC component (CRIB_2017); 2^{BC}, PTS system sucrose-specific EIIBC component (CRIB_1461); 3, ß-fructofuranosidase with RDD family protein (CRIB_1459 and CRIB_1460); 4, fructokinase (CRIB_152 and CRIB_1458); 5; ABC-type transporter (CRIB_148-CRIB_150); 6, ß-fructofuranosidase (CRIB_151); 7, glucokinase (CRIB_1849); 8, glucose 6-phosphate isomerase (CRIB_140); 9, fructose 1,6-bisphosphatase (CRIB_45 and CRIB_2020); 10, 6-phosphofructokinase ; (CRIB_104); 11, fructose-bisphosphate aldolase (CRIB_2223); 12, triosephosphate isomerase (CRIB_189); 13, glyceraldehyde-3-phosphate dehydrogenase (CRIB_187); 14, phosphoglycerate kinase; 15, phosphoglycerate mutase (CRIB_1223) and 2,3-bisphosphoglycerate-independent phosphoglycerate mutase (CRIB_190); 16, enolase (CRIB_191); 17, pyruvate kinase (CRIB_105); 18, L-lactate dehydrogenase (CRIB_684); 19, formate acetyltransferase (CRIB_2141); 20, pyruvate-flavodoxin oxidoreductase (CRIB_2021); 21, phosphate acetyltransferase (CRIB_2171); 22, acetate kinase (CRIB_1927); 23, bifunctional aldehyde-alcohol dehydrogenase (CRIB_2231); 24, fatty aldehyde dehydrogenase (CRIB_2231); 25, L-fucose permease (CRIB_1294); 26, L-fucose isomerase (CRIB_1298); 27, L-fuculokinase (CRIB_1297); 28, L-fuculose phosphate aldolase (CRIB_1297); 29, lactaldehyde reductase (CRIB_1300); ?, possible mechanisms of external fructose accumulation (external degradation, or export).

The gene cluster involved in glycolysis (CRIB_186-CRIB_191) was most abundantly expressed in the conditions that support the highest growth rates determined by the highest cell density reached in the time period that was measured (glucose, followed by FOS; Fig. 3). This was also reflected in the fact that expression of genes encoding proteins involved in replication such as ribosomal proteins, proteins involved in cell wall biosynthesis and general cell division processes were most strongly expressed during growth in the presence of glucose and to a lesser extent FOS. Other genes involved in the central sugar metabolic pathways (e.g. CRIB_1849, CRIB_140, CRIB_2223, and CRIB_105) were upregulated in these conditions, albeit not significantly differentially regulated. This suggests that these genes are less tightly regulated at the transcriptional level, probably because they are also involved in other processes than sugar degradation [351]. The metabolic model suggests that this is indeed the case, as some of the enzymes produce intermediates which can be consumed by fatty acid biosynthesis and amino acid biosynthesis processes.

Altogether, the transcriptome of *R. ilealis* CRIB^T grown on FOS was very similar to its transcriptome when grown on glucose (Fig. 2), with only 18 genes significantly upregulated during growth in the presence of FOS compared to glucose (Table S4). Apparent was the upregulation of the gene clusters that code for proteins involved in the transport and degradation of the respective sugars or their derivatives (Fig. 3). In the presence of glucose, the glucose-specific PTS system (CIRB_2017-CRIB_2018) was significantly upregulated, together with its associated transcriptional regulator (CRIB_2019). In turn, in the presence of FOS, two clusters predicted to be involved in sucrose degradation (CRIB_148-CRIB_152 and CRIB_1458-1461) were significantly upregulated. The third gene cluster predicted to be involved in sucrose degradation (CRIB_1399-1400) was not significantly regulated during growth on FOS. However, it should be noted that these genes are located in a cluster functionally annotated to melibiose metabolism and are most likely regulated by the transcriptional regulator in this cluster. In addition to the two sucrose degradation clusters, a transport cluster of unknown function (CRIB_1506-CRIB_1509) was upregulated during growth on FOS, albeit only significantly when compared to growth on glucose. During growth in the presence of L-fucose, the gene cluster predicted to be involved in L-fucose degradation (CRIB_1294-CRIB_1298) was significantly upregulated, including the gene encoding the corresponding transcriptional regulator (CRIB_1299). An overview of the main carbohydrate degradation pathways regulated in the different conditions is given in Figure 4.

During growth on glucose, L-lactate dehydrogenase (CRIB_684) was significantly upregulated, albeit not significantly compared to growth on FOS. This enzyme catalyses the reduction of pyruvate resulting in the production of L-lactate and the reoxidation of the NADH formed during glycolysis. Only at the time point of 24 h, lactate was observed (Table 2). This suggests that at time point ~8-10 h the cells were starting to regenerate NAD by upregulating this gene. In the presence of L-fucose, NAD^+ regeneration is achieved via the reduction of lactaldehyde to 1,2-propanediol by lactaldehyde reductase (CRIB_1300), which was upregulated in the presence of L-fucose together with the L-fucose degradation gene cluster. In the spent medium of L-fucose grown cells, 1,2-propanediol was already seen at time point ~8-10 h, whereas no lactate production was observed. Another way to regenerate NAD^+ is to reduce pyruvate to ethanol (Fig. 4). In the presence of both glucose and FOS, an upregulation was seen for the gene encoding the bifunctional aldehyde/alcohol dehydrogenase (CRIB_2231), which converts acetyl-CoA to ethanol. However, in none of the samples, ethanol was measured by HPLC analysis.

During growth on FOS, a small gene cluster (CRIB_601-CRIB_603) that includes a gene encoding an alternative sigma factor was significantly upregulated. This was also apparent in the control culture that was inoculated with FOS-preconditioned cells. This suggests that in the presence of FOS (or its derivatives sucrose or fructose) transcription is also regulated by RNA polymerase promoter recognition.

Expression and regulation of other environmentally relevant functions in *R. ilealis* CRIB^T

Noteworthy was the significant upregulation of a gene cluster related to iron transport (CRIB_892-CRIB-898) during growth on glucose and FOS compared to growth on L-fucose. The significance of this gene cluster for carbohydrate utilization is not known, however, several enzymes could be identified in the genome of *R. ilealis* CRIB^T that use different forms of iron as cofactor, for example the hydrogenases involved in hydrogen metabolism [352], several ferredoxins, and the L-threonine dehydratase (CRIB-426) that was significantly upregulated during growth on L-fucose. As multiple transporters involved in the transport of iron compounds were predicted, it is possible that the uptake of iron provides a competitive advantage to other microbes that are dependent on iron for respiration and other metabolic processes [353].

Prevalence of *R. ilealis* in human datasets

R. ilealis was found to be a natural and abundant inhabitant of the rat small intestine, specifically of the ileum [309]. To study its prevalence in humans, 16S rRNA amplicon sequencing datasets were investigated for the presence of *R. ilealis*-like 16S rRNA gene sequences. Unfortunately, with respect to composition analysis of human ileum samples, there is only a limited number of datasets available due to the sampling difficulties that are the result of the inaccessibility of this part of the intestinal tract. In the dataset published by [331], a paediatric human dataset with samples from both healthy individuals and inflammatory bowel disease patients, we were not able to identify any *Romboutsia*-like 16S rRNA gene sequences. In the dataset published by [330], one of the biggest 16S rRNA gene datasets published to date that includes samples obtained from multiple gastrointestinal locations (ileal and rectal biopsies and faecal samples) from both healthy individuals and inflammatory bowel disease patients, only a limited number of *R. ilealis*-like 16S rRNA gene sequences could be identified. In this dataset, the genus *Romboutsia* could be identified in two samples, with a relative abundance of 0.1% and 0.2%. In 173 cases the family *Peptostreptococcaceae* could be identified in these samples, but it was not possible to differentiate between the genera *Romboutsia* and *Intestinibacter*, due to 100% identity of their rRNA gene sequences in this region. The *Peptostreptococcaceae*-positive samples were obtained from both healthy and diseased individuals (including at least one with ileal overgrowth and a *Peptostreptococcaceae* abundance of 46%). It should be noted that both datasets contain only sequence data from paediatric ileal biopsy samples and therefore only mucosa-associated microbiota could be studied limited to a human population <17 years of age, which could explain the low prevalence of *R. ilealis*-like 16S rRNA gene sequences. Unfortunately, due to the limited number of available human datasets and the low prevalence of *R. ilealis*-like 16S rRNA gene sequences in ileal biopsy samples, it was not possible to find positive or negative correlations between prevalence and/or abundance of *R. ilealis* and specific human diseases.

Discussion

Gerritsen *et al*. [309] have shown by 16S rRNA gene sequence-based analysis that *R. ilealis* CRIB^T is a dominant member of the small intestine microbiota in rats, especially in the ileum. The genomic and transcriptomic analysis of *R. ilealis* CRIB^T reported here provides new insights into the genetic and functional potential of this inhabitant of the small intestine. Genomic analysis revealed the presence of metabolic pathways for the utilization of a wide array of simple carbohydrates in addition to a multitude of carbohydrate uptake systems that included a series of PTS systems, carbohydrate specific ABC transporters, permeases and symporters. This is in agreement with prior observations by [70], who reported that the small intestinal microbiome is enriched for genes involved in the consumption of simple carbohydrates. However, small disagreements with prior observations were also observed. An enrichment for amino acid metabolism [70] was not visible in *R. ilealis* CRIB^T, and considerable less COGs could be classified than for the average small intestinal bacterium [179].

Since the small intestine is an environment in which environmental conditions change quickly due to the varying food intake of the host, microorganisms in this environment must be able to respond rapidly to such changes. As previously mentioned, the high number of rRNA operons found in the genome of *R. ilealis* CRIB^T is an indication that this this strain is indeed able to adapt its metabolism quickly in response to changing conditions, as a high rRNA copy number has been associated with this trait [333]. Considering the small intestinal habitat, we chose to focus on key pathways involved in the utilization of specific diet- and host-derived carbon sources by whole-genome transcriptome analysis.

Degradation of FOS and its possible role in cross-feeding

In the intestinal tract, the diet-derived carbohydrates that the host is unable to digest are important sources of energy for many microbes. In return, the host is dependent on the degradation of food-derived indigestible component by microbes for the release of certain essential metabolites (e.g. SCFA). Here we examined the growth of *R. ilealis* CRIB^T on FOS, a relatively simple oligosaccharide that is indigestible by the host, and the metabolites that were released. The transcriptome of *R. ilealis* CRIB^T grown on FOS was very similar to its transcriptome when grown on glucose, a monosaccharide used by the majority of microbes present in the intestinal tract. This is not surprising considering that glucose in addition to fructose is one of the two subunits present in FOS. Noteworthy was the accumulation of fructose in the culture supernatant during growth of *R. ilealis* CRIB^T on FOS. Based on the genomic analysis there are no apparent reasons why fructose should not be metabolized as all the necessary metabolic enzymes are present. However, it has been previously observed that *R. ilealis* only grows weakly on D-fructose [316]. The absence of a fructose-specific transporter, which could be identified in close relatives that are able to grow on D-fructose, might explain the fructose accumulation during growth of *R. ilealis* CRIB^T on FOS.

Differential gene expression analysis demonstrated the apparent FOS-induced upregulation of two separate gene clusters that were predicted to be involved in sucrose transport and degradation. However, based on the genomic analysis no apparent pathways could be identified to be responsible for FOS degradation. A simple explanation for the observed growth on FOS could be extracellular degradation of FOS, followed by import of sucrose and/or glucose into the cell. Fructan degradation by extracellular enzymes is described for other (intestinal) microbes [354]. The observed accumulation of fructose during growth of *R. ilealis* CRIB^T on FOS supports the hypothesis of extracellular degradation. However, no extracellular fructansucrase or glucansucrase could be predicted. Furthermore, no new candidates for this activity could be identified via the differential gene expression analysis described here. However, one possible candidate could be the predicted beta-fructofuranosidase present in the PTS system-containing sucrose degradation gene cluster. Next to the beta-fructofuranosidase-encoding gene, a gene was found to which no function could be assigned, but that was predicted to have a transmembrane region and a domain which could be involved in transport. Given that both loci overlap by a few nucleotides, and that the overlap is within a homopolymer region, it is possible that both loci form one protein due to ribosomal slippage on the homopolymer [355]. This could possibly lead to an external membrane-bound enzymatically active protein, which would explain the accumulation of fructose. Future studies with mutant strains might shed more light on the specific contribution of the two predicted sucrose degradation gene clusters to the degradation of FOS, or even longer fructans (e.g. inulin), by *R. ilealis* CRIB^T. Altogether, these results might indicate a possible role for *R. ilealis* CRIB^T in intestinal cross-feeding networks by releasing D-fructose during growth on fructans like FOS, which can function as growth substrate for other microbes or be directly absorbed by the host.

Fucose degradation and its advantages

Besides diet-derived carbohydrates, also host-derived carbohydrates are an important source of energy for some microbes. Unlike other members of the family *Peptostreptococcaceae*, *R. ilealis* CRIB^{T} is able to grow on L-fucose, a predominantly host-derived carbon source [316]. The transcriptome analysis confirmed the presence of a functional L-fucose degradation pathway, similar to the pathways previously identified in other intestinal inhabitants such as *E. coli* [356], *Bacteroides thetaiotaomicron* [357] and *Roseburia inulinivorans* [358]. By gene sequence homology a similar pathway was found in *Clostridium perfringens* and the more closely related *C. sordellii* (Fig. S2). L-fucose is a common sugar present within the intestinal environment, since it is a monosaccharide that is an abundant component of many N- and O-linked glycans and glycolipids produced by mammalian cells, including the fucosylated glycans that are found at the terminal positions of mucin glycoproteins [359]. Fucosylated mucin glycoproteins are especially found in the (human) ileum [360, 361]. For both intestinal commensals and pathogens the ability to utilize L-fucose has been demonstrated to provide a competitive advantage in the intestinal environment [357, 362]. In *R. ilealis*, all enzymes for L-fucose degradation are present in one cluster, however, no fucosidase-encoding gene could be identified, which means that *R. ilealis* is not able to release L-fucose units from fucosylated glycans (e.g. mucin) by itself. Hence, in the intestinal environment *R. ilealis* is dependent on free L-fucose monosaccharides released by other microbes. Furthermore, a gene cluster involved in degradation of sialic acid [343, 363, 364] was predicted from the genome, but no extracellular sialidase could be identified, which is similar to what has been found for *C. difficile* [344]. This suggests that also for sialic acid, a common residue found in mucin glycoproteins, *R. ilealis* CRIB^{T} seems to be dependent on the activity of other microbes. However, this also suggests that by its ability to use L-fucose and sialic acid monosaccharides, *R. ilealis* CRIB^{T} is dependent for these host-derived sugars that are released by the action of extracellular enzymes of with mucus-degrading microbes like *B. thetaiotaomicron* or *Akkermansia muciniphila*. Besides niche competition with other commensals, fucose utilization may also be important in niche competition with pathogens. It was recently suggested that the host is able to regulate fucosylation of its intestinal epithelial cells in response to pathogen-induced stress and that microbes that are able to use fucose as an energy source may contribute to the protection of the host against infections by endogenous pathogens [365].

Regulation of carbohydrate catabolism

In the intestinal environment *R. ilealis* CRIB^{T} will encounter a wide array of carbohydrates that are either continually or transiently present. Prioritization of carbohydrate utilization is partly achieved at the transcriptional level by the selective expression of genes. The primary mechanism by which bacteria regulate the utilization of non-preferred carbohydrates in the presence of preferred carbon sources is known as carbon catabolite repression (CCR), a hierarchical system for coordinating sugar metabolism [366]. The fact that, compared to glucose and FOS, L-fucose is utilized by a pathway that does not directly involve fructose-1,6-bisphosphate, a key metabolite in the regulation of CCR of Gram-positive bacteria, made it possible to study CCR by either glucose or FOS. The transcriptome analysis suggests that some genes and operons in *R. ilealis* CRIB^{T} were indeed subject to CCR in response to the presence of glucose. For example, two gene clusters predicted to be involved in hexuronate metabolism (CRIB_649-CRIB_652 and CRIB_2244-CRIB_2249), pathways that make the use of Dglucuronate and D-galacturonates as sole carbon source possible, were significantly upregulated during growth in the presence of L-fucose compared to growth on glucose (Table S5). In addition, the gene cluster predicted to be involved in sialic acid utilization (CRIB_613-CRIB_616) was downregulated in the presence of glucose as well. Furthermore, when comparing the expression of the gene cluster involved in L-fucose degradation during growth on glucose relative to the growth in the absence of a carbon source (control condition), this gene cluster appeared to be under CCR as well, in the presence of glucose (Table S5). These results suggest that in *R. ilealis* CRIB^T, multiple gene clusters that are involved in the use of alternative carbon sources are subject to CCR.

Expression and regulation of niche-specific functionalities in R. ilealis CRIB^T

Microbes residing in the intestinal tract have to withstand the harsh environmental conditions specific for the intestine. In this context, it was interesting that we identified a urease gene cluster in *R. ilealis* CRIB^T (CRIB_1381-CRIB_1388), expression of which appeared to be induced in carbon source limiting circumstances. The fact that this gene cluster was significantly upregulated when grown in the absence of an additional carbon source compared to growth on glucose, possibly suggests CCR of the urease gene cluster. However, upregulation of this gene cluster in the absence of an exogenous carbon source might also be a possible mechanism. Urea in the intestinal tract is derived from the breakdown of amino acids. *Helicobacter pylori* is a well-known example where urease activity contributes to the survival of the bacterium in the acidic environment of the stomach [367]. For some of the urease-positive bacteria, this enzyme has been shown to act as a virulence factor as it is responsible for urea hydrolysis that leads to increased pH and ammonia toxicity [349]. However, for commensal intestinal bacteria ureases can probably function as colonization factors as well, as they contribute in general to acid resistance and thereby play a role in gastrointestinal survival [367]. Urea is released into all parts of the intestinal tract via diffusion from the blood, but it has been reported that pancreatic excretions and bile are a main route of entry [368]. So far, we have not been able to demonstrate urease activity in *R. ilealis* CRIB^T [316]. However, different mechanisms for the expression of urease have been identified in other microbes: constitutive, inducible by urea, or controlled by nitrogen source availability [347]. For *C. perfringens* for example, the urease activity, which is plasmid borne, was shown to be only expressed in nitrogen-limiting conditions [369]. The increased urease gene expression by *R. ilealis* CRIB^T observed in the control condition, in the absence of an additional carbohydrate, suggests an alternative mechanism for regulation of urease gene expression.

Conclusions

We are just starting to elucidate the composition and function of the microbial communities in the mammalian small intestine. Recently we have reported the isolation and characterization of *R. ilealis* CRIB^T from the small intestine of a rat [316]. In rats, this species was identified to be a dominant member of the ileal microbiota [309]. Here we applied a holistic systems biology approach, involving several fields of experimental and theoretical biology, to study *R. ilealis* CRIB^T. In conclusion, *R. ilealis* CRIB^T is a strain that is able to utilize an array of carbohydrates using different and partially redundant pathways. Its ability to use host-derived sugars that are liberated by other microbes suggests that *R. ilealis* CRIB^T is dependent on mucus-degrading microbes, like *B. thetaiotaomicron* or *A. muciniphila*. In contrast, it has only limited ability to *de novo* synthesize amino acids and vitamins, and hence the organism shows an adaption to a nutrient-rich environment in which carbohydrates and exogenous sources of amino acids and vitamins are abundantly available. In addition, we were able to pinpoint potential mechanisms that might enable this organism to survive in the competitive small intestinal environment. These mechanisms include bile salt hydrolase and urease enzymes, which enhance the organism's ability to handle in particular small-intestinal conditions.

It has to be emphasized that the results presented in this study correspond to one specific strain and that different strains belonging to the same species could possibly encode for different functions, including utilisation of specific glycans as previously described by [370]. However, a deeper investigation of key players in the intestinal tract like *R. ilealis* CRIB^T and others will lead to a better understanding of how the microbial communities in us function as a whole. The more we understand how each organism works, and how they interact, the better we get an insight into these environments and can predict how nutrition will influence our health and well-being.

Acknowledgements

The authors thank Hans Heilig (Wageningen University, Laboratory of Microbiology) for his help with genomic DNA isolations and Jasper Koehorst (Wageningen University, Laboratory of Systems and Synthetic Biology) for his help with the genome annotation.

Funding

This work was supported by a grant of SenterNovum (Agentschap NL), an agency of the Dutch Ministry of Economic Affairs (FND-07013). B Hornung is supported by Wageningen University and the Wageningen Institute for Environment and Climate Research (WIMEK) through the IP/OP program Systems Biology (project KB-17-003.02-023). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary information

Supplementary information can be found online at:

https://doi.org/10.7717/peerj.3698

Competing Interests

J Gerritsen is currently an employee of Winclove Probiotics. However, these associations do not influence the objectivity, integrity and interpretation of the results that presented in this manuscript. Sacha A.F.T. van Hijum is an employee of NIZO, Kernhemseweg 2, 6718 ZB, Ede, the Netherlands and Vitor A.P. Martins dos Santos is an employee of LifeGlimmer GmbH, Markelstrasse 38, Berlin, Germany. Willem M. de Vos and Hauke Smidt are Academic Editors for PeerJ.

Chapter 4: Isomalto/malto-polysaccharides maintain normal gut functioning while promoting growth and activity of beneficial bacteria

Manuscript in preparation:

Klaudyna Borewicz[*], Bastian Hornung[*], Fangjie Gu, Pieter H. van der Zaal, Henk Schols, Peter J. Schaap, Hauke Smidt. "Isomalto/malto-polysaccharides maintain normal gut functioning while promoting growth and activity of beneficial bacteria"

*Contributed equally

Abstract

Isomalto/malto-polysaccharides (IMMPs) are a novel type of soluble dietary fibres with a prebiotic potential capable of promoting growth of beneficial microbes in the gut. However, the mode of action of IMMPs remains unknown. Previous studies on IMMPs showed an increase in total bacteria, especially lactobacilli, and higher production of short chain fatty acids (SCFA) when IMMPs were fed to rats or used during *in vitro* fermentation. In this study, we investigated with metatranscriptomics how IMMPs with different amounts of α-(1→6) glycosidic linkages affected microbial function during incubation with human faecal inoculum. We showed that microbial community dynamics during fermentation varied depending on the type of IMMP used and that the observed changes were reflected in the community gene expression profiles. Based on metatranscriptome analysis, members of *Bacteroides*, *Lactobacillus* and *Bifidobacterium* were the predominant degraders of IMMPs, and the increased activity of these bacteria correlated with high amounts of α-(1→6) glycosidic linkages. We also noted an increase in relative abundance of these bacteria and an activation of pathways involved in SCFA synthesis. Our findings could provide a baseline for more targeted approaches in designing prebiotics for specific bacteria and to achieve more controlled modulation of microbial activity towards desired health outcomes.

Introduction

The human gut is home to a diverse ecosystem inhabited by bacteria, archaea, viruses and eukaryotes, which play an important role in their host's health and well-being [371-373]. These organisms interact with each other and with the host via a complex network of relations, and knowing the mechanisms of these interactions and how to influence them might provide a useful tool for refining the function of this ecosystem to promote homeostasis and to strengthen host's immunity against infections [374]. Currently there are only a few ways to manipulate the composition and function of the gut microbiota. These range from mild measures, such as the implementation of various dietary regimes and the use of dietary supplements, especially pro- and prebiotics [300], to more extreme ones, such as the use of antibiotics [375] or faecal transplantations [376]. Prebiotics are complex carbohydrates, often soluble dietary fibres, that cannot be digested by human enzymes but are readily used by the colonic microbiota and provide a health benefit for the host [281]. A range of different prebiotics may preferably stimulate growth and activity of specific microbial groups (e.g. butyrogenic bacteria [377]), leading to the production of different metabolites with health-supporting effects. However, the exact mode of action of most prebiotics remains unknown and their specific impact on microbial interactive networks needs to be investigated.

Isomalto/malto-polysaccharides (IMMPs) comprise a novel class of soluble dietary fibres with prebiotic potential. These fibres are synthetized from starch by enzymatic conversion of α-(1→4) glycosidic linkages into α-(1→6) glycosidic linkages by 4,6-α-glucanotransferase (GTFB) from *Lactobacillus reuteri* 121 [378]. The resulting α-(1→6) linkages present in IMMPs make these fibres resistant to digestion by human digestive enzymes in the small intestine. As such, this modified starch can pass undigested into the large intestine where it is fermented by the resident microbes capable of breaking down the α-(1→6) glycosidic linkages. This property of the IMMPs makes them potentially interesting as a prebiotic food ingredient capable of modulating the intestinal microbiota and exerting health promoting effects onto the host. A previous study has reported an increased production of short chain fatty acids (SCFA), especially acetate and propionate, when IMMPs were used as a carbon source for microbial *in vitro* fermentation with human faecal inoculum as the microbial source [378]. In this study, we investigated the effects of three different IMMPs on microbial composition and function during *in vitro* batch fermentations with faecal inoculum from healthy human adults. Here we show that specific changes of the microbiota, such as growth of *Bifidobacterium* and *Lactobacillus* can be attributed to the IMMPs, and that these changes are also reflected at the transcriptomic level, i.e. upregulation of specific gene groups, as well as in enzymatic activity and increase in production of SCFA.

Materials and Methods

In vitro fermentation; design and sampling

The faecal inoculum stock was prepared at TNO (Zeist, The Netherlands) from fresh faeces of seven healthy adult donors. The stock was mixed, aliquoted and stored anaerobically at -80 °C [379]. Sterile 20 mL anaerobic serum bottles were filled with 10 mL of the Standard Ileal Efflux Medium (SIEM; Tritium Microbiology, Eindhoven, The Netherlands). The SIEM was prepared according to Rösch *et al.* [380], but omitting the carbon source and Tween 80. The modified SIEM medium contained 40% (v/v) BCO medium, 1.6% (v/v) salt solution, 0.8% (v/v) MgSO4 (50 g/L), 0.4% (v/v) cysteine hydrochloride (40 g/L), 0.08% (v/v) vitamin solution and 10% (v/v) MES buffer (1 M, pH 6.0) in water. Before inoculation, a faeces stock aliquot was mixed with SIEM at 1:10 v/v and incubated overnight at 37 °C. The activated inoculum was then added to the fermentation bottles at 1% (v/v) final concentration. Three different IMMP fibres were tested, with 27% (IMMP-27), 94% (IMMP-94) and 96% (IMMP-96) of α-(1→6) glycosidic linkages as compared to the total amount of glycosidic linkages. In addition, a pre-treated IMMP-27 (IMMP-dig27) sample was included after it had been digested with α-amylase and amyloglucosidase to imitate passage through the small intestine [381]. Samples were prepared and processed in duplicate with fibres added to individual fermentation bottles at a final concentration of 10 mg/mL. Flasks were incubated at 37 °C, and 0.5 to 2 mL of each culture was removed at different sampling time points, depending on the experiment.

In experiment A, cultures supplied with two different prebiotic fibres (IMMP-27 and IMMP-94) and one control culture without any substrate (IMMP blank) were monitored over 48 hours (in duplicate), and aliquots were removed at time points 0 (up to 15 min after addition of the prebiotic), 24 h and 48 h. In experiment B, cultures were supplied with two other prebiotics, IMMP-dig27 and IMMP-96, and one culture was left with no prebiotic (IMMP blank). Experiment B was monitored for 48 hours, and samples were taken at 6 h, 12 h, 24 h and 48 h (in duplicate). An aliquot of the activated blank inoculum was taken at time point 0, just before the addition of the IMMP. All samples (18) from experiment A were subjected to metatranscriptomic sequencing. In experiment B the metatranscriptomics sequencing was done for the activated inoculum at time point t0 and for the treatment groups at all time points (17). Samples for metatranscriptomics were harvested and immediately stabilized in RNAprotect (Qiagen, Hilden, Germany) following the manufacturer's instructions, and bacterial pellets were stored at -80 °C for up to three weeks before further processing.

RNA extraction and Illumina sequencing

Total RNA was extracted by using the beat beating - TRIzol - column method modified from Kang *et al*. [382]. Briefly, bacterial pellets were re-suspended in 100 μL TE buffer (30 mM Tris-HCl, 1 mM EDTA, pH = 8.0) containing 15 mg/mL Lysozyme, 10 U/mL of Mutanolysin and 100 μg/mL of Proteinase K. Samples were vortexed for 10 s and incubated at room temperature for 10 min, and 400 μL of RLT buffer (Qiagen, Hilden, Germany) containing 4 μL of β-mercaptoethanol was added. Samples were then vortexed, mixed with 500 μL of TRIzol Max reagent (Invitrogen, Carlsbad, CA, USA) and homogenized with 0.8 g of sterilized 0.1 mm zirconia beads for three min (3 × 1 min with cooling in between) at 5.5 ms using a bead beater (Precellys 24, Bertin Technologies). Following the beating step, samples were cooled on ice, gently mixed by inverting the tube with 200 μL of ice cold chloroform for 15 s and centrifuged for 15 min at 4 °C at 12,000 × g. The aqueous phase containing total RNA was transferred to fresh tubes and mixed with an equal volume of 70% ethanol. The mixture was placed on a Qiagen RNeasy mini column (RNeasy Mini Kit, Qiagen, Hilden, Germany) and centrifuged at 8,000 × g for 15 s to bind RNA into the column. Filtrate was discarded, and the RNA binding step was repeated until the complete sample was filtered through the column.

The columns were rinsed with 350 µL of RW1 buffer (RNeasy Mini Kit, Qiagen, Hilden, Germany), and 80 µL DNAse I solution (Roche, Manheim, Germany) was applied to the column and incubated for 15 min at RT to digest DNA. The columns were rinsed twice with 350 µL RW1 buffer, and twice with 700 µL of RPE buffer (RNeasy Mini Kit, Qiagen, Hilden, Germany), following with a final wash with 80% ethanol. Columns were dried by a 2 min centrifugation at maximum speed, and total RNA was eluted with 30 µL of DNAse/RNAse free water. The total RNA concentrations were measured spectrophotometrically with an ND-1000 spectrophotometer (NanoDrop® Technologies, Wilmington, DE, USA), and residual DNA concentrations were measured with the Qubit® dsDNA BR Assay Kit (Life Technologies, Leusden, the Netherlands). Samples which contained over 10 ng/µL DNA contamination were treated with the Turbo DNAfree® Kit (Ambion, Bleiswijk, Netherlands) following manufacturer's instructions and purified using the RNeasy Mini Kit. Total RNA quality was evaluated using the Experion RNA StdSens kit (Biorad Laboratories INC, USA), total RNA concentrations were measured with NanoDrop® and DNA contamination concentrations were measured with the Qubit®dsDNA BR Assay Kit. Between 3-5 µg of total RNA from each sample was used for mRNA enrichment with the RiboZero Bacterial rRNA Removal Kit (Illumina, San Diego, CA, USA), and the quality and quantity of enriched mRNA was assessed as described above for total RNA. Between 200-500 ng of enriched mRNA was used for cDNA production using the ScriptSeq®v2RNA-Seq Library Preparation Kit (Epicentre, Madison, WI, USA), FailSafe®PCR Enzyme Mix (Epicentre, Madison, WI, USA) and ScriptSeq®Index PCR Primers (Epicentre, Madison, WI, USA) for amplification and barcoding of di-tagged cDNA. The PCR product presence was confirmed with gel electrophoresis using the FlashGel® System (Lonza, Rockland, ME, USA). PCR products were then purified with the HighPrep® PCR kit (MagBio Genomics, Gaithersburg, MD, USA) and concentrations of indexed cDNA were measured using the Qubit®dsDNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA). Approximately 28 ng of DNA from each sample was added to a pool, and final volume of each library was adjusted to 25 µL using the HighPrep® PCR kit. Two libraries were prepared containing either 17 or 18 samples, with final concentrations of 20 ng/µL in each library. Libraries were sent for single end 150 bp Illumina HiSeq2000 sequencing (GATC, Konstanz, Germany).

Bioinformatic processing, read assembly and annotation

The bioinformatics workflow was adapted from Davids *et al.* [182]. SortMeRNA v1.9 [209] software was used to screen the metatranscriptome data against all databases deployed with the program and to remove rRNA reads. Adapters were trimmed with cutadapt v1.2.1 [383] using default settings. Quality trimming was performed with PRINSEQ Lite v0.20.0 [384] with a minimum sequence length of 40 bp and a minimum quality of 30 on both ends of the read, and as mean quality. All reads containing more than three Ns or non-IUPAC characters were discarded.

Reads from experiment A (Suppl. Figure S1) were pooled and assembled with IDBA_UD version 1.1.1 [385] using two rounds of assembly; firstly, with the options –min_count 200 and – min_support 5, and secondly, the reads, which could not be mapped to this assembly with bowtie2 v2.0.6 [327], standard parameters, were extracted, and assembled with standard options, but with the output from the previous run provided as long reads. Contigs with an A/T content of >80% were removed from the final assembly. Because both experiments A and B were performed with aliquots from the same inoculum, we did not include reads from experiment B in the assembly, but rather mapped reads to the assembly generated from reads obtained from experiment A as described below. Prodigal v2.5 was used for prediction of protein coding DNA sequences with the option for meta samples [191]. Protein sequences were annotated with InterProScan 5.4-47.0 [198] on the Dutch science grid (offered by the Dutch National Grid Initiative via SurfSara), and enriched by adding EC numbers using PRIAM version March 06, 2013 [386]. Carbohydrate active enzymes were predicted with dbCAN release 3.0 [196]. Further enrichment for EC numbers was obtained by matching all InterProScan derived domain names against the BRENDA database (download 13.06.13) [387] and using a text mining algorithm that included removal of the non-alphanumerical characters (colons, commas, brackets, etc.), partial and generic terms (type, terminal, subunit, domain, enzyme, like, etc.), as well as other smaller modifications. Details are provided in Supplementary Materials and Methods.

Read counts from experiment A and B (Figure S1) were obtained with Bowtie2 v2.0.6 [327] using default settings. BAM files were converted with SAMtools v0.1.18 [388], and gene coverage was calculated with subread version 1.4.6 [389]. Read mappings to the RNA-assemblies were inspected with Tablet [390].

Taxonomic assignments

RNA sequences from the metatranscriptome assembly were compared with Blast 2.2.29 [391] against the NCBI NT database (download 22.01.2014) using standard parameters, besides an E-value of 0.0001, to the human microbiome (download 08.05.2014), NCBI bacterial draft genomes (download 23.01.2014), NCBI protozoa genomes (download 08.05.2014), and the human genome (download 30.12.2013, release 08.08.2013, NCBI Homo sapiens annotation release 105). Taxonomy was estimated with a custom version of the LCA algorithm as implemented in MEGAN [392], but with the following changes: only hits, which exceeded a bit-score of 50 were considered, and of these, only hits with a length of more than 100 nucleotides and which did not deviate more than 10% from the longest hit were accepted.

From all sequences from the assembly, which did not have a match in any of the former blast analyses, another run with the – blastn option was performed against the same databases, and in case this did not yield any results, a blastp of the predicted proteins was performed against a custom version of the KEGG Orthology database (http://www.genome.jp/kegg/ko.html, download 25.04.2014). Taxonomic assignment was again performed with the LCA algorithm, and for the blastp run only hits which did not deviate by more than 10% from the hit with the maximum identity were considered.

Differential expression

Differential expression analysis was performed at genus level in R version 3.1.1 [393] with the TCC package release 1.6.5 [394], with 36 iterations and the combination of tmm normalization and edgeR, with an FDR=0.1. Only genes with a q-value (multitest corrected p-value) of less than 0.01 in any of the relevant comparisons were considered to be significantly differentially expressed, unless otherwise mentioned.

Metabolic mapping

Two rounds of clustering were performed to detect patterns in the expressed genes (Figure S2). All genera, which either had an average read count of >=10 per gene, or which exceeded 1% of all reads in any given condition, were clustered into groups based on relative counts per group using the k-means algorithm in Scipy version 1.6.1 [395]. To determine the stability of the clustering, 50 iterations with a clustering between 1 and 20 clusters were performed, with the option "iter" set to 100.000. Afterwards the average cluster support per amount of clusters over all the iterations was computed, and additionally, the clustering was investigated with a custom python implementation of clustergrams [396]. Within the clustered genera, genes with similar expression patterns were identified with the DBSCAN algorithm [397]. Clustering on expression patterns was performed with ELKI 0.7.0~20150828 [225], the –minpts parameter was fixed to 3 and the epsilon parameter was varied in percentages. Final clustering was evaluated using the Tau index as implemented in ELKI, and the clustering result with the best Tau was chosen, unless a lower Tau led to better cluster separation.

Only genes which were differentially expressed in at least one sampling time point in any of the incubations (i.e. Ino.BL, IMMP-27, IMMP-94, IMMP-96, IMMP-dig27), were considered in the clustering analysis. Genes were normalized per row before the clustering. All derived EC numbers were mapped with custom scripts onto the KEGG database [227] and visualized with Python Scipy version 1.6.1 and NumPy version 0.9.0 [395]. Correlations were calculated with the mentioned versions of Scipy/NumPy. Differentially expressed genes were mapped separately for groups of interest, and changed functions were derived from visual inspections. Cofactor requirements were investigated with the Expasy database [398].

Data accessibility

The raw data has been uploaded to the EBI under project number PRJEB13209.

Results

We performed two *in vitro* batch fermentation experiments to investigate the influence of different IMMPs on human faecal microbiota. Our aim was to understand how the IMMPs containing different amounts of α-(1→6) glycosidic linkages were broken down by bacteria over time, and how the chemical structure of these compounds affected the functional dynamics of the microbial community during fermentation. Experiment A included fermentation of IMMPs of varying percentage of α-(1→6) glycosidic linkages (27%, IMMP-27; 96%, IMMP-96) at three different time points. This was complemented by experiment B that was performed with IMMP with 94% α-(1→6) linkages (IMMP-94) and IMMP-27 after treatment with α-amylase and amyloglucosidase (IMMP-dig27). Furthermore, in experiment B an additional set of time points was evaluated to provide a more detailed understanding of microbial community dynamics. In both experiments a control blank that did not receive any IMMP substrate was included. We then performed metatranscriptome sequencing of all these samples, and assembled the resulting data into one reference metatranscriptome. Afterwards, machine learning techniques were applied to identify groups of similarly behaving bacteria and to discover consistent dynamic patterns in gene expression.

Quality control and statistics

The metatranscriptome was sequenced and subjected to a quality control process before the data was further analysed (Figure S1). As a result, 320 million reads (89% of the raw reads and 54% of all bases) passed the quality check and were used for assembly into contigs. In experiment A, the assembly yielded over 140,000 contigs, with more than 200,000 protein coding genes, and contained, on average, 81% of the input reads (range 71% - 85%) per sample. Read counts for experiment B were acquired by mapping to the same assembly obtained from experiment A (Table S1), and showed the same average mapping rate (81%, range 71% - 89%). After mapping, the biological replicates within each experiment showed a spearman correlation of on average 86% (range 78% - 93%), indicating good reproducibility within the sets of samples from the same treatment group.

Community structure and activity patterns

Taxonomic classification to at least the superkingdom of bacteria was assigned to 190,000 of the 200,000 genes obtained from the RNA-assemblies. Less than 3,000 genes were assigned to eukaryotes and less than 2,000 to Archaea. Of the bacterial groups, most genes were assigned to the orders Bacteroidales (>67,000), Clostridiales (>40,000), Lactobacillales (27,000) and Enterobacteriales (>14,000). The genus with the highest number of assigned genes was the genus *Bacteroides* (>54.000; Figure 1).

To identify bacterial activity patterns, we focused on RNA reads for which a KEGG Orthology (KO) or EC identifiers could be assigned. The percentage of reads with defined KO or EC ranged from 42% to 83% for different samples. Most of the expression data with assigned KO or EC identifiers came from 22 bacterial groups, of which 12 could be assigned to a known genus, and only a small number of genes was assigned to minor groups (3%), unclassifiable sequences (3%), and sequences not classifiable beyond the superkingdom bacteria (3.5%). In the activated inoculum at the start of the incubation (t0), unclassified Enterobacteriaceae were the most active group (Figure 1). However, once the incubation had started, the relative activity of *Bacteroides* increased in all treatment groups. In all samples combined across all treatment groups and time points, 39% of all expression data came from the genus *Bacteroides* and 27% from unclassified Enterobacteriaceae. Overall, the relative abundance of different bacterial groups based on the metatranscriptome data corresponded to the pattern in the relative abundance of different taxa based on the 16S rRNA gene analysis described previously by Gu *et al.* [381](Figure 2).

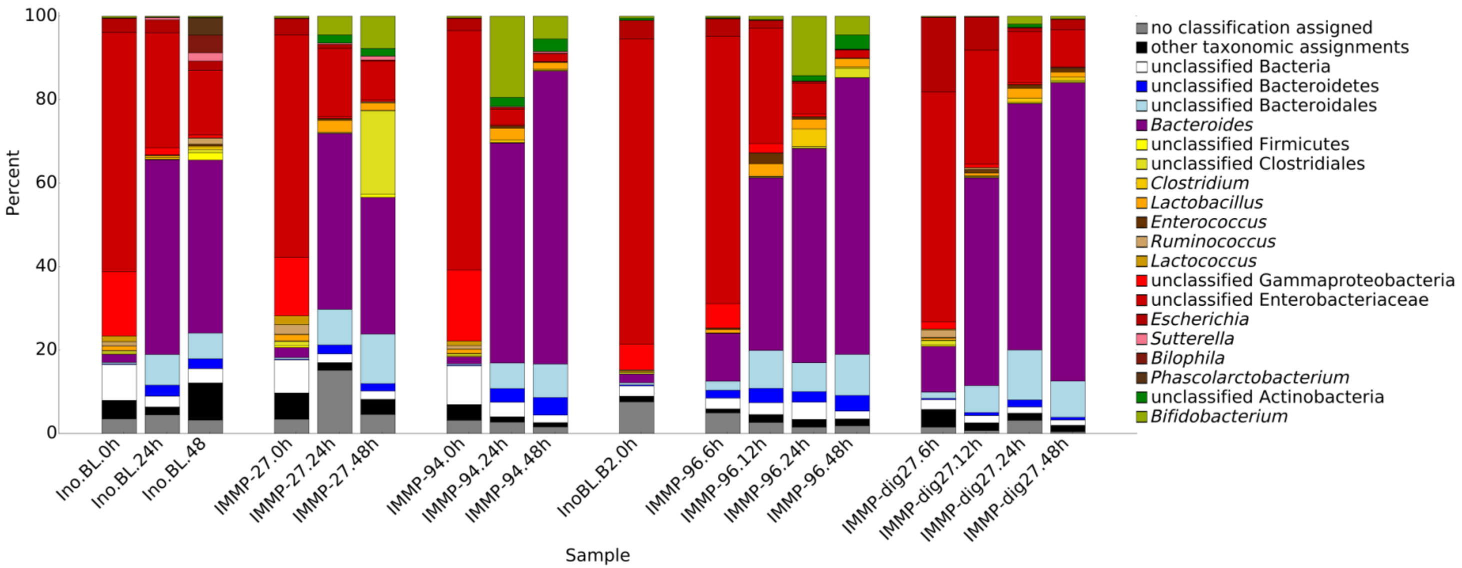


Figure 1: Average relative transcript expression of different genus level taxa in incubations sampled at time points 0 h, 6 h, 12 h, 24 h and 48 h. When the taxonomic assignment could not be made at genus level, the lowest classifiable taxonomy assignment was used for display. Low abundance genera are summarised as “Other genera” for display purposes.

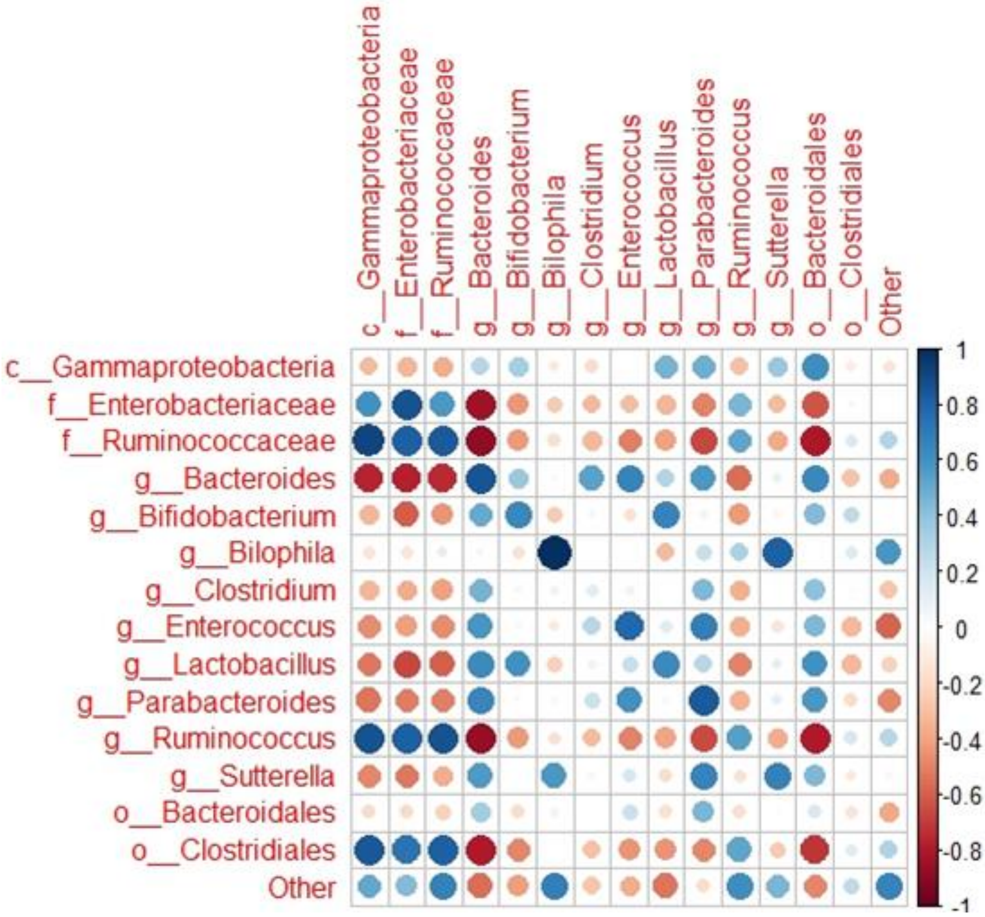


Figure 2: Correlation between the relative activity of the main bacterial groups based on metatranscriptome data, and their relative abundance based on 16S rRNA gene sequencing data (Gu *et al.,* unpublished). In case when genus level assignment was ambiguous, unclassified fraction within the next higher taxonomic level was used.

Global and IMMP specific co-occurrence of taxa

It is known that in microbial ecosystems bacterial taxa occupy different niches and co-exist forming a complex network of co-dependencies. We wanted to assess whether, based on the metatranscriptome data, we could identify bacterial groups which co-occurred in our samples and in relation to specific IMMPs. We performed clustering analysis based on mRNA reads from all samples in our dataset to test for global co-occurrence patterns. We showed that clustering into nine groups was most stable. An overview of organism assignment per cluster, with number of assigned genes and differentially expressed genes is provided in Table S2. One of these clusters was present in all t0 samples, but decreased or was absent at all other time points. This cluster consisted mostly of reads assigned to *Ruminococcus* and *Lactococcus* as well as reads that could be largely classified as contamination from the sampling (e.g. *Homo, Mus, Bos*, unclassified Mammalia). The second and third cluster consisted mainly of genera that also include many probiotic organisms, i.e. *Bifidobacterium, Lactobacillus*, and *Enterococcus,* and sequences, which could not be classified beyond a related higher order (e.g. unclassified Bifidobacteriaceae, unclassified Lactobacillaceae). These clusters also contained a related phage group (*Myoviridae,* mainly *Lactobacillus* phages), and an unrelated genus (*Fusobacterium*). The identified genera in cluster two and three showed an increasing pattern in terms of relative transcript abundance in all the cultures which were supplied with IMMP substrates, whereas relative transcript abundance was decreased or undetected in the control cultures without IMMPs. The fourth cluster was dominated by *E. coli* and related higher order classifications (e.g. unclassified Enterobacteriaceae), together with other enterobacteria such as *Enterobacter, Citrobacter* and *Klebsiella,* and the unrelated genus *Eubacterium*. This cluster was mainly present in the samples without prebiotics, and declined in the samples with prebiotics. The fifth cluster was dominated by *Bacteroides*, and showed an increase with time in all incubations. This cluster also included *Parabacteroides, Prevotella, Flavobacterium*, and *Desulfosporosinus*. The sixth cluster consisted only of *Clostridium*/unclassified Clostridia, which showed some increase with time in all incubations. The seventh cluster contained *Anaerostipes* and related higher order classifications (unclassified Clostridiales, unclassified Lachnospiraceae) and showed a similar pattern as cluster six. No clear pattern was seen for the eighth cluster consisting of *Corynebacterium, Ethanoligenes, Odoribacter*, and *Sutterella*. Finally, the ninth group consisted of different bacterial genera, some of which also containing known pathogens (*Bilophila, Phascolarctobacterium*), some related to non-carbohydrate metabolizing bacteria (*Acidaminococcus*), and some known gut symbionts like *Veillonella* and *Megasphaera*. This group was common in samples of incubations without any prebiotics at 48 h, and was nearly absent in all the other samples.

Detection of specific gene expression patterns

Besides the co-occurrence of bacterial groups, the specific gene expression patterns within these groups were investigated as well, based on the optimal gene clustering for all bacterial groups using DBSCAN. The clustering with the optimal tau was chosen for all bacterial groups, except for the genus *Enterococcus*, for which a suboptimal tau lead to better cluster separation. As a result, the DBSCAN gene clustering analysis revealed the presence of three main patterns in the expression in nearly all observed bacterial groups (Figure S3). These three patterns comprised in all cases at least 80% of all investigated genes, which were not considered noise. The first pattern was present in all incubations, and was characterized by genes which were expressed only at t0, and not expressed at any later time points. The second pattern was found only in the control group and only at 48 h. The third, and the most common pattern found in all experimental groups included genes that were not expressed at t0, but showed upregulation at the later time points during incubation. This pattern was characteristic for genes assigned to the genera *Enterococcus* and *Bacteroides*, which showed big gene clusters increasingly expressed over time in all treatment groups including the control group. *Bifidobacterium/Lactobacillus* and *Clostridium* also showed the same pattern, but only in the groups where IMMPs were present. *Eubacterium hallii,* showed the same gene expression pattern, but only in the group supplemented with IMMP-27 (Figure S3).

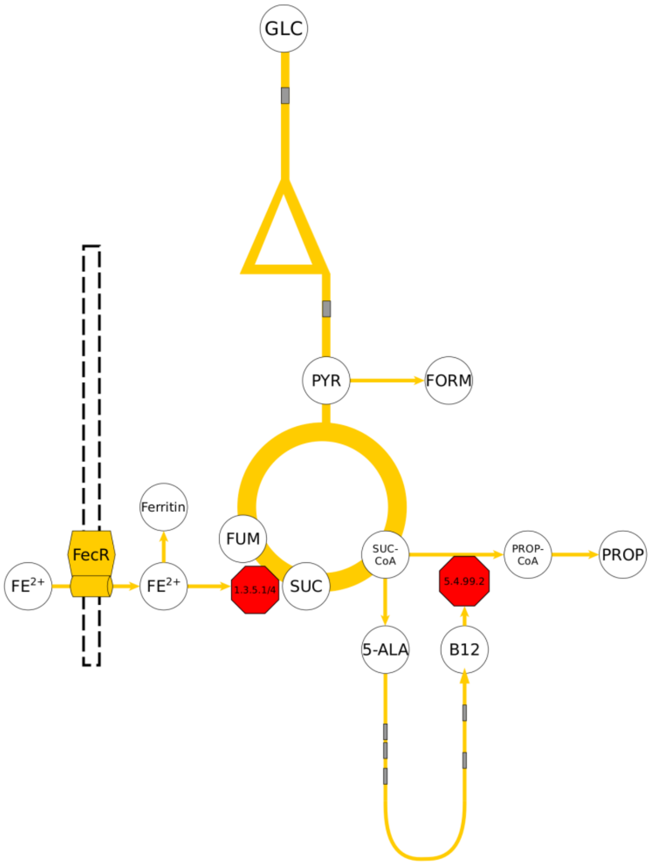
The expression levels of genes assigned to a specific bacterial group indicates its contribution to utilising the specified substrate, or its by-products. The high overall relative activity of bifidobacteria (and unclassified Bifidobacteriaceae), lactobacilli, enterococci, and unclassified Actinobacteria was positively correlated with the presence of IMMPs (Figure 1). Contrary, the activity of unclassified Proteobacteria, *Prevotella*, *Sutterella*, *Acinetobacter*, *Eggerthella*, *Acidaminococcus*, *Streptococcus*, *Phascolarctobacterium*, and *Bilophila* was negatively associated with the presence of IMMPs, as compared to the control group.

General metabolic effects of IMMP

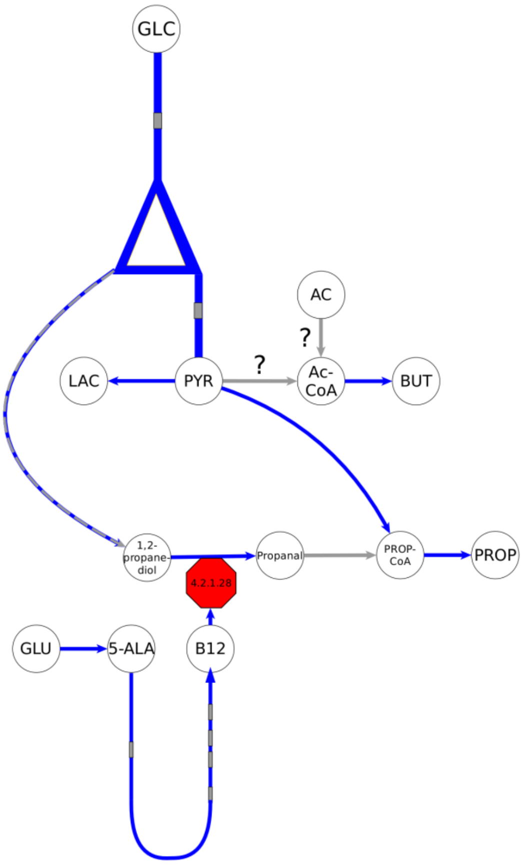
We wanted to further investigate the activity of the bacterial groups associated with the fermentation of different IMMPs. Our analysis of the metabolic clusters revealed that five bacterial groups found in the faecal inoculum, namely *Bifidobacterium/Lactobacillus, Enterococcus, Bacteroides, Clostridium,* and *Eubacterium hallii,* showed a considerable upregulation of general metabolic pathways like glycolysis, nucleic acid or fatty acid biosynthesis, as compared to the gene expression at t0. When we compared metabolic patterns between different bacterial groups, the groups exhibited overall different metabolic patterns. Members of the genus *Bacteroides* active in our incubations showed at first a unique partial upregulation of Vitamin B12 metabolism. An investigation of the cofactor requirements showed that Vitamin B12 in *Bacteroides* is essential for methionine synthase and methylmalonyl-CoA mutase, the latter of which produces methylmalonyl-CoA from succinyl-CoA (Figure 3).

Figure 3: Overview of the metabolism of specific microbial groups observed in the samples taken during *in vitro* fermentation of different IMMPs by human faecal inoculum All samples show in general the same patterns for all organisms, besides for *Eubacterium hallii*, which only showed expression in the samples with IMMP-dig27. The genus *Enterococcus* showed the same pattern as *Bifidobacterium/Lactobacillus*, but at lower relative transcript abundance. Grey indicates that certain genes were not differentially expressed within a pathway. 5-ALA = 5-Aminolevulinate, AC = Acetate, Ac-CoA = Acetyl-CoA, BUT = Butyrate, FORM = Formate, FUM = Fumarate , GLC = Glucose, LAC = Lactate, PROP = Propionate, PROP-CoA = Propanoyl-CoA, PYR = Pyruvate, SUC = Succinate, SUC-CoA = Succinyl-CoA

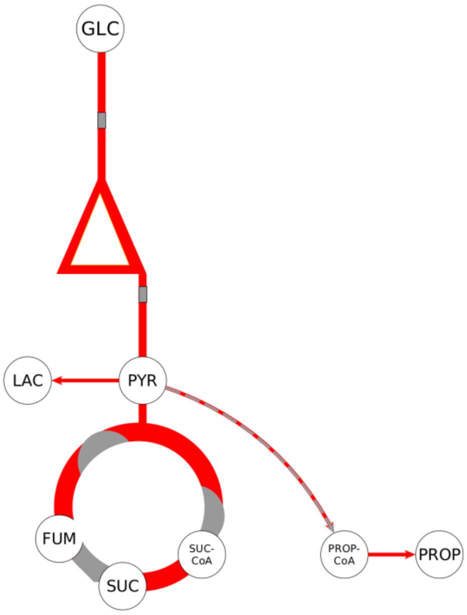
Bacteroides



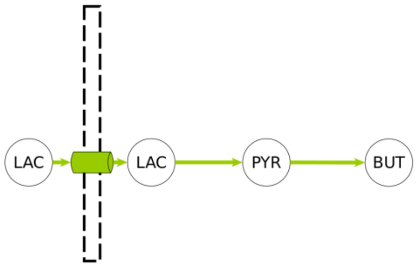
Clostridium



Bifidobacterium
Lactobacillus



Eubacterium hallii



Methylmalonyl-CoA mutase is involved in propionate biosynthesis, and our data showed that the whole pathway for propionate biosynthesis was, in fact, upregulated. The data further showed that many genes coding for proteins involved in iron scavenging were also upregulated (e.g. FecR). One of the genes coding for an enzyme with iron requirements was that encoding succinate dehydrogenase, which converts succinate into fumarate. This function, as well as all others in the TCA cycle, showed upregulation in all samples tested. The genus *Clostridium* also showed an upregulation of genes involved in Vitamin B12 production, but the biosynthesis occurred via glutamate, whereas in the *Bacteroides* group it was produced via succinate. The genes in the pathway for propionate production were overall upregulated (production via acetyl-CoA, not succinyl-CoA), similar to the genes in lactate and butyrate production pathways. The only other enzyme requiring Vitamin B12 in the microbiome was a multimer of propanediol dehydratase or glycerol dehydratase (ambiguous taxonomic assignment), which are both involved in the breakdown of glycerol/glycerone phosphate to propanol/propionate/1,3-propanediol. However, a full upregulation of either pathway was not observed. The *Bifidobacterium*/*Lactobacillus* group and the *Enterococcus* group showed upregulation of genes related to production of lactate from pyruvate, and the *Bifidobacterium*/*Lactobacillus* group also showed upregulation of genes encoding proteins involved in butyrate production, but it is unclear if butyrate would be directly produced from pyruvate, or derived from external acetate. *Eubacterium hallii*, on the other hand, showed high activity related to converting lactate into butyrate, as also shown previously [399]. In addition, our data indicated that formate was produced by the *Enterococcus* and *Bacteroides* populations.

Microbial groups directly involved in the degradation of the IMMPs

In order to gain insight into which bacterial groups are directly involved in degradation of different IMMPs, we used the KEGG reference pathway for starch and sucrose metabolism [227]. We surveyed our data for the expression of the genes encoding enzymes that are known to be involved in sucrose and starch metabolism. More specifically we focused on genes encoding enzymes from glycoside hydrolase family 13 (http://www.cazy.org/GH13_bacteria.html), as this family includes a number of bacterial proteins shown to be essential in degradation of similar compounds, such as isomaltooligosaccharides (IMO) [400]. The majority of genes listed in the KEGG starch and sucrose metabolism pathway were detected in our transcriptome data (Figure S4), as well as some additional genes in glycoside hydrolase family 13 (EC 3.2.1.135, 3.2.1.68 and 3.2.1.11), which were not listed in the KEGG pathway, but which are known to be activated during the degradation of pullulan and dextran [401-404]. It is interesting to note that the relative contribution of these starch and sucrose metabolism genes to the total number of genes from each sample did not correlate with the presence or absence of IMMPs in the samples. The only exception was incubation with pre-treated IMMP-27, in which starch and sucrose metabolism genes reached 10% at 12 h and about 12% at 48 h, whereas in other groups they ranged between 4 to 5% (Figure S5). Despite of the similarities in the overall expression of the starch and sucrose metabolism genes in all samples, we could see differences in the relative abundance of genes coding for specific enzymes depending on the IMMP used, and the duration of the fermentation (Figure S6).

One of the aims of this study was to better understand the functional dynamics of the bacterial communities during IMMP degradation. Previously reported HPAEC and HPSEC analyses [381] showed that the degradation of IMMP-94 and IMMP-96 occurred between 12 h and 24 h of the incubation. At 24 h and 48 h we noted an increase in the expression of genes coding for enzymes that might be directly involved in the hydrolysis of α-(1→6) glycosidic linkages, namely EC 3.2.1.10 – oligo-1,6-glucosidase, EC 3.2.1.11 – dextranase, and EC 3.2.1.33 – amylo-α-1,6- glucosidase (Figure S7a,b). There was also an increase in the expression of genes coding for enzymes that can hydrolyse α-(1→4) glycosidic linkages, mainly the EC 3.2.1.1 – α-amylase, EC 3.2.1.20 – α–glucosidase 4-α-glucanotransferase, and EC 2.4.1.25 – 4-α-glucanotransferase. Since IMMP-27 contains lower amounts of α-(1→6) linkages, its degradation also involves the activation of the same genes, however, the expression levels of the genes encoding enzymes which hydrolyse α-(1→6) linkages were much lower (Figure S7a,b). Bacterial groups that contributed the most to the primary degradation of IMMP's α-(1→6) linkages were *Lactobacillus*, *Bifidobacterium* and *Bacteroides*, all expressing the genes encoding EC 3.2.1.10 oligo-1,6-glucosidase and EC 3.2.1.11 dextranase. On the other hand, the metatranscriptomic data suggests that α-(1→4) linkages were hydrolysed mainly by *Bacteroides*, unclassified Bacteroidales, unclassified Enterobacteriaceae, *Lactobacillus* and *Bifidobacterium* via EC 3.2.1.1 alpha-amylase and EC 2.4.1.1 glycogen/amylophosphorylase (Figure S8). Based on the transcript data, *Bifidobacterium* and *Lactobacillus* were mainly active in the degradation of IMMP-94 and IMMP-96 at 24 h (Figure 4, and Figure S8). These genera were also active in degradation of IMMP-27 and the pre-treated IMMP-27, but their relative contributions were much lower (Figure 4, and Figure S8). The breakdown of IMMPs at 24 h and 48 h was otherwise dominated by *Bacteroides*, with the exception of pre-treated IMMP-27 at 48 h, which showed a high level of expression of genes assigned to unclassified Enterobacteriaceae. Figure 4 summarises our model of IMMP degradation and confirms the specialised role of lactobacilli and bifidobacteria in hydrolysis of α-(1→6) linkages. It also reveals the important contribution of *Bacteroides* as both, primary and secondary degraders of IMMPs and their by-products.

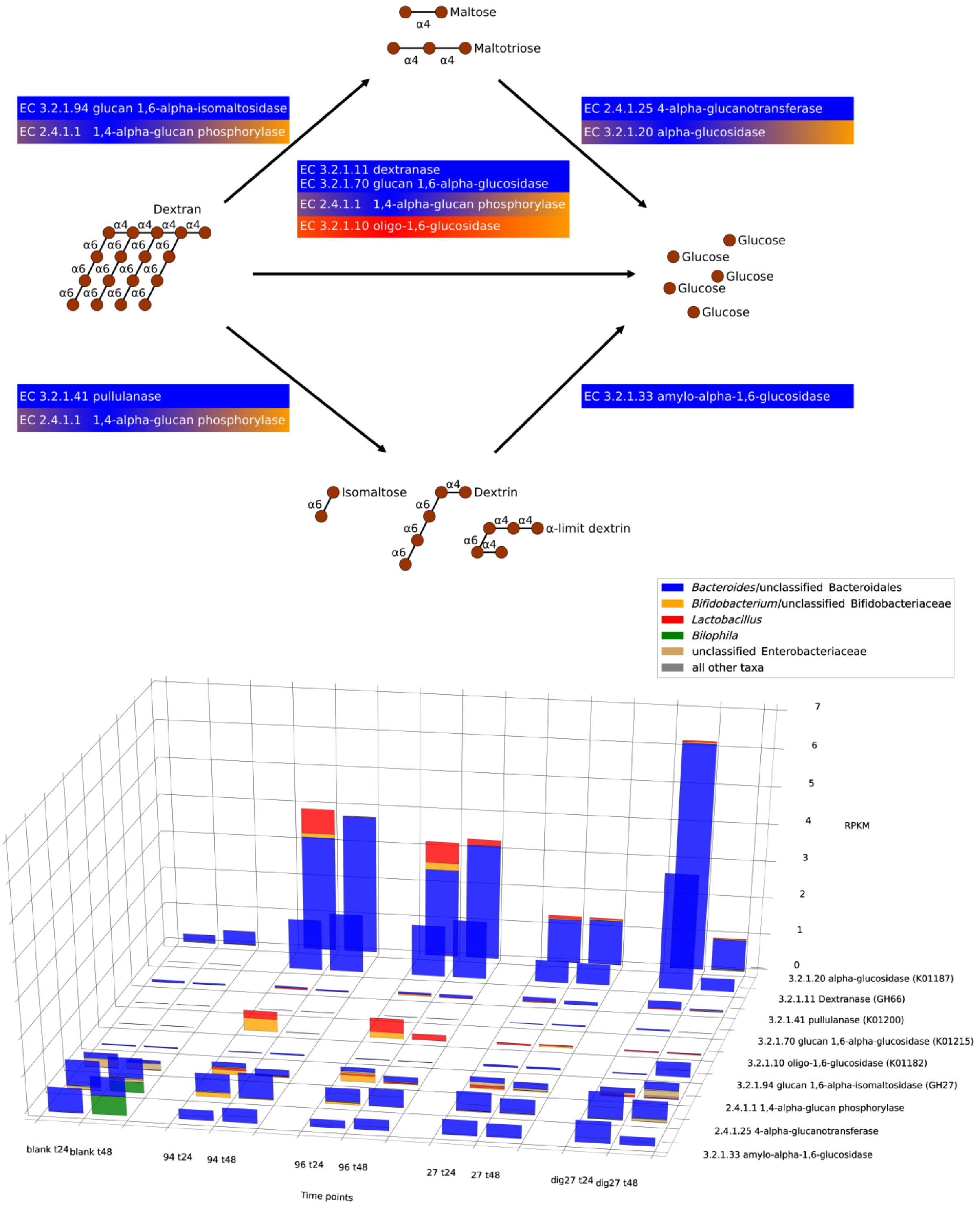


Figure 4: Overview over the main degradation pathways starting from dextran. Colours in the top panel indicate the main contributors to a reaction. The bottom panel shows the overall expression in reads per kilobase per million (RPKM) per organism at time points 24 and 48h for all conditions. If an enzyme could not be identified by its associated EC number, then the KO or CAZy identifier used for identification are given in brackets.

Discussion

Prebiotic food components should be resistant to host's gastric enzymes, fermentable by the host's intestinal microbiota and capable of promoting growth and activity of bacterial groups associated with health [405]. The IMMPs seem to fulfil all these criteria [281, 406, 407]. Earlier studies demonstrated that hydrogenated and high DP IMMPs are not or little-digestible by rat gastric enzymes [407], and that diets containing IMMPs are associated with higher numbers of lactobacilli, and an overall increase in the number of intestinal bacteria [408]. Moreover, a recent study with human inoculum reported that IMMPs can be fermented by human large intestinal microbiota and that SCFAs, in particular acetate and propionate, are produced, indicating that IMMPs may stimulate activity of probiotic groups [378]. This is in accordance with earlier findings from a small human trial that showed an increased level of bifidobacteria in subjects who received IMOs in their diets [406].

In our study, we confirmed the prebiotic character of the IMMPs and showed that the specific effect of different IMMPs on human faecal microbiota composition and activity varied during *in vitro* fermentation, depending on the relative amount of α-(1→6) glycosidic linkages present in the substrate. When IMMP-94 and IMMP-96 were used as a carbon source, we observed a strong upregulation of genes in the probiotic cluster, specifically genes assigned to bifidobacteria and lactobacilli. Furthermore, high relative activity of these bacteria corresponded with an increase in their relative abundance as estimated by rRNA gene sequencing [381]. In contrast, when the pre-treated IMMP-27 was used as a substrate, the relative activity of bifidobacteria and lactobacilli was lower, these bacteria were less active in the control, and their activity peak in the presence of IMMP-27 was delayed to 48 h. Interestingly, all IMMP treatment groups showed a time lag between the maximum relative activity, and the increase in the corresponding bacterial relative abundance as measured by rRNA gene-targeted community analysis [381]. For example, the maximum activity of bifidobacteria was observed at 24 h of incubation when IMMP-94 and IMMP-96 were used as substrates. Yet, bifidobacteria reached their highest relative abundance only at 48 h when their relative activity had already decreased. The relative activity of lactobacilli followed a pattern similar to that of bifidobacteria in all treatment groups, except for incubations with IMMP-94 where lactobacilli showed maximum relative activity at 12 h, whereas bifidobacteria activity peaked at 24 h. Relative activity of *Bacteroides* was very high in all groups, regardless of the incubation time, presence and type of the IMMP that was used as a carbon source. *Bacteroides* spp. are known to be generalists that are able to break down a wide array of carbon sources [409]. Bifidobacteria and lactobacilli are often more specialised and can grow on substrates that are chemically not accessible to other bacteria in the microbial ecosystem. This may be the reason that these groups show delayed activity in relation to *Bacteroides*, as only after the depletion of the easily accessible IMMP fractions containing the α-(1→4) glycosidic linkages the bacterial groups capable of utilising the α-(1→6) glycosidic linkages gained a competitive advantage. It is known that bacteria can sense specific polysaccharides and produce specific sets of enzymes according to their individual nutrient prioritization schemes [71]. The patterns of activity and growth in the presence of IMMP-27, pre-treated IMMP-27 and in the control group may confirm this hypothesis, as we observed increased relative abundance of *Parabacteroides*, *Sutterella, Parasutterella*, *Enterococcus*, unclassified Lachnospiraceae *Incertae Sedis*, *Eggerthella* and few other groups [381]. High relative abundance of these groups could be explained by the presence of residual α-(1→4) glycosidic linkages, or the presence of products generated during the enzymatic conversion of the α-(1→4) glycosidic linkages into α-(1→6) glycosidic linkages during the IMMP pre-treatment process, or by more efficient scavenging on other bacteria or their metabolites.

While some of the beneficial bacteria increased in relative abundance and activity with the presence of IMMPs, we also noted that the exclusive use of these prebiotics put a selective pressure on other beneficial microbes. For example, *Lactococcus lactis* and *Ruminococcus bromii* - two specialized beneficial degraders, did not show any survival in our samples [381]. This can be explained by the lack of suitable substrate for both species, given that neither any simple mono- or disaccharides (for *Lactococcus* [410]) nor type II or III resistant starch (for *Ruminococcus* [411]) were present in this experiment. Although both organisms can be considered a probiotic, they were not stimulated in the particular prebiotic environment tested here, enforcing the notion that prebiotics can selectively stimulate activity and growth of specific groups, whereas in general, a diverse diet may be necessary to comprehensively support a stable community of commensal microbes.

In our study we also observed a clear effect of having no carbohydrate source in the control samples. With the absence of the prebiotics, there was a switch of the community from processing carbohydrates to utilising amino acids [412], as indicated by the increase of relative abundance of *Acidaminococcus* [381]. In addition, there was an increase of *Bilophila* in the control samples, which is an organism previously associated with gut dysbiosis [413].

IMMP Degradation Model

A total of 130 families of glycoside hydrolases, 22 families of polysaccharide lyases, and 16 families of carbohydrate esterases have been described, and many of these enzymes are encoded only by the genomes of microbes (www.cazy.org) [195]. We surveyed our data for the presence of genes encoding the enzymes that are known to be involved in sucrose and starch metabolism, mostly genes from glycoside hydrolase family 13. Few studies up to date looked at the genetics and enzymology of degradation of IMMPs mainly in lactobacilli [402, 414, 415], bifidobacteria [416] and *Bacteroides*. However, microbial species in the gut do not act in isolation, but rather interact with each other through a network of syntrophic interactions often making the utilization of the substrate more effective [417]. Metabolic potential and fermentation efficiency vary between different species, and complete degradation of IMMPs in the gut is a result of different bacterial groups working together in a complementary fashion, likely leading to the formation of microbial food chains [417, 418]. Certain bacterial groups may show a higher activity at specific degradation steps, as measured by the expression of specific genes coding for enzymes required to catalyse given reactions. This is also visible in our experiments. The expression of oligo-1-6-glucosidase encoding genes was dominated by lactobacilli and bifidobacteria when IMMP-94 and IMMP-96 were used as a substrate, whereas *Bacteroides* and unclassified Bacteroidales were also highly active in the presence of IMMP-27 or IMMP-dig27. Similar patterns could be observed in expression of other genes that code for enzymes involved in sucrose and starch metabolism (Suppl. Figure S8). While some of the carbohydrate breakdown steps were dominated by known probiotic genera, many of the primary and secondary degradation processes were also performed by members of *Bacteroides*. Our data showed that once the fermentation started, one of the very specialized enzymes, dextranase, was produced only by *Bacteroides*. Other processes were found reliant on multiple genera as based on the gene expression data. For example, the breakdown of the IMMP27 and IMMP-dig27 to maltose and maltotriose by ɑ –amylases was dominated by *Bacteroides*, whereas the further metabolisation was performed also by bifidobacteria and lactobacilli. Furthermore, other groups such as enterobacteria or *Parabacteroides* were not involved in most of these breakdown processes, but still constituted viable populations in the communities. Their functional role in the community is, however, not clear.

Metabolites of fermentation

Experimental results showed that the administration of IMMPs lead to an increased production of different SCFAs, mainly acetate and succinate [381]. While succinate normally does not accumulate in this medium [60], the excess of substrate [419], high CO_2 levels, and the upregulation of all the necessary steps [61] in our metabolic mapping, including the necessity for iron, could explain such accumulation. In addition, previous studies showed that succinate accumulation is associated with oversupply of complex substrates [286], such as prebiotics, or in our case IMMPs or when further metabolisation of succinate is unnecessary [60]. It is also possible that lack of Vitamin B12, which is necessary for propionate production [61], and for which an upregulation could be observed, resulted in the accumulation of succinate instead of propionate. However, we are unable to conclude the exact reason based on our data. One of the other propionate production pathways, the acrylate pathway [287], could not be detected in the data. However, it is tempting to speculate that the production of propionate proceeded via the direct fermentation of pyruvate via 3-hydroxypropionate and acryloyl-CoA in the currently studied fermentation. This pathway has not been described before, but it is potentially visible in the data, with just a few reactions missing. Furthermore, the potential of producing propionate via 1,2-propanediol directly through methylglyoxal is indicated in the data. Unfortunately, no definite conclusions can be drawn due to missing steps in the metabolism of the involved populations (*Bifidobacterium/Lactobacillus* for the former, and *Clostridium* for both), however, the possibility of these alternative pathways should be investigated. Besides succinate, propionate and acetate, also lactate and butyrate were observed as metabolites [381].

Dietary fibres, including modified starches such as IMMPs offer a promising, non-invasive way to intentionally manipulate gut microbiota composition. Investigations of whole bacterial communities and understanding of the mechanisms by which microorganisms interact to degrade different dietary carbohydrates are essential for our ability to manipulate gut microbiota to benefit our health. We showed how IMMPs can increase the relative abundance and activity of beneficial bacteria, making these novel prebiotics potentially useful in improving host’s health from the aspect of nutrition, to achieve prevention or even alleviation of diseases.

Acknowledgements

The authors thank Jasper Koehorst (Wageningen University, Laboratory of Systems and Synthetic Biology) for his help with the transcriptome annotation, and Maria Suarez-Diez and Edoardo Saccenti (Wageningen University, Laboratory of Systems and Synthetic Biology) for helpful discussions. The authors want to thank Bartholomeus van den Bogert (Wageningen University, Laboratory of Microbiology) for the design of Figure S1.

This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

Funding

This research was performed in the public-private partnership CarboHealth coordinated by the Carbohydrate Competence Center (CCC, www.cccresearch.nl) and financed by participating partners and allowances of the TKI Agri&Food program, Ministry of Economic Affairs. This work was furthermore supported by Wageningen University and the Wageningen Institute for Environment and Climate Research (WIMEK) to BH through the IP/OP program Systems Biology [project KB-17-003.02-023].

Supplementary information

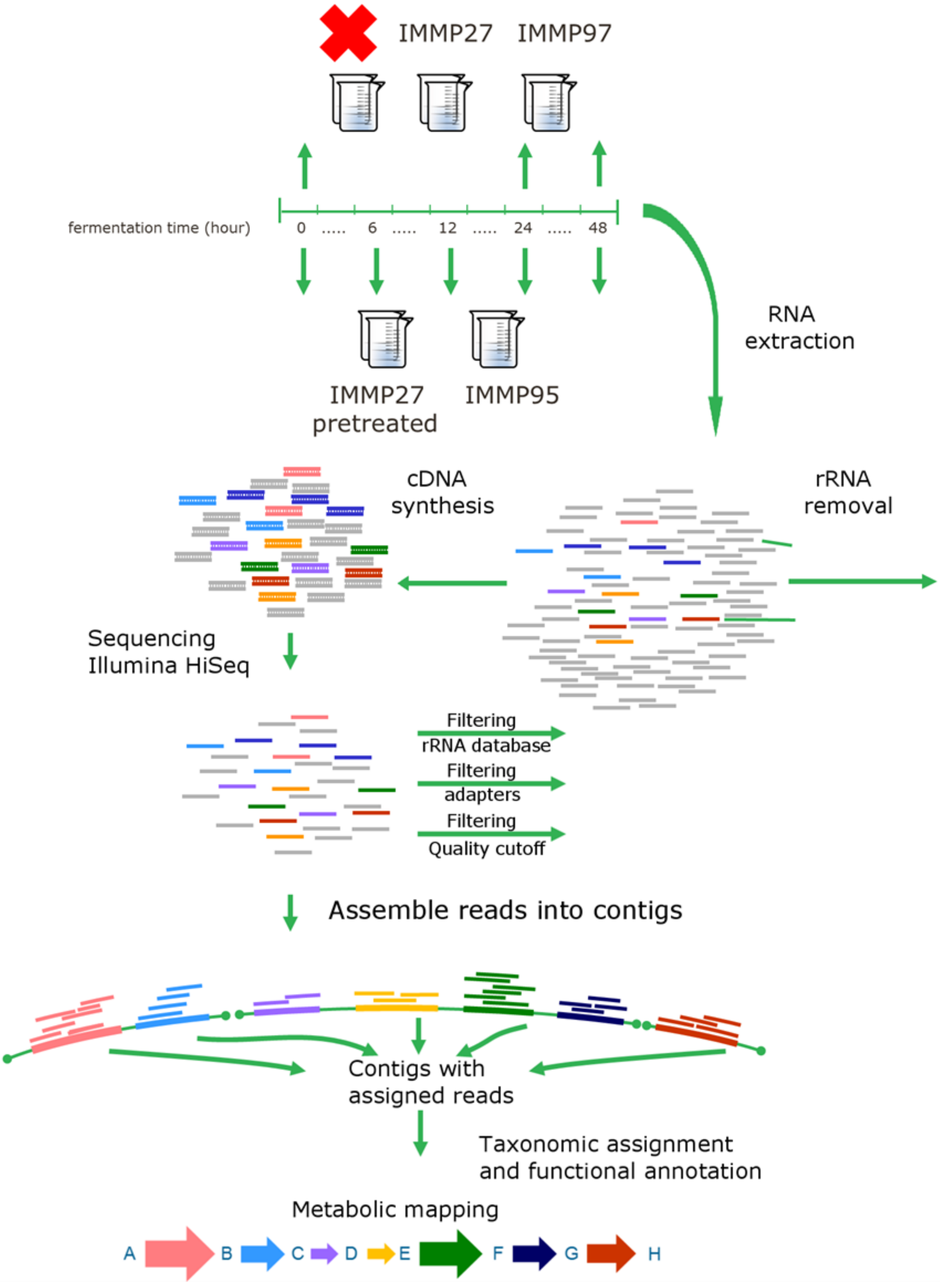


Figure S1: Figure S1: Experimental design.

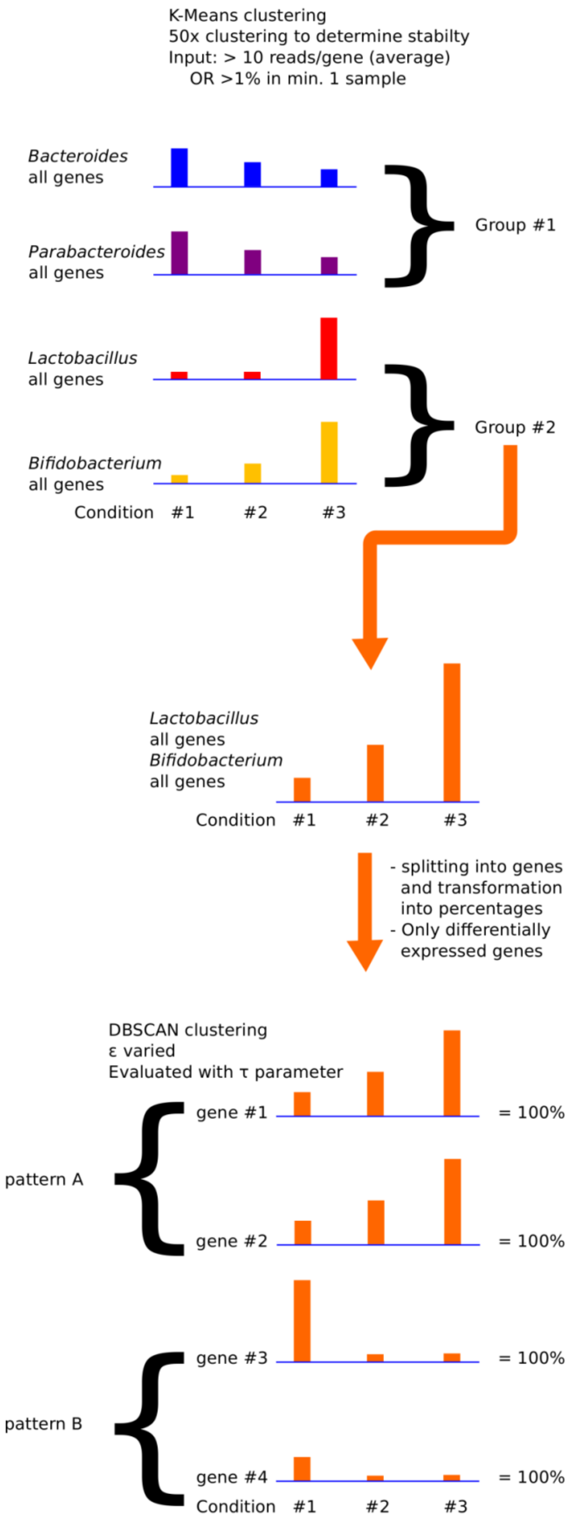


Figure S2: Overview over the clustering procedure. First, expression was lumped at the genus level. On the accumulated expression data k-means clustering was performed, until a stable clustering was achieved. The genes of the grouped genera were afterwards subjected to DBSCAN clustering. The stability of the clustering was evaluated with the Tau-parameter. Only genes, which were at least once differentially expressed, were used in the clustering process to reduce the noise.

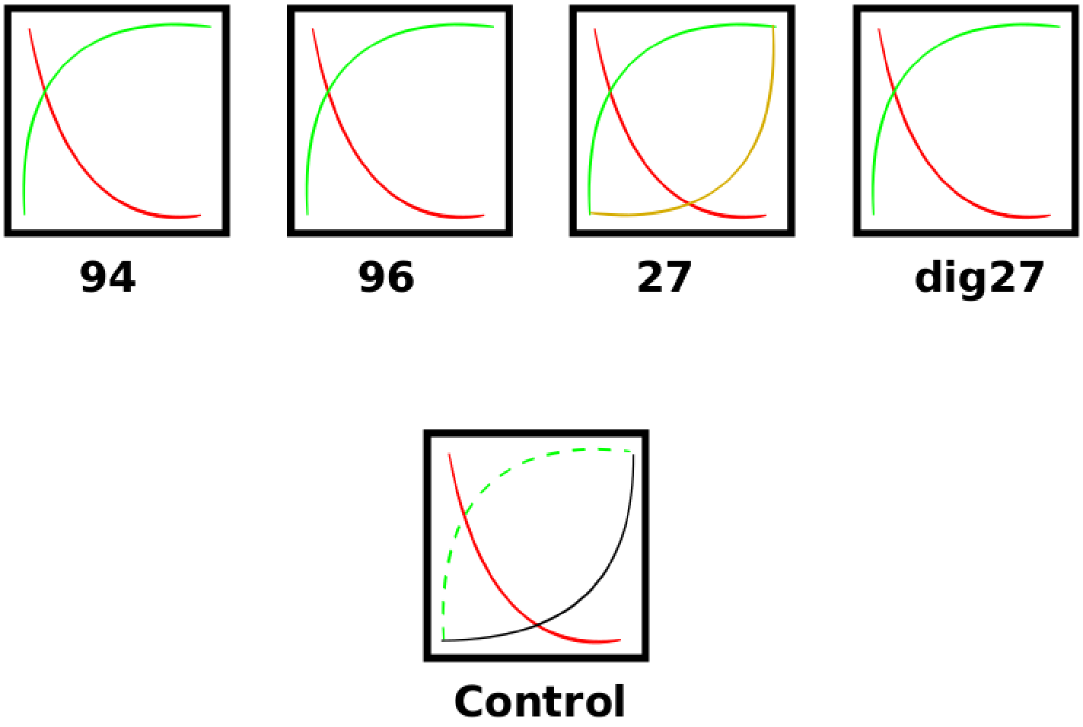


Figure S3: Overview of the main gene expression patterns. All groups (*Bacteroides, E.coli, Lactobacillus/Bifidobacterium, Enterococcus*; besides *Eubacterium hallii*) showed in all prebiotic conditions increase in relative transcript abundance in roughly the same proportion (green). Some groups (*Bacteroides, Escherichia*) also showed comparable increase in expression in the control condition (dotted green line). Furthermore, all groups showed a downregulation of certain genes in all conditions (red), and an upregulation of a group of genes in the control condition (black). *Eubacterium hallii* showed only increase in transcript abundance at the last time point with the prebiotic IMMP-27 (yellow).

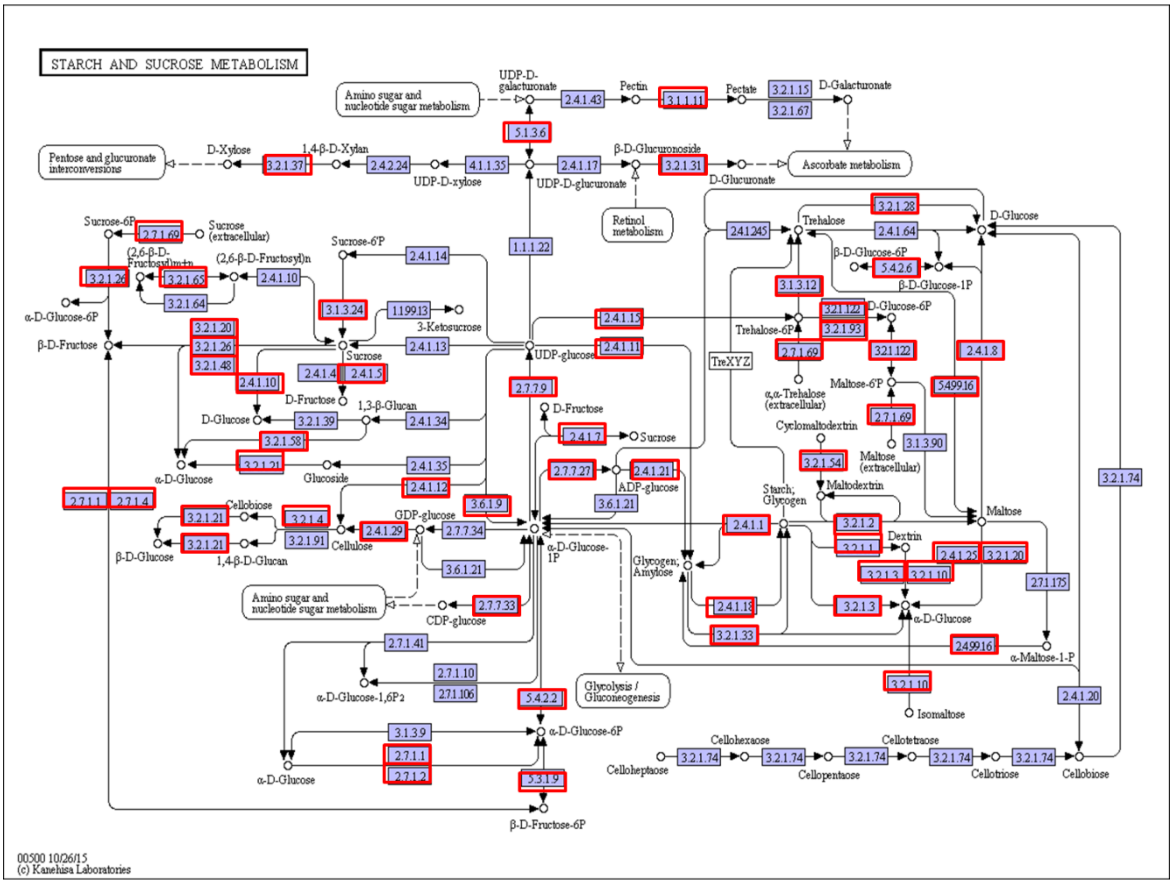


Figure S4: Starch and sucrose metabolism enzymes detected in the data.

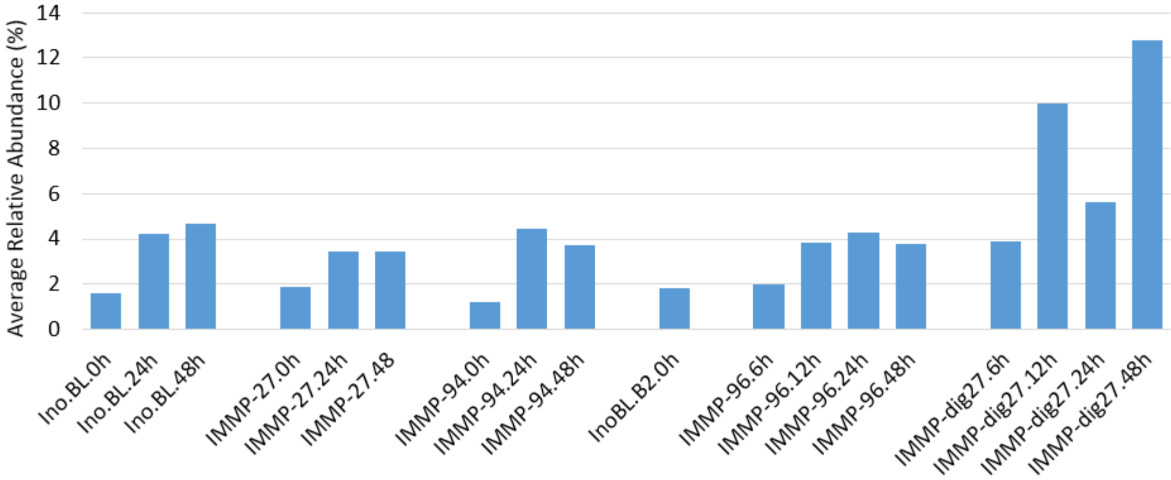


Figure S5: Relative abundance (percentage) of starch and sucrose metabolism enzyme encoding genes detected in the metatranscriptome data.

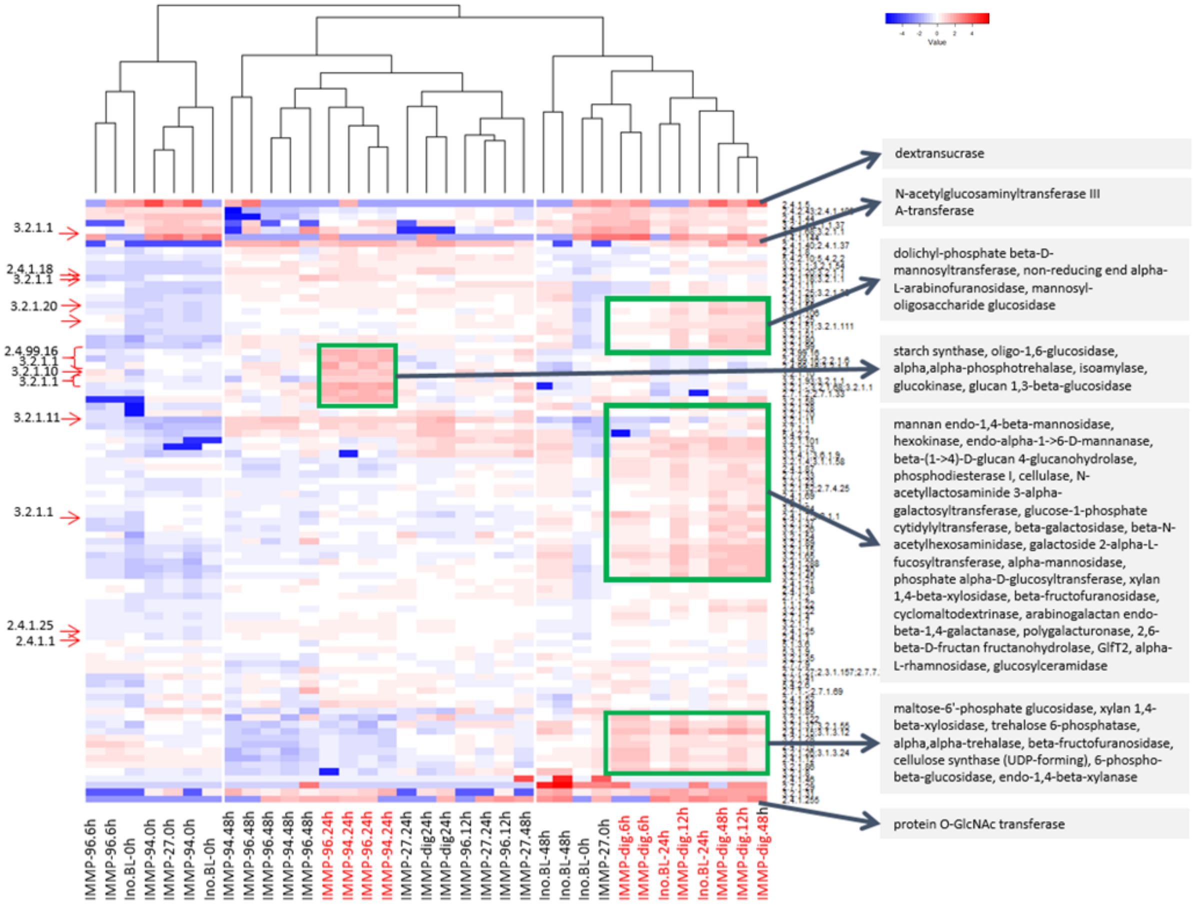


Figure S6: Heatmap of log10 transformed relative abundances of expressed genes detected in our data coding for starch and sucrose metabolism enzymes. Samples clustered based on the similarities between the up and down regulated genes. The red arrows indicate selected genes that code for enzymes described in our IMMP degradation model. Green boxes highlight the gene upregulation patterns for different IMMPs at various incubation times.

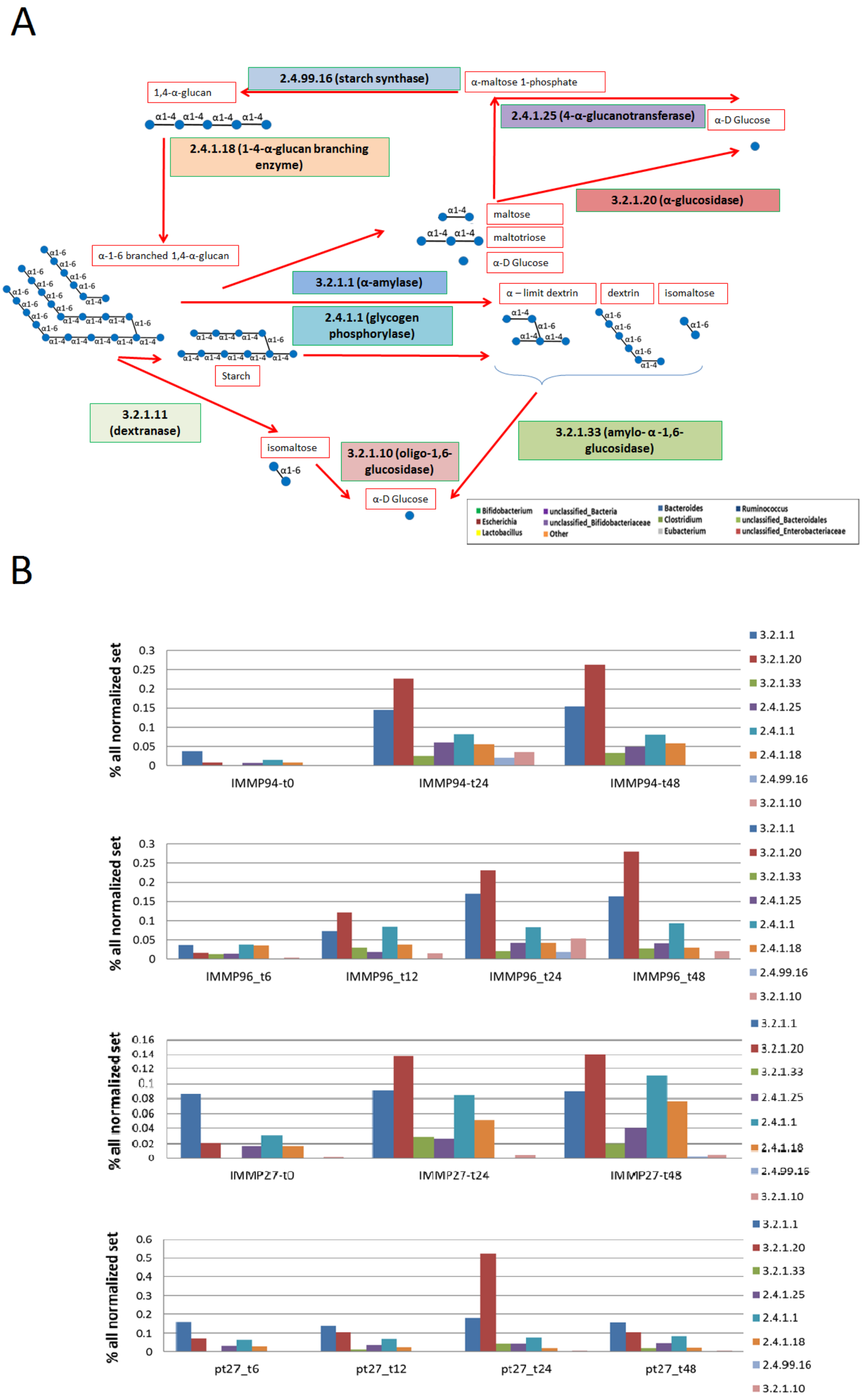
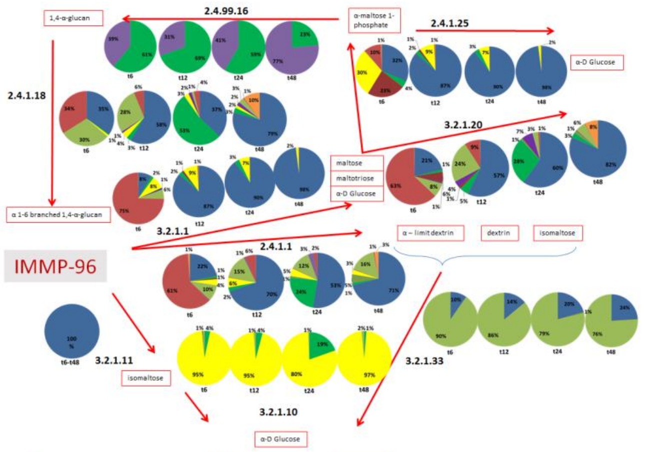
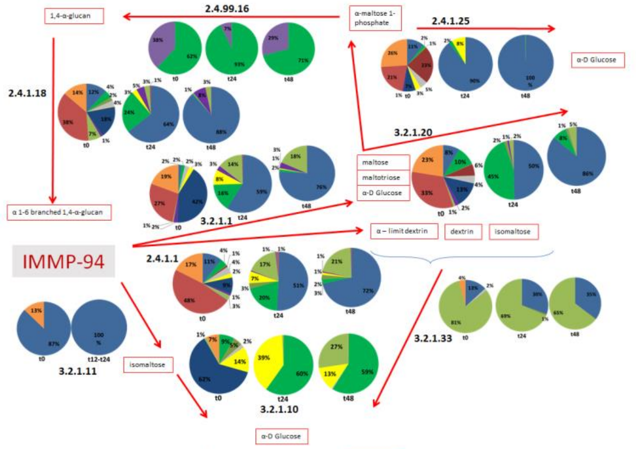
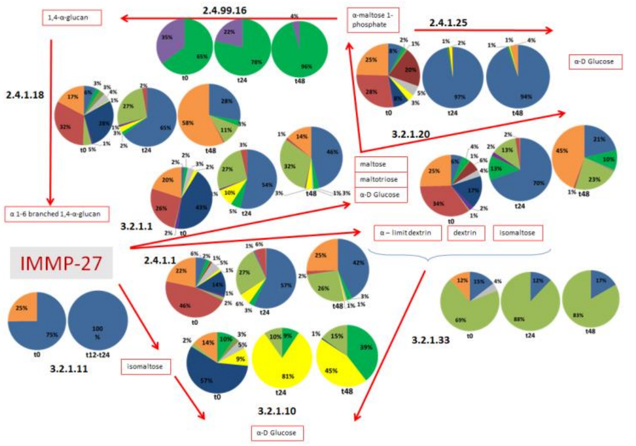


Figure S7: IMMP degradation model. a. main enzymes involved in the pathway, b. relative abundance of transcripts of genes coding for enzymes needed for IMMP degradation.


1,4-α-glucan
2.4.99.16
α-maltose 1-phosphate
2.4.1.25
α-D Glucose
2.4.1.18
3.2.1.20
maltose
maltotriose
α-D Glucose
α 1-6 branched 1,4-α-glucan
3.2.1.1
2.4.1.1
α – limit dextrin
dextrin
isomaltose
IMMP-96
3.2.1.11
isomaltose
3.2.1.33
3.2.1.10
α-D Glucose
t6
t12
t24
t48
t6-t48


1,4-α-glucan
2.4.99.16
α-maltose 1-phosphate
2.4.1.25
α-D Glucose
2.4.1.18
3.2.1.20
maltose
maltotriose
α-D Glucose
α 1-6 branched 1,4-α-glucan
3.2.1.1
α – limit dextrin
dextrin
isomaltose
IMMP-94
2.4.1.1
3.2.1.11
isomaltose
3.2.1.33
3.2.1.10
α-D Glucose
t0
t24
t48
t12-t24


1,4-α-glucan
2.4.99.16
α-maltose 1-phosphate
2.4.1.25
α-D Glucose
2.4.1.18
3.2.1.20
maltose
maltotriose
α-D Glucose
α 1-6 branched 1,4-α-glucan
3.2.1.1
α – limit dextrin
dextrin
isomaltose
IMMP-27
2.4.1.1
3.2.1.11
isomaltose
3.2.1.33
3.2.1.10
α-D Glucose
t0
t24
t48
t12-t24

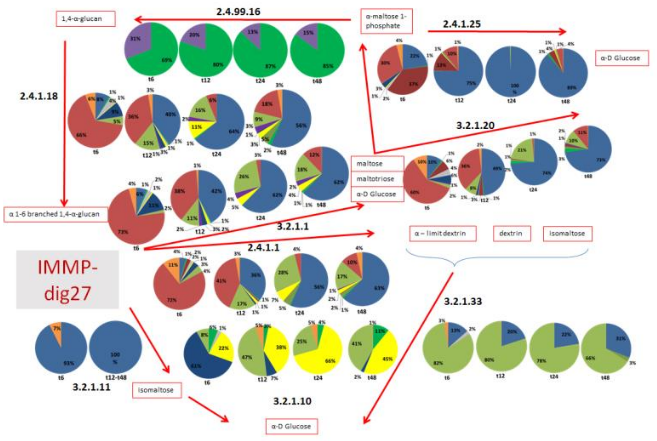

1,4-α-glucan
2.4.99.16
α-maltose 1-phosphate
2.4.1.25
α-D Glucose
2.4.1.18
3.2.1.20
maltose
maltotriose
α-D Glucose
α 1-6 branched 1,4-α-glucan
3.2.1.1
α – limit dextrin
dextrin
isomaltose
IMMP-dig27
2.4.1.1
3.2.1.11
isomaltose
3.2.1.33
3.2.1.10
α-D Glucose
t6
t12
t24
t48
t12-t48

Figure S8: Relative contribution of different bacterial groups to expression of genes coding for the enzymes in the IMMP degradation pathway. Colour coding as in Figure S7.

Table S1. Overview over the RNA-seq metrics

Condition	total reads	rRNA	% rRNA	non rRNA	trimmed bases due to adapters	% of bases trimmed due to adapters	sequences passing prinseq quality filtering	% passing prinseq quality filtering	mean length	Total % of bases passing ALL filtering steps	Mapping rate in % to assembly
Blank, repl. 1, t0	17182356	388147	2,26	16794209	713631983	28,14	15120013	90,03	108,34	63,56	79,06
Blank, repl. 2, t0	12843968	234407	1,83	12609561	620076640	32,57	11339128	89,92	101,95	60,00	75,99
IMMP-27, repl. 1, t0	18661592	402676	2,16	18258916	844672334	30,64	16454635	90,12	104,71	61,55	71,29
IMMP-27, repl. 2, t0	21152485	438493	2,07	20713992	881059209	28,17	18728000	90,41	108,06	63,78	74,42
IMMP-94, repl. 1, t0	30405866	555904	1,83	29849962	1087733238	24,13	26908763	90,15	113,98	67,25	79,06
IMMP-94, repl. 2, t0	19354896	777385	4,02	18577511	588584413	20,98	16575459	89,22	119,53	68,24	80,79
Blank, repl. 1, t24	27195831	243329	0,89	26952502	917731871	22,55	24835747	92,15	114,96	69,99	81,28
Blank, repl. 2, t24	26279510	263802	1,00	26015708	1199020108	30,52	23724640	91,19	103,99	62,59	81,43
IMMP-27, repl. 1, t24	12540536	1832716	14,61	10707820	587094636	36,31	9467522	88,42	93,59	47,10	85,32
IMMP-27, repl. 2, t24	23402662	2094144	8,95	21308518	775851341	24,11	18915069	88,77	111,03	59,83	86,95
IMMP-94, repl. 1, t24	14390413	901227	6,26	13489186	729841441	35,83	12186486	90,34	96,64	54,56	85,41
IMMP-94, repl. 2, t24	27352727	1655263	6,05	25697464	1235521747	31,84	23084321	89,83	102,74	57,80	84,52
Blank, repl. 1, t48	22532893	822320	3,65	21710573	1368385387	41,74	19533106	89,97	88,47	51,13	79,64
Blank, repl. 2, t48	23432237	902055	3,85	22530182	1200860127	35,3	20328428	90,23	97,88	56,61	80,17
IMMP-27, repl. 1, t48	24004294	4361225	18,17	19643069	1175519132	39,63	16574957	84,38	92,59	42,62	81,24
IMMP-27, repl. 2, t48	21275156	8177863	38,44	13097293	691799504	34,98	11702497	89,35	98,7	36,19	80,22
IMMP-96, repl. 1, t48	28735621	11137710	38,76	17597911	1296700269	48,8	14765189	83,9	81,2	27,82	83,92
IMMP-96, repl. 2, t48	35679453	11638379	32,62	24041074	1511937150	41,65	20304145	84,46	92,59	35,13	85,36
IMMP-96, repl. 1, t6	18066769	185180	1,03	17881589	390649819	14,47	15736624	88	126,02	73,18	75,76
IMMP-96, repl. 2, t6	20993700	55456	0,26	20938244	412525567	13,05	18327046	87,53	128,66	74,88	71,55
IMMP-96, repl. 1, t12	35944607	70088	0,20	35874519	1270799851	23,46	31773426	88,57	113,35	66,80	76,36
IMMP-96, repl. 2, t12	22773537	64983	0,29	22708554	964291453	28,12	20307424	89,43	106,59	63,37	76,79
IMMP-96, repl. 1, t24	17241186	1380292	8,01	15860894	385783659	16,11	13984773	88,17	122,24	66,10	87,1
IMMP-96, repl. 2, t24	19319950	1502708	7,78	17817242	607265419	22,57	15673285	87,97	111,85	60,49	86,3
IMMP-96, repl. 1, t48	34060532	145402	0,43	33915130	1420500568	27,74	30260173	89,22	107,64	63,75	75,81
IMMP-96, repl. 2, t48	31992519	119756	0,37	31872763	1450909006	30,15	28116504	88,21	105,1	61,58	72,91
Blank, repl. 1, t0	14780189	291372	1,97	14488817	458214833	20,94	12514201	86,37	111,32	62,84	84,1
IMMP-dig27, repl. 1, t6	21177851	113633	0,54	21064218	1266290906	39,81	17927736	85,11	89,44	50,48	81,34
IMMP-dig27, repl. 2, t6	14730499	91528	0,62	14638971	464265511	21	12827318	87,62	113,79	66,06	81,04
IMMP-dig27, repl. 1, t12	19229982	120092	0,62	19109890	353792739	12,26	16872478	88,29	128,38	75,09	81,28
IMMP-dig27, repl. 2, t12	19183847	184071	0,96	18999776	449569556	15,67	16837679	88,62	123,83	72,46	81,79
IMMP-dig27, repl. 1, t24	19000262	346531	1,82	18653731	573909821	20,38	16612325	89,06	117,09	68,25	82,71
IMMP-dig27, repl. 2, t24	18981915	260033	1,37	18721882	214922072	7,6	16320453	87,17	135,84	77,86	83,29
IMMP-dig27, repl. 1, t48	19947312	1144473	5,74	18802839	508216740	17,9	16742425	89,04	120,37	67,35	88,6
IMMP-dig27, repl. 2, t48	11159596	475984	4,27	10683612	160119185	9,93	9455765	88,51	132,49	74,84	89,42
Average	21366411,23	1514013,714	6,33	19852397,51	801840435,8	25,74	17591935,06	85,99	106,19	60,89	78,66
Total	747824393	52990480		694833913	28064415252		615717727				

Table S2. Assignment of genus-level taxa per cluster, showing the amount of assigned genes and differentially expressed genes over all conditions

Organism	**Genes assigned**	**Genes differentially expressed**
Cluster 1		
Ruminococcus	9904	6220
Lactococcus	8211	5263
unclassified_Gammaproteobacteria	373	285
Bos	329	263
N/A	208	85
unclassified_Mammalia	198	67
unclassified_Bovidae	107	41
Clostridiales	4	0
Bacteria	3	1
Eukaryota	2	1
Gammaproteobacteria	1	1
Viruses	1	0
Cluster 2		
Lactobacillus	8655	7279
Bifidobacterium	7817	5849
unclassified_Actinobacteria	265	195
unclassified_Bifidobacteriaceae	116	98
Fusobacterium	70	47
unclassified_Lactobacillaceae	52	38
Myoviridae	36	33
Cluster 3		
Enterococcus	2580	1927
unclassified_Lactobacillales	1154	1012
unclassified_Bacilli	537	473
unclassified_Enterococcaceae	139	136
Cluster 4		
unclassified_Enterobacteriaceae	11350	9255
unclassified_Bacteria	6997	4494
Eubacterium	4884	3849
Escherichia	2972	2386
unclassified_Proteobacteria	683	530
Salmonella	59	47
Shigella	51	9
Enterobacter	36	29
Citrobacter	31	14
Vibrio	26	9
Cluster 5		
Bacteroides	53749	49101
unclassified_Bacteroidales	10243	9067
Parabacteroides	1749	919
Prevotella	302	233

Bacteria	193	133
Desulfosporosinus	32	11
Flavobacterium	30	27
Cluster 6		
Clostridium	4244	3228
unclassified_Clostridia	114	69
Cluster 7		
unclassified_Clostridiales	10819	3356
unclassified_Lachnospiraceae	1030	219
Anaerostipes	127	6
Clostridiales	39	5
Cluster 8		
N/A	8225	4083
Sutterella	3598	2645
unclassified_Bacteroidetes	770	681
unclassified_Betaproteobacteria	135	77
Odoribacter	94	76
Ethanoligenens	28	21
Corynebacterium	20	8
Cluster 9		
Bilophila	3853	2614
unclassified_Firmicutes	3152	1898
Phascolarctobacterium	1328	8
unclassified_Selenomonadales	244	2
Acidaminococcus	174	4
unclassified_Acidaminococcaceae	117	0
Selenomonas	109	19
Veillonella	88	8
Megamonas	64	12
unclassified_Veillonellaceae	56	3
Pelosinus	53	4
Megasphaera	53	9
Desulfitobacterium	51	15
Anaeromusa	33	0
Acetonema	32	3
Mitsuokella	20	5

Text mining

Further EC numbers were derived by text mining and matching all InterProScan derived domain names against the BRENDA database (download 13.06.13) [387]. The text mining algorithm included lower casing all characters, removal of non-alphanumerical characters (colons, commas, brackets, apostrophes, dashes, terminal points), removal of partial and generic terms (type, terminal, subunit, domain, enzyme, like, hypothetical, conserved, operon, active site, enzyme, probably, central, 51 kd, respiratory chain, c terminal, n terminal), rejection of overly generic final result terms (kinase, cytochrome, protein, methyltransferase) and reduction of certain terms (deletion of PEP/pyruvate binding; removal of "prokaryotic" in "prokaryotic cytidylate kinase"; "family" in "cytidilate kinase family"; "phosphorylating" in "glyceraldehyde phosphate dehydrogenase phosphorylating"; "iron containing" in "iron containing alcohol dehydrogenase"; "zinc containing" in "zinc containing alcohol dehydrogenase"; "manganese containing" in "manganese containing catalase"; "20 kd" in "nadh ubiquinone oxidoreductase 20 kd"; replacement of "carboxyltransferase" with "carboxylase" in "pyruvate carboxyltransferase"). Furthermore, all terms, which were only of length one, were also removed, in case the remaining name contained more than two words. On some domain names a manual curation was performed, and overly generic identifications (e.g. matching PF12847 "Methyltransferase domain" with e.g. EC 2.1.1.124 with alternative name "Protein Methyltransferase I") were rejected.

Chapter 5: The rumen metatranscriptome landscape reflects dietary adaptation and methanogenesis in lactating dairy cows

This chapter is adapted from:

Bastian Hornung*, Bartholomeus van den Bogert*, Mark Davids, Vitor A.P. Martins dos Santos, Caroline M. Plugge, Peter J. Schaap, Hauke Smidt, 2018. “The Rumen Metatranscriptome Llandscape Reflects Dietary Adaptation and Methanogenesis in Lactating Dairy Cows”. In bioRxiv. https://doi.org/10.1101/275883

*Contributed equally

Abstract

Methane eructed by ruminant animals is a main contributor to greenhouse gas emissions and is solely produced by members of the phylum *Euryarchaeota* within the domain *Archaea*. Methanogenesis depends on the availability of hydrogen, carbon dioxide, methanol and acetate produced, which are metabolic products of anaerobic microbial degradation of feed-derived fibers. Changing the feed composition of the ruminants has been proposed as a strategy to mitigate methanogenesis of the rumen microbiota.

We investigated the impact of corn silage enhanced diets on the rumen microbiota of rumen-fistulated dairy cows, with a special focus on carbohydrate breakdown and methanogenesis. Metatranscriptome analysis of rumen samples taken from animals fed corn silage enhanced diets revealed that genes involved in starch metabolism were significantly more expressed while archaeal genes involved in methanogenesis showed lower expression values. The nutritional intervention also influenced the cross-feeding between *Archaea* and *Bacteria*.

The results indicate that the ruminant diet is important in methanogenesis. The diet-induced changes resulted in a reduced methane emission. The metatranscriptomic analysis provided insights into key underlying mechanisms and opens the way for new rational methods to further reduce methane output of ruminant animals.

Introduction

Reduction of global greenhouse gas (GHG) output is necessary to prevent a further increase in global warming, which is predicted to result in multiple detrimental effects for the environment and human affairs [420]. The necessary measures are focused on the industrial and agricultural sectors in developed countries, with the aim to reduce carbon dioxide, methane and other GHG emissions. One of the predominant sources of methane emission, estimated to be as high as ~35% of the total anthropogenic methane emissions worldwide [63, 64], is the agricultural sector, and especially the eructation by ruminant animals [421].

Ruminal microbes play a pivotal role in the breakdown of animal feed and contribute between 35 to 50% of the animal's energy intake [94]. The ruminal microbial composition is complex, with diverse populations including bacteria, archaea, fungi, and protozoa. Their functional capacity is vast and has not yet been fully elucidated [422, 423].

Notwithstanding the ruminal microbial complexity, methane is solely produced by a few members of the phylum *Euryarcheota* belonging to the *Archaea* [36]. It has been shown that a change in diet can have a significant effect on the methane emissions of ruminants [424, 425], but the mechanisms that drive this change are not fully understood. The methanogenic archaea are not directly involved in the breakdown of the feed, but rely on their relationships with other community members that provide the necessary substrates for methanogenesis like hydrogen, formate and methanol.

Microbial ecology in cows and other ruminants has been investigated using 16S ribosomal RNA (rRNA) genes as molecular markers [426, 427], the sheep rumen microbial metatranscriptome has been investigated [428], and in cows specialized and general microbial functions have been examined [423, 429-434]. Understanding the mechanisms that influence cow rumen methanogenesis requires community-level analysis of active metabolic functions, however, a comprehensive analysis of diet-dependent effects on the functional landscape of the rumen microbiota is lacking. Here we investigated the effect of feed composition on bovine rumen activity patterns with a special focus on methane metabolism. By analysis of the rumen metatranscriptome landscapes in animals fed mixed grass silage (GS) and corn silage (CS) diets, we were able to elucidate the impact of the diet on the expression of methanogenic pathways and on the relationships of methanogens with other community members.

Materials and Methods

Study design and sampling

The study design has been described in detail by Van Gastelen *et al*. [424]. Briefly, the experiment was performed in a complete randomized block design with four dietary treatments and 32 multiparous lactating Holstein-Friesian cows. Cows were blocked according to lactation stage, parity, milk production, and presence of a rumen fistula (12 cows). Within each block cows were randomly assigned to 1 of 4 dietary treatments. All dietary treatments had a roughage-to-concentrate ratio of 80:20 based on dry matter. In the four diets, the roughage consisted of either 100% GS (**GS100**), 67% GS and 33% CS (**GS67**), 33% GS and 67% CS (**GS33**), or 100% CS (**GS0**; all dry matter basis).

This study, including the rumen fluid sampling, was conducted in accordance with Dutch law and approved by the Animal Care and Use Committee of Wageningen University.

Sample collection and processing

In total, samples from 12 rumen fistulated cows, three per dietary treatment, were used for metatranscriptome analysis. Rumen fluid was collected 3 hours after morning feeding on day 17 of the experimental period (for further details regarding the whole experimental period, see [424]). The samples were obtained as described previously [435], and collected from the middle of the ventral sac. The rumen fluid samples were immediately frozen on dry ice and subsequently transported to the laboratory where the samples were stored at -80°C until further analysis.

For RNA extraction, 1 ml rumen fluid was centrifuged for 5 min at 9000 g, after which the pellet was re-suspended in 500 µl TE buffer (Tris-HCl pH 7.6, EDTA, pH 8.0). Total RNA was extracted from the resuspended pellet according to the Macaloid-based RNA isolation protocol [436] with the use of Phase Lock Gel heavy (5 Prime GmbH, Hamburg) [437] during phase separation. The aqueous phase was purified using the RNAeasy mini kit (Qiagen, USA), including an on-column DNAseI (Roche, Germany) treatment as described previously [436]. Total RNA was eluted in 30 µl TE buffer. RNA quantity and quality were assessed using NanoDrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, USA) and Experion RNA Stdsens (Biorad Laboratories Inc., USA).

rRNA was removed from the total RNA samples using the Ribo-Zero^{TM} rRNA removal Kit (Meta-Bacteria; Epicentre, Madison, WI, USA) using 5 µg total RNA as input. Subsequently, barcoded cDNA libraries were constructed for each of the rRNA depleted samples using the ScriptSeq™ Complete Kit (Bacteria; Epicentre) according to manufacturer’s instructions in combination with Epicentre’s ScriptSeq Index PCR Primers.

The barcoded cDNA libraries were pooled and sent to GATC Biotech (Konstanz, Germany) for 150 bp single end sequencing on one single lane using the Illumina HiSeq2500 platform in combination with the TruSeq Rapid SBS (200 cycles) and TruSeq Rapid SR Cluster Kits (Illumina Inc., San Diego, CA, USA).

Bioinformatics

The general workflow for data quality assessment and filtering was adapted from [182]. rRNA reads were removed with SortMeRNA v1.9 [209] and all included databases. Adapters were trimmed with cutadapt v1.2.1 [383] using default settings except for an increased error value of 20 % for the adapters. The latter was chosen considering that with the default setting of 10% adapter sequences could still be found after trimming. Quality trimming was performed with PRINSEQ Lite v0.20.0 [384] with a minimum sequence length of 40 bp and a minimum quality of 30 at both ends of the read and as mean quality. All reads with non-IUPAC characters were discarded as were all reads containing more than three Ns. Details on the RNAseq raw data analysis can be found in Supplementary Table 1. The log files with the used commands can be found in supplementary file 1 and the used python script in supplementary file 2. The raw data was deposited at EBI ENA, and can be accessed under accession numbers ERS685245 - ERS685256.

Assembly and annotation

All reads which passed the quality assessment were pooled and cross-assembled with IDBA_UD version 1.1.1 with standard parameters [385]. A second dataset was added to the assembly to increase coverage (see supplementary materials & methods for details on this dataset). Prodigal v2.5 was used for prediction of protein coding DNA sequences (CDS) with the option for meta samples [191]. Proteins were annotated with InterProScan 5.4-47.0 [198] on the Dutch Science Grid. The annotation was further enhanced by adding EC numbers via PRIAM version March 06, 2013 [386]. Carbohydrate active modules were predicted with dbCAN release 3.0 [196]. Further EC numbers were derived by text mining and matching all InterproScan derived domain names against the BRENDA database (download 13.06.2013) [387]. Further details on the text mining can be found in the supplementary materials & methods.

Reads were mapped back to the assembled metatranscriptome with Bowtie2 v2.0.6 [327] using default settings. The resulting BAM files were converted with SAMtools v0.1.18 [388], and gene coverage was calculated with subread version 1.4.6 [438]. Read mappings to the contigs were inspected with Tablet [390]. The log files with the used commands for mapping and counting can be found in supplementary file 1 and the used python script in supplementary file 2. The whole read table including all annotations can be found in supplementary file 3.

Taxonomic assignments

All assembled contigs were analysed by blastn [177] against the NCBI NT database (download 22.01.2014) with standard parameters, except for an e-value of 0.0001, and against the human microbiome (download 08.05.2014), the NCBI bacterial draft genomes (download 23.01.2014), the NCBI protozoa genomes (download 08.05.2014), the human genome (download 30.12.2013, release 08.08.2013, NCBI *Homo sapiens* annotation release 105) and the genomes of *Bos mutus*, *Bos taurus* and *Bubalus bubalis* (download 21.05.2014). Taxonomy was estimated with the LCA algorithm as implemented in MEGAN [392], but with changed default parameters. Only hits exceeding a bitscore of 50 were considered, and of these only hits with a length of more than 100 nucleotides and that did not deviate more than 10% in length from the longest hit.

For contigs, which did not retain any hits after the filtering described above, another run with blastp of the associated proteins was performed against a custom download of the KEGG Orthology (KO) database (download 25.04.2014). Taxonomic assignment was again performed with the LCA algorithm, but only hits were considered, which did not deviate by more than 10% from the hit with the maximal identity.

All taxa, which were attributed to the phylum Chordata, kingdom Viridiplantae or to artificial constructs were considered to be contaminations and were automatically removed, as well as any proteins in which the annotation contained the word "microvirus". Furthermore, contigs that had a length of less than 300 nucleotides and which did not contain any proteins with a functional domain (disregarding the coils database) were discarded. Contigs belonging to the Illumina spike in PhiX phage were manually removed.

A compact schematic representation of the workflow is provided in Figure 1.

Statistical analysis

Differential expression was calculated in R version 3.1.1 [393] with the edgeR package release 3.0 [211]. Only genes, which had at least 50 reads mapped in all ten samples together were considered, and only genes with a p-value and q-value <0.05 in any of the comparisons were considered to be significantly differentially expressed. Furthermore, samples from cow #14 and #511 were excluded from the statistical analysis, due to dermal antibiotic treatment and due to feeding aberrations. To examine missing links within pathways, a q-value <0.1 was also considered (referred to as "lenient approach"). The used input file, the R script with the commands, output tables and MA plots can be found in supplementary file 4. To determine whether transcription levels corresponded to the diet components, the differentially expressed genes were sorted for each gene by diet group with increasing GS content, and an increasing or decreasing isotonic regression was fit on the data. An R^2 value of ≥0.8 was considered to be indicative of an increasing or decreasing profile, respectively, and all other values were considered to indicate that gene expression followed another, irregular, profile. Regression values and assignment of profile can be found in supplementary file 3. Isotonic regressions were computed in Python with scikit-learn version 0.15.2 [439]. Spearman rank correlation between the samples and Mann-Whitney U-test were calculated in Python with Scipy version 1.6.1 and NumPy version 0.9.0 [395].

Metabolic mapping

All derived EC numbers were mapped with custom scripts onto the KEGG database [227] and visualized using Python Scipy version 1.6.1 and NumPy version 0.9.0 [395] together with matplotlib version 1.4.3 [440].

Differentially expressed genes were investigated separately for microbial groups, which showed changes over multiple genes per pathway, and changed functions were determined by manual inspection of the KEGG maps.

Availability of data and material

All data has been deposited at the European Nucleotide Archive (ENA) under accession numbers ERS685245 - ERS685256 and ERS710560 - ERS710568

Results

Four experimental groups of three cows each were fed a control diet that contained only GS as roughage, and three different CS-enhanced diets for twelve days (Figure 1). From day 13 – 17, methane emission was measured using a respiration chamber, showing a significant reduction of methane emission with increasing CS proportion in the diet [424]. This decrease accounted for approximately 10% of the cows' methane emission. The analysis by van Gastelen *et al*. [424] showed that the dry matter feed intake of the different treatment groups did not differ significantly. Therefore the reduction in methane emission was not based on the available energy, but rather on the composition of the different diets.

Rumen fluid was collected at day 17, and used for microbial RNA extraction, mRNA enrichment and RNAseq. The complete set of RNAseq reads was cross-assembled into a single metatranscriptome. To determine activity per phylogenetic group the *de novo* assembled transcripts/genes were assigned to a taxonomic rank, and relative expression levels were obtained for four groups of animals fed different diets. Gene functional assignments were subsequently used to assess potential metabolic changes as predicted from the gene expression profiles observed in animals fed the four different diets, with a focus on carbohydrate breakdown, short chain fatty acid (SCFA) production and methane metabolism.

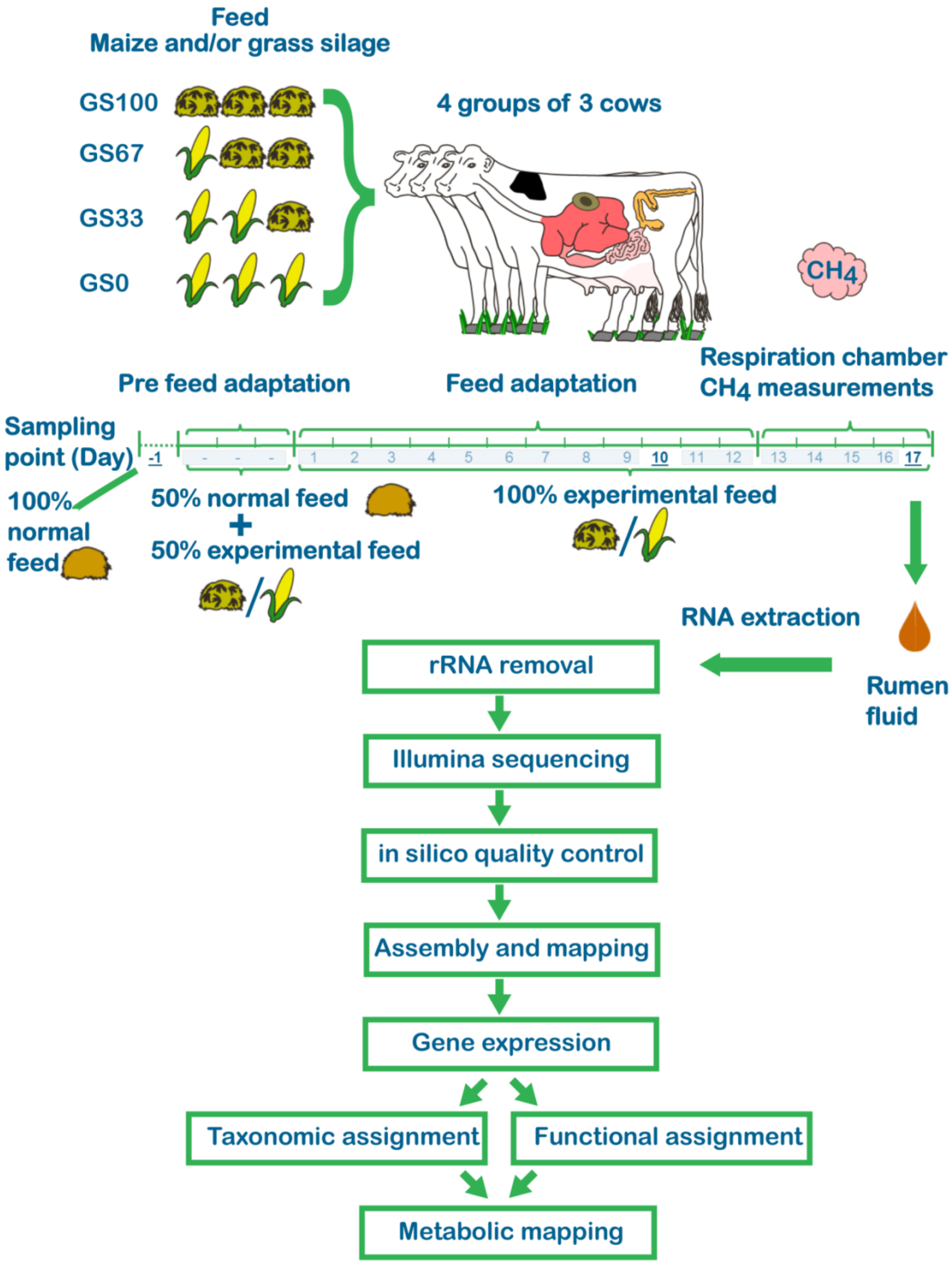


Figure 1: Study design. Four groups of three cows were allowed to adapt to one of four different experimental diets for twelve days. From day 13 – 17 methane emission was measured using a respiration chamber. Rumen fluid was collected at day 17 and used for microbial RNA extraction. See Methods section for details.

Sequence, assembly and annotation metrics

In total more than 160 million reads were obtained from twelve rumen fluid samples.

On average, 22.5% (Standard Deviation (SD) 6.15%) of all reads obtained per sample passed all filtering steps, retaining 18.5% of the total raw reads. Of these filtering steps, the filtering for rRNA sequences had the most impact, and removed the majority of the reads with an average of more than 63% (SD 8.75%) (all details are given in Supplementary Table 1**).** The majority of these rRNA reads (min. 96%) were matched to sequences from eukaryotes.

The assembly yielded 712,246 contigs with in total 866,052 protein coding sequences, a length of 414,768,486 bp and an N50 of 596. While the longest contig had a size of 54,845 bp, most contigs (645,026, 90.1%) were smaller than 1000 bp. A total amount of 30 million reads, on average 58% (SD 8.75%) of the reads per sample which passed quality filtering, could be mapped back to the assembly (see Supplementary Table 1; in the following, expression values will be given relative to the amount of mapped reads, referred to as "overall expression").

For 556,705 of the predicted protein encoding sequences a domain (excluding "Coils" domains) could be predicted. To 85,404 protein encoding sequences an EC number could be assigned.

A taxonomic classification could be obtained for 635,892 protein encoding sequences (73%), of which 282,074 could be classified at genus level. In total 1152 genera were detected, and additional 190 taxonomic assignments above the genus level were retrieved. 24 groups (at different taxonomic ranks) accounted for more than 58% of the total expression data (Figure 2). These groups included 13 genera (*Bacteroides, Butyrivibrio, Clostridium*, *Entamoeba, Entodinium*, *Eubacterium*, *Faecalibacterium*, *Fibrobacter*, *Methanobrevibacter*, *Methanosphaera*, *Plasmodium*, *Prevotella*, *Ruminococcus*) and 11 sequence clusters (not including the data assigned to the 13 genera) that could only be assigned at higher taxonomic levels (Archaea, Bacteria, Bacteroidales, Bacteroidetes, Clostridia, Clostridiales, Coriobacteriaceae, Eukaryota, Firmicutes, Methanobacteriaceae, Peptostreptococcaceae).

Fungal genes could be detected, but accounted for less than 0.1% of the overall expression. 184,991 genes without a taxonomic assignment accounted for 32% of the total expression. To only 34,731 of these genes (18.7%) any type of domain (excluding "Coils" domain) could be assigned, and only 2685 of these had an EC number assigned. Most present domains within the proteins encoded by taxonomically not assigned genes were generic domains (e.g. membrane lipoprotein attachment site, MORN repeat, P-loop containing nucleoside triphosphate hydrolase, WD40 repeat, etc.) without more specific functions.

Methanogens were represented by sequence assemblies that could be assigned to *Methanobrevibacter smithii, Methanobrevibacter ruminantium* and *Methanosphaera stadtmanae.* Reads mapping to protein coding genes assigned to methanogens captured on average 6.2% of the overall expression. In general, the overall taxonomic expression profile of the methanogens did not seem to change considerably between the different diets (Figure 2). When expression was summarized at genus level (or otherwise deepest taxonomically assigned group, as given in Figure 2, with minor groups treated together as “all other taxonomic groups”), the lowest correlation between all samples was 0.85. All microbial groups included in Figure 2 were furthermore tested (after exclusion of cows 511 and 14, due to mentioned aberrations) for statistically significant differences between animals fed the different diets (Mann–Whitney *U* test, p<0.05, not multi-test corrected), which was rejected for 150 out of 156 tests. None of the differences were statistically significant after multi-test correction (Bonferroni).

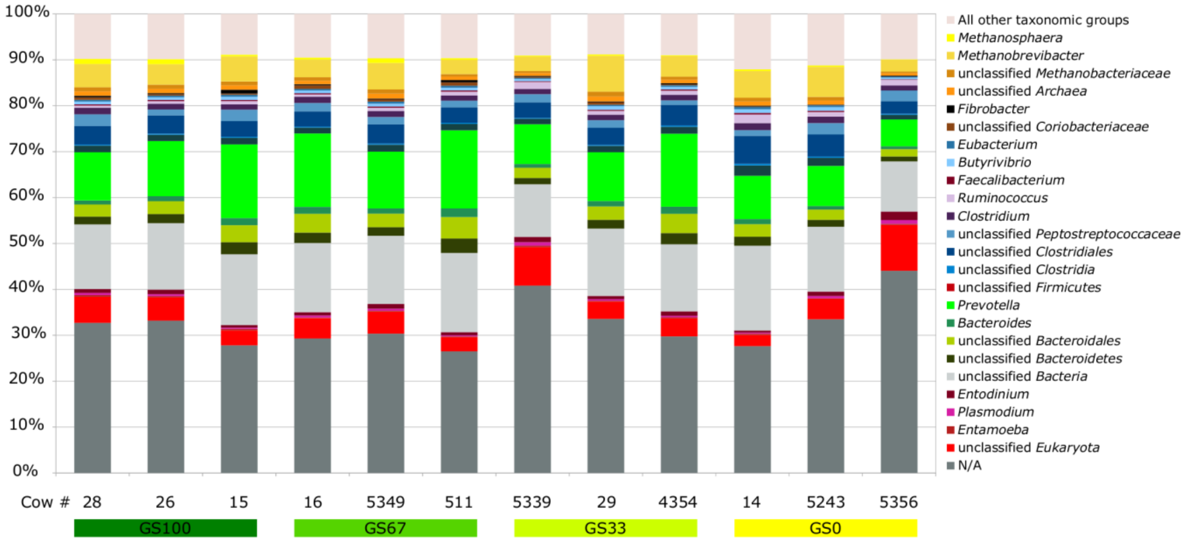


Figure 2: Taxonomic composition of metatranscriptome landscapes observed in animals fed one of four different diets. Diets and cows are indicated on the X-axis, taxonomic groups (at genus level, or otherwise deepest classification) are colour coded, see legend for details. N/A: No taxonomic rank could be assigned.

Differential expression analysis of the rumen microbiomes

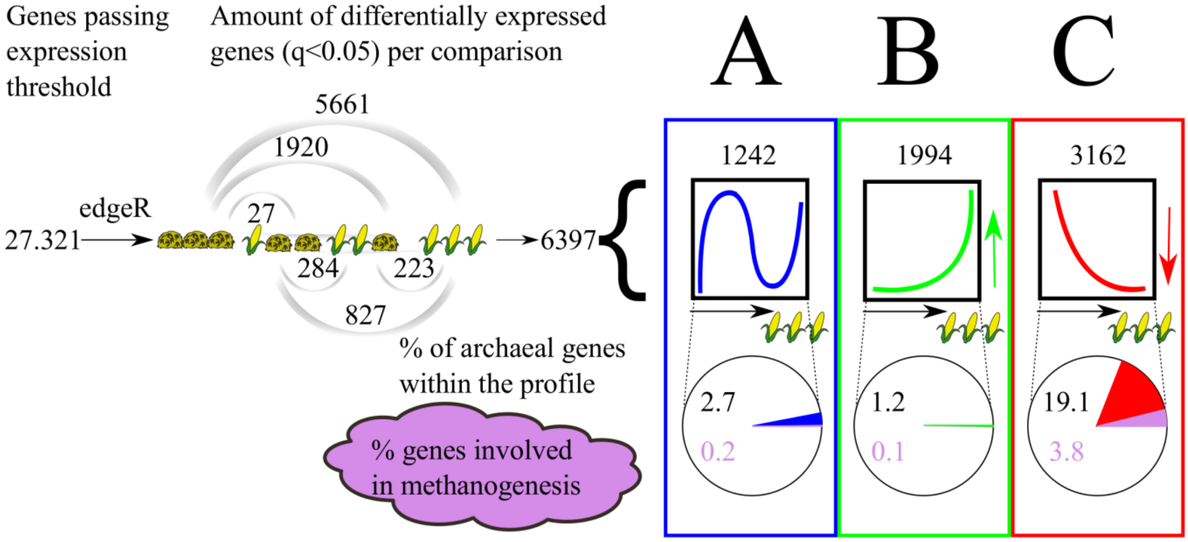

Genes passing expression threshold
Amount of differentially expressed genes (q<0.05) per comparison
A
B
C
5661
1920
27
edgeR
27.321
284
223
827
6397
1242
1994
3162
% of archaeal genes within the profile
% genes involved in methanogenesis
2.7
0.2
1.2
0.1
19.1
3.8

Figure 3: Differential expression analysis of the rumen microbiomes. Overview of the number of genes that were found to be differently expressed in pairwise comparisons of metatranscriptomes derived from animals fed different diets. Three profiles are distinguished: **Profile A**, genes with an expression, which does not follow a dietary pattern. **Profile B**, genes which are upregulated with increasing amounts of corn silage (CS). **Profile C**, genes, which are downregulated with increasing amounts of CS. Furthermore the results show that with the increase of CS, archaeal genes were mainly downregulated considerably affecting methane metabolism.

In total, 27,731 genes, which passed a set threshold for having captured at least 50 reads over all conditions combined, were subjected to the differential expression analysis, and 6397 were differentially expressed in at least one comparison (q<0.05). Three corn silage (CS) enhanced diet-induced expression profiles were distinguished (via regression analysis with isotonic regression), i.e. genes with an unlinked expression profile (profile A, 1241 genes), an induced expression corresponding to the amount of CS in the diet (profile B, 1994 genes), and a reduced expression corresponding to the amount of CS in the diet (profile C, 3162 genes) (Figure 3). Three heatmaps of all genes (per profile) can be found in supplementary file 5, displaying the overall trends within the data.

Taxonomic and functional analysis of the three diet induced expression profiles

Genes grouped into the three different expression profiles were investigated for their taxonomic and functional classification.

For profile A, i.e. genes that did not follow a diet specific expression profile, most genes were related to general energy metabolism/carbohydrate breakdown and ribosomal protein production, as well as transport reactions. No other major functions seemed to be affected in the diet-unspecific way characteristic of profile A, and most of the genes within this group could be linked to the Clostridiales*,* but also to Bacteroidales, Actinobacteria and Archaea.

Most predominantly represented taxa among genes following transcription profile B were bacteria belonging to the order Clostridiales, and to a lesser extent the genera Prevotella, Proteobacteria and Actinobacteria, but more than half of the differentially expressed genes could not be classified below kingdom level. The most affected functions were ribosomal protein production (mainly Eukaryota), and nucleotide metabolism in different groups, including the Eukaryota. The almost complete lack of genes associated with *Archaea* and/or methanogenesis among the genes with expression profile B indicated that there was hardly any increase of methanogenic activity with the increase of CS.

Among the genes exhibiting a lower expression upon increasing the amount of CS in the diet (profile C), the main represented microbial groups included three different methanogens (*Methanobrevibacter smithii*, *Methanobrevibacter ruminantium*, *Methanosphaera stadtmanae*), members of the genus *Prevotella*, and many genes, which could not be classified beyond the order Clostridiales. Functional profiling showed that the most downregulated processes were related to methanogenesis, electron transport and regulatory processes in the *Archaea*, as well as general metabolic functions like glycolysis, ATP generation or ribosomal protein production in all affected groups. Increased expression could also be observed for nine genes encoding putative non-ribosomal peptide synthase (NRPS) modules, among which three were taxonomically linked to *M. ruminantium* whereas the other six NRPS modules could not be classified beyond the kingdom bacteria.

With an increase of CS in the diet, Eukaryota appeared to show a decrease in their expression of genes encoding glycosylhydrolases (GH) and glycosyltransferases (GT). Furthermore, they also showed differential regulation of genes associated with movement abilities and cilia/cytoskeleton assembly, chaperons and ribosomal proteins in response to the diet changes. Most of the sequences (71.9%) assigned to the Eukaryota could not be classified below the kingdom level. For example, of the 85 differentially expressed genes encoding proteins involved in cilia/cytoskeleton assembly, only 12 could be assigned to a rank more specific than the kingdom level. Within all the classified eukaryotic sequences that showed consistent downregulation with increasing CS in the diet, the phylum Apicomplexa was the most represented, whereas the family of Ophyoscolecidae (*Entodinium, Epidinium*) showed a specific downregulation of GH encoding genes.

Microbial starch and cellulose metabolism in cows fed with different diets

The expression of genes related to the breakdown of different complex carbohydrates differed considerably between animals fed different diets. Profile A did not include major changes in genes coding for carbohydrate degradation associated enzymes.

For genes following expression profile B, an increase of CS in the diet mainly lead to the increased expression of genes encoding different extracellular binding proteins in the genera *Ruminococcus, Bifidobacterium* and *Entodinium*, as well as an increase in the expression of genes coding for starch binding modules (CAZy classes CBM25 and CBM26) and alpha-amylases (GH13).

Most carbohydrate-metabolism associated genes affected by an increase in CS in the diet, however, followed expression profile C. With an increase of the CS in the feed, a downregulation of multiple genes involved in the breakdown of plant cell walls and their constituents could be observed, such as all the steps involved in cellulose degradation [441]. Expression of genes encoding endocellulases (CAZy classes GH5, GH9, GH45; mainly assigned to *Fibrobacter*), catalysing the first step of cellulose breakdown, was most affected, followed by genes that code for exocellulases (GH48, *Ruminococcus*) and beta-glucosidases (GH3), catalysing the second and the last step of cellulose breakdown, respectively, as well as genes encoding cellulose binding modules (e.g. CBM4, CBM13). Downregulation of the expression of genes encoding proteins involved in the breakdown of hemicellulose constituents (xylan, mannan, galactan/pectate, rhamnose) could also be observed, including genes encoding endo-1,4-beta-xylanases (GH10, GH11), beta-mannanase (GH26), pectate lyase (PL3), alpha-L-rhamnosidase (GH78), beta-1,4-galactan binding (CBM61), and xylan binding modules (e.g. CBM35). Expression of genes related to transport of glucose into the cells was also downregulated (monosaccharide transporters, EC 3.6.3.17). An overview of differentially expressed genes encoding glycosylhydrolases and carbohydrate-binding modules, including their taxonomic distribution, is presented in Figure 4 and Figure 5, respectively.

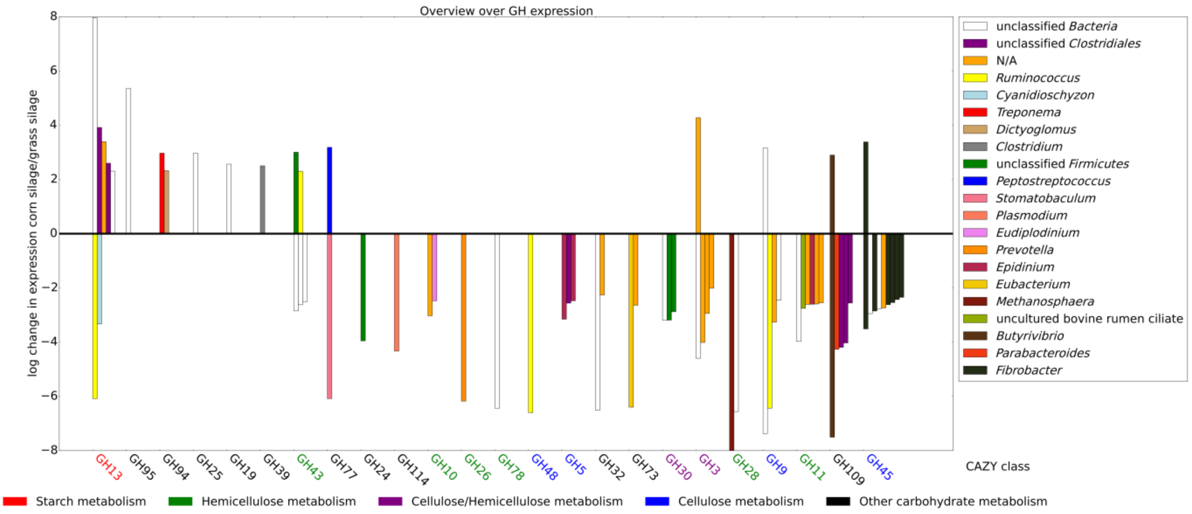


Figure 4: Log10 fold changes in expression of differentially expressed glycosylhydrolase encoding genes in a comparison of the 100% corn silage diet (GS0) versus the 100% grass silage diet (GS100). Positive values indicate an upregulation of gene expression in the corn silage diet. N/A: No taxonomic rank could be assigned. Colour-coding of bars indicate different taxonomic groups, whereas colour-coding of protein families indicate their involvement in the metabolism of different carbohydrates.

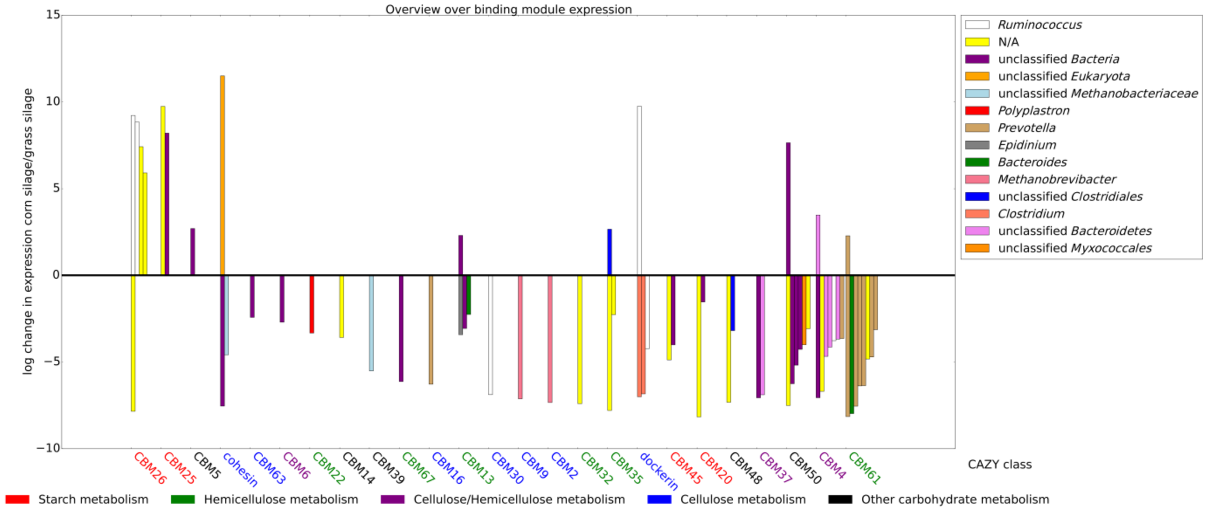


Figure 5: Log10 fold changes in expression of differentially expressed carbohydrate binding module encoding genes in a comparison of the 100% corn silage diet (GS0) versus the 100% grass silage diet (GS100). Positive values indicate upregulation of gene expression in the corn silage diet. N/A: No taxonomic rank could be assigned. Colour-coding of bars indicate different taxonomic groups, whereas colour-coding of protein families indicate their involvement in the metabolism of different carbohydrates.

With the increase of CS, a downregulation (profile C) could be observed for *susC* and *susD* genes coding for starch binding proteins, and which could be assigned to the phylum Bacteroidetes, mainly in the genus *Prevotella*. A downregulation of expression of genes encoding proteins involved in cellulose binding was also found, including e.g. sortases, cohesins, dockerins, extracellular binding and calcium binding domains, which potentially could belong to a cellulosome [442, 443]. This was mainly observed for genes assigned to the families Cellulomonadaceae, Clostridiaceae, Lachnospiraceae and Ruminococcaceae. Many functionally similar downregulated protein-coding genes could not be assigned to a taxonomic rank below the superkingdom level, mainly in the bacteria. The downregulation of a gene encoding a cohesin module was also detected in the *Archaea*, as well as the upregulation in the expression of a cohesin and dockerin module with an increase of CS.

Microbial short chain fatty acid metabolism in cows fed different diets

The production of SCFAs is an important function of the rumen microbiome. These metabolites are taken up by the host and serve as an energy source [94], have a considerable effect on methane production [444], and affect the pH, which in turn has an influence on the animal's wellbeing [445]. In the study by Gastelen *et al*., only a small significant reduction in the SCFA butyrate was reported, with other SCFAs not changing significantly.

Increased expression upon an increase of CS in the diet (profile B) was found for genes coding for proteins which are involved in the conversion of acetyl-CoA to crotonyl-CoA, which is part of butyrate synthesis. This increase was found within the family Lachnospiraceae. The total expression of this family was on average 1.9% in all samples.

Reduced expression upon an increase of CS in the diet (profile C) was observed for genes encoding proteins involved in butyrate metabolism, and again mainly for genes assigned to the Lachnospiraceae. Several of the downregulated genes encode proteins catalysing the reactions from pyruvate to crotonyl-CoA, via acetyl-CoA, acetoacetyl-CoA and (S)-3-hydroxybutanoyl-CoA. Genes that code for enzymes catalysing the last steps to butyrate via crotonyl-CoA and butanoyl-CoA were also present in the assembly, but were not found to be differentially expressed in any of the conditions. Thus, the here presented data provide an inconclusive picture regarding the regulation of genes encoding proteins involved in ruminal butyrate production. Furthermore, consistent differential expression patterns could also not be observed for genes involved in the formation of the other SCFAs acetate or propionate. Genes encoding SCFA transporters were present in the assembly, but were not differentially expressed. Overall, these observations are in line with the fact that total SCFA concentration was found to be not affected by increasing CS in the diet, with only a minor, albeit significant increase in the molar proportion of butyrate [424].

Expression of archaeal genes involved in methane metabolism

A considerable amount of differentially expressed genes in the *Archaea* was found to encode proteins involved in methane metabolism. Based on the RNAseq data almost the complete pathways leading to methanogenesis could be reconstructed (Fig. 6). Closer inspection revealed that with an increase of CS in the diet, nearly all genes of the methanogenesis pathways were downregulated in a subset of the *Archaea* (expression profile C). Of the four possible methanogenic pathways, those for the production of methane from methanol/hydrogen, as well as from formate/carbon dioxide and hydrogen were affected. Proteins for the utilization of trimethylamines into methane could be detected in the dataset, but were not differentially expressed between animals fed the different diets. The pathway for methanogenesis from acetate was absent in the dataset.

Among genes assigned to *Methanosphaera stadtmanae,* genes coding for proteins involved in the conversion of methanol to methyl-CoM (methanol-corrinoid protein Co-methyltransferase**,** EC 2.1.1.90) and of methyl-CoM to methane (methyl-CoM reductase, EC 2.8.4.1) showed consistent downregulation with increasing CS in the diet (Figure 6).

Compared to changes observed for *Methanosphaera stadtmanae*, the change in the transcription pattern of genes encoding proteins involved in methanogenesis from hydrogen and formate in *Methanobrevibacter smithii* was more extensive. More specifically, the expression of genes associated with the methanogenesis pathway with formate/hydrogen was downregulated in nearly all steps (besides formylmethanofuran-tetrahydromethanopterin N-formyltransferase, EC 2.3.1.101), following expression profile C. In addition, expression of several genes encoding proteins involved in the biosynthesis of coenzyme F420 was downregulated with an increasing amount of CS in the diet (profile C). Some of these reactions could only be assigned to taxonomic levels above species, but the placements of these functions in the metabolic network indicate that they most probably can also be assigned to *M. smithii.*

Expression of genes encoding transporters for formate uptake were also downregulated (profile C), as well as genes involved in other processes related to methanogenesis, e.g. the general production of ATP, electron transport via the membrane, and sodium transport.

Nearly none of the genes that could be assigned to the third detected major methanogen in the dataset, *Methanobrevibacter ruminantium*, showed considerable downregulation, however, it should be noted that several archaeal genes, including several genes encoding proteins involved in methanogenesis, could not be classified at the species level and therefore it cannot be excluded that some of these in fact also belong to this species. Differential regulation of genes assigned to a potential syntrophic partner of *M. ruminantium, Butyrivibrio proteoclasticus* [446], could only be detected in a few genes. Genes assigned to other formate producing organisms were also present in the data, pointing towards their potential involvement as syntrophic partners, however, no differential expression was observed for these genes, making deduction of possible syntrophic connections difficult.

Further analysis of the data at the functional level showed downregulation of the expression of genes encoding proteins linked to the production of necessary substrates for methanogenesis. Expression of one of the genes encoding a subunit of pyruvate formate lyase (EC 2.3.1.54) that catalyses the production of formate from pyruvate was downregulated in a bacterium in the order Clostridiales, which could not further be classified, as well as in *Eubacterium hallii*. At the same time, several genes encoding proteins involved in the degradation of cellulose were found downregulated in animals fed CS-containing diets (profile C), and could be assigned to a not further classifiable bacterium in the order Clostridiales as well as *Ruminococcus flavefaciens, Fibrobacter succinogenes*, and several other bacteria/eukaryotes. A downregulation of genes that code for proteins involved in the production of the other substrates needed for methanogenesis, hydrogen and carbon dioxide, could not be detected.

Interestingly, using a more lenient approach (see Methods) a downregulation of expression of a gene for the production of the second major substrate for methanogenesis, methanol, was observed. More specifically, an unspecified Clostridiales bacterium showed decreased expression of a gene encoding pectinesterase (EC 3.1.1.11), catalysing the degradation of pectin to pectate and methanol.

An overview of the metabolic consequences of the observed changes in gene expression profiles is provided in Figure 6. A version of this figure with more details and a table with all reactions and assigned genes can be found in supplementary figure 1 and supplementary file 6.

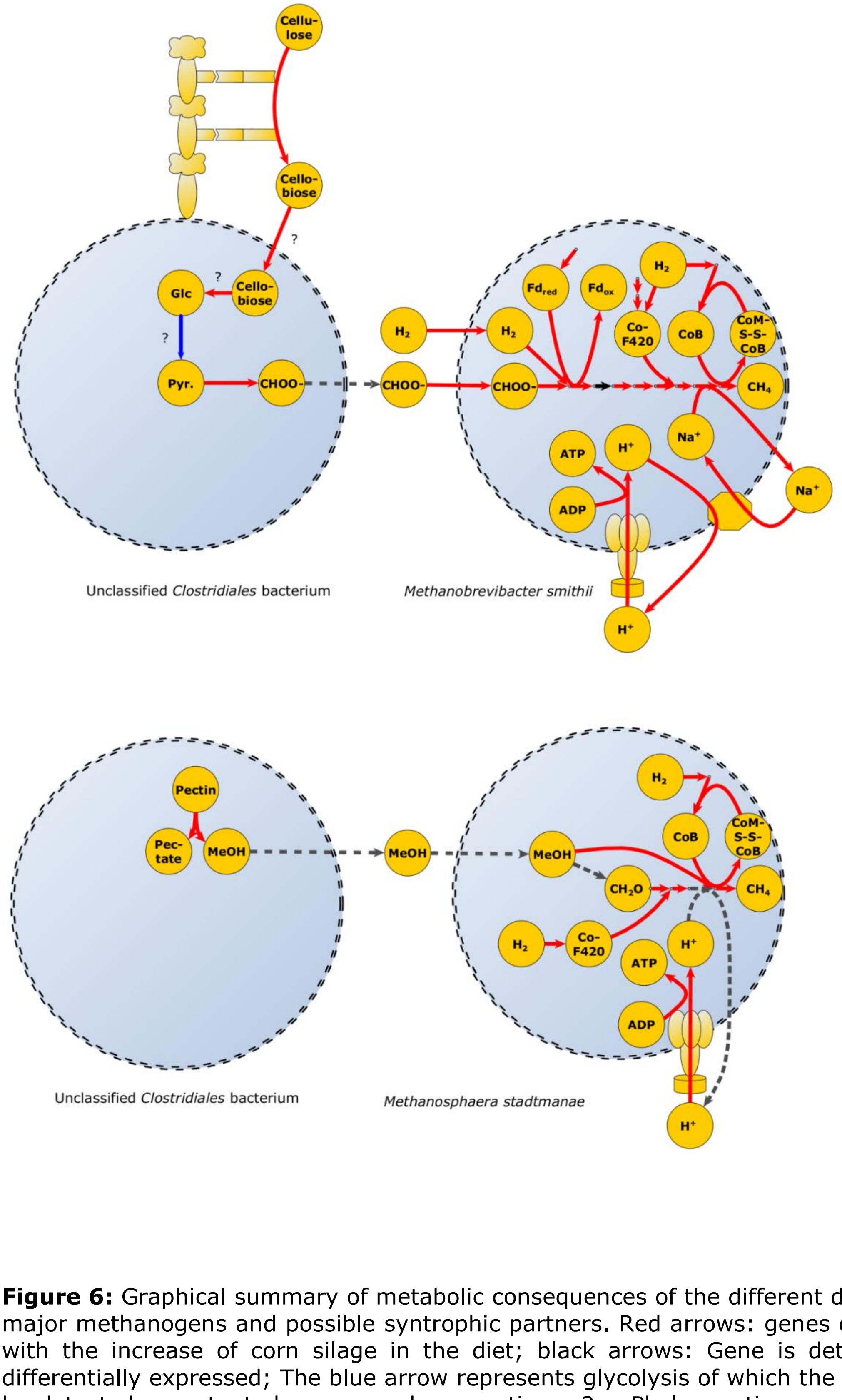


Figure 6: Graphical summary of metabolic consequences of the different diets in the two major methanogens and possible syntrophic partners. Red arrows: genes downregulated with the increase of corn silage in the diet; black arrows: Gene is detected but not differentially expressed; The blue arrow represents glycolysis of which the majority could be detected; punctuated arrow: orphan reactions; ? = Phylogenetic association unclear. Pyr. = Pyruvate, $CHOO^-$ = Formate, FD_{ox} = oxidized Ferredoxin, FD_{red} = reduced Ferredoxin, CH_2O = Formaldehyde.

Discussion

How feed affects methanogenesis

The rumen microbiome is a complex ecosystem, and its dynamics are determined by many variables. Most investigations to date have been focussed on the community composition and changes therein in response to different perturbations. In a recent metagenomic study by Roehe *et al*. [433] on animals fed similar diets as the ones tested here the authors found no considerable effect on the composition of the microbiome. Here we show that in response to a diet change, gene expression within a microbiome and consequently the metabolic profile may change. Differential expression analysis revealed that although there were no extensive changes visible within the overall community expression, in line with what has previously been noticed for the sheep rumen [428], major effects could be seen regarding the expression of genes related to methane metabolism, which are also in agreement with genes which were prior identified within the metagenomics dataset by Roehe *et al*. and related publications [433, 447, 448]. In two of the three methanogens identified in the dataset a coordinated downregulation of genes involved in methanogenesis as response to increased CS in the diet could be observed. Thus not only isolated single nodes involved in methanogenesis, but whole pathways were downregulated. We further found evidence for a possible syntrophy between these methanogens and several yet unidentified members of the rumen community belonging to the order of Clostridiales, which might contribute to the production of the necessary substrates (formate, methanol) for the methanogens, which was also discussed (albeit with potentially different syntrophy partners) in a related setup by Parmar *et al*. [449]. Additionally we observed a downregulation of cellulose degradation functions with increased CS in the diet. For *M. ruminatium*, we did not see a significant response to the diet changes nor did we see a significant response in possible syntrophic partners. Thus it may be that in addition to diet changes other types of biological effectors are necessary to further influence the process. Our findings are also in contrast to those reported by Shi *et al*.[428], who concluded that in the sheep rumen the supply of hydrogen is the determining factor for methane output, whereas in the present study the supply of other substrates seem to have a bigger influence.

We further observed community wide responses to the change in the main energy/carbon source, with a shift in the involved glycosylhydrolases over multiple organisms and phylogenetic branches. Nevertheless, we did not observe a response in all members of the microbial community. While there was a definite downregulation of certain processes like methanogenesis, these processes were not affected in all organisms. To this end, it should be noted that the total gene count assigned to members of the Archaea greatly exceeded the size of currently known individual archaeal genomes, suggesting the presence of multiple strains of the same species in this environment [450, 451]. Not all of these strains seemed to be affected by the different diets, as there were also instances of pathways, which did not show a differential regulation at all. As already observed here for the different species of methanogens, which were potentially affected because their syntrophic partners were affected, this could also be the case for the different strains of the same species, which might inhabit different niches in the rumen. It cannot be expected that e.g. methanogens living intracellularly within protozoa [452] are in the same way affected as free living methanogens are, and that populations living closer to the substrates, i.e.

those associated with the fibre fraction, will show the same behaviour as populations in the liquid fraction of the rumen [453]. Finally, as overall a reduction of methane production by ~10% was observed in this study when comparing animals fed either the GS or CS diets, it is perceivable that not all pathways and microorganisms are affected to an extend that would be detectable in significant differences in gene expression levels, also considering the relatively small sample size of three animals per experimental group.

Unexpected findings and limitations

Several findings in this study were surprising, at least at first glance.

As shown in Figure 2, and also shown by the statistical testing, the overall expression profile did not change significantly. A major change in the supplied feed was expected to result in significant changes though. Also the study of Roehe *et al*. [433] showed no considerable changes in the relative abundance of organisms in a similar setting. We showed that the main changes are not within a taxonomic group, but rather the expression patterns per taxonomic group, which also explains the findings by [433].

There are also concerns that differential expression analysis in communities could not reflect actual differential expression, but rather a change in organism abundance, leading to wrongly perceived changes in expression. Since in this dataset the overall expression profile per group did not statistically significantly change (although the small sampling size gives only limited power to detect this change), this is likely not an issue, and genes detected as differentially expressed are probably truly differentially expressed.

The overall taxonomic composition itself as shown in Figure 2 in general agrees with previous findings, as most of the major taxonomic groups were reported previously [454]. This is also the case for the methanogens, which are similar to the ones commonly found the rumen of cows [455] and other ruminants [456, 457]. Despite this, it should be noted that the genes assigned to *Methanobrevibacter smithii* most likely belong to a related species/group of *Methanobrevibacter*, since *M. smithii* itself is not a dominant member of the rumen microbiota, but the closest sequenced relative of the species appearing in the rumen [458].

As shown in Figure 3, we also recovered changing expression profiles, which did not correspond with the diets. We were not able to find any specific functional background for these profiles, and suspect that some organisms are influenced more by the surrounding community members and not primarily by the diet, or maybe inhabit very specific niches. This would be in agreement with the findings in Figure 4 and 5, which show that a minor amount of carbohydrate active enzymes and binding modules show expression profiles against the expected trend, e.g. increase in expression of some cellulose degrading enzymes while less cellulose is fed [459, 460]. It could also be possible that this change in expression reflects a change in metabolic strategy. As response to e.g. the lower abundance of cellulose in the environment, the affected organisms could attempt to downregulate the expression of genes coding for cellulose binding modules with low affinity, and upregulate the expression for genes coding for modules with high affinity. This mechanism is similar to the regulation of carbohydrate transporters in different organisms [459, 460]. Additionally it needs to be considered that initial annotations might not always be correct. We found an increase in cohesin and dockerin coding modules with an increase of starch in the diet. These components are primary known as cellulosome components, but non-cellulosomal origin of these modules has been reported before [461, 462]. Furthermore one of these modules was found in the Archaea, which are not known to harbour either cellulosomal complexes or their starch counterparts. The same issue holds for the downregulation in expression of the genes coding for different starch binding proteins, *susC* and *susD*, which have been found to not only be starch binding, but also cellulose binding [463].

Another finding, which was obvious in the investigated data, is the substantial decrease in expression of genes coding for proteins involved in cytoskeleton assembly in different Eukaryota. As several Archaea are endosymbionts of Protozoa, it can be speculated that an experimental change, which has an impact on the symbionts, will also affect their host [452] (although this relationship is also not entirely clear [464]). General cellular processes, like replication, in which the cytoskeleton is involved, will then probably be directly affected, and this has been observed before in a different setting with intracellular Archaea [465]. Recently the high abundance of these proteins in the rumen proteome also have been demonstrated [466].

At last, the biggest limitation on this study are the lack of sequencing depth and little replication. The former was mainly caused by the inefficiency of the ribosomal rRNA depletion. The method used could not remove all rRNA, due to the diversity of unknown eukaryotic sequences, which resulted in a lower sequencing depth than expected. Also due to the low number of replicates, an arbitrary cutoff for the tested genes had to be applied, which is common practice and can help in some settings to increase power [467], and therefore it was not possible to find more subtle changes in the expression levels (e.g. a change in transcription levels of the butyrogenic pathway). Therefore this work mainly focused on changes within more highly expressed genes, and most changes were also not dependent on single p-values, but supported by expression changes in multiple genes. This has still lead to the ability to track the impact of diet on methane production, which was the aim of this study, and other effects, which were not initially expected, could still be observed. It still needs to be pointed out that the amount of replication was very small and probably too small for this type of experiment, and that many changes, including not only subtle ones, were potentially missed due to this setup.

In summary, in this study we found a significant effect of a dietary change on the gene expression in the cow rumen. A substantial fraction of the affected genes was related to methane emission, showing that a decrease in cellulose in the diet decreased the gene expression of methane related pathways. The here presented metatranscriptomic analysis is in agreement with the experimental measurements, which showed a decrease in methane emissions with the diet change [424], suggesting that a change in the feed regime can have a positive effect on microbial GHG emissions.

Acknowledgements

The authors thank Jasper Koehorst for his help with the transcriptome annotation, and Jesse van Dam, Maria Suarez-Diez and Edoardo Saccenti for helpful discussions.

This computational work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

Funding

B. Hornung was supported by Wageningen University and the Wageningen Institute for Environment and Climate Research (WIMEK) through the IP/OP program Systems Biology (project KB-17-003.02-023).

This work was furthermore supported by funding from the Top Institute Food and Nutrition, Wageningen, The Netherlands, a public-private partnership on pre-competitive research in food and nutrition. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary information

Supplementary information can be found online at:

https://doi.org/10.1101/275883

Chapter 6: General discussion and future perspectives

The final goal of this study was to model the make-up and behaviour of microbial communities, and to study how they could be affected by different environmental conditions such as host diet or nutrient availability. Although this final goal has not been fully achieved yet, various steps along the path were successfully taken during this work. The results of the research described in this thesis, as well as potential reasons for current bottlenecks will be discussed, and perspectives for future research will be provided.

Issues with simulating meta-metabolism

In the course of this PhD thesis multiple times it was attempted to perform simulations on the metabolism of the various microbiomes. Genome-scale, constraint-based metabolic models (GSMM) were chosen over ODE- or agent-based models, due to the fact that metabolism seemed to be the most interesting choice in most of the cases, especially given the available data (although mathematical models could have been an interesting choice too [468]). There are various options to construct GSMMs [241, 469], each with their own advantages and limitations. The easiest type of model is the supra-organism [241]. In this type of model species boundaries are not considered, and all possible metabolic reactions are included in one model, basically modelling the whole community as one organism. The advantage of this approach is that it simplifies the overall model, and makes exact species delineation/binning unnecessary, but has the disadvantage that interactions cannot be studied. The second type of model considers all organisms separately, but still considers them to be one entity, with a common optimization goal for the whole community [469]. The last type of model considers each organism separately, also in respect to how the organisms behave, and gives each organism its own optimization goal (e.g. [254]).

All of these models have lead already to interesting findings. Greenblum *et al*. [470] constructed a supra-organismal model from the faecal metagenome of healthy individuals and IBD patients. Their network analysis showed that both states were associated with specific attributes in the periphery of the network, indicating differences in nutrient utilization of the gut community. In the publication by Stolyar *et al*. [471] a model was constructed of a consortium, containing a lactate utilizer and a methanogen. This model could correctly capture the interaction between both organisms, including metabolite exchanges, and showed also that this type of exchange was essential for the growth of the system. In a similar approach, Ponce-de-Leon *et al*. [472] showed how an endosymbiotic bacterial consortium is tightly integrated with its host, and that the exchange of nutrients is critical for its survival. Overall, this shows that exciting results can be obtained with these models, independent of what type is chosen, as long as it is appropriate for the interrogated question.

The first approach, the supra-organismal model, was not really considered for this work (only for intermediate testing and checking), since it was not necessarily the goal to optimize a community, or to investigate it as a whole, but also to understand the interactions between community members. This also turned out to be important in this work (e.g. chapter 4, different behaviours of *Bacteroides* and *Clostridiales* groups, or chapter 5, potential interactions of *Methanobrevibacter smithii* with another community member).

The choice between the second and third type of model was initially not made, since the third type only really became available during the course of this work and hence, the second model was initially considered. In the process of the efforts to implement such models, numerous issues were encountered (not necessarily with the methodology, but also with the investigated habitats), which made this choice obsolete, and which will be discussed in the following paragraphs.

The first challenge was to obtain sufficiently processed data. Despite the fact that transcriptomic data was always available, none of the projects was in a state where it could be used for modelling, because e.g. it was unclear which reactions were occurring at which rates. The bioinformatics processing, which would potentially lead to the gain of such information, took over most of the work described in this PhD thesis. The remaining time was at the end insufficient to pursue this goal, since performing reasonable GSMM simulations and solving all associated problems with the related models can be a thesis subject on its own [473].

The second hurdle, which was also encountered by other scientists working in this field [248, 473-475] is the need for manual curation. It is not possible to sufficiently perform manual curation for such large datasets, and it will also lead to inconsistencies due to differing opinions and expertise of the curating domain experts. In turn, the automatic annotation of (meta)-genomes (or transcriptomes) is in general not sufficient for high quality simulations, without having any manual curation. This issue has multiple sub-issues, which will be explained in the following paragraphs.

The first issue concerns the ambiguity of EC numbers, on which these model-based simulations would ideally be based upon. While EC numbers were initially created to provide a structured system of unique identifiers for chemical reactions [476], this turns out to be too imprecise for modelling (despite still being maybe the most sensible system). The best known example here would be the alcohol dehydrogenase EC 1.1.1.1. The definition of the IUBMB is "(1) a primary alcohol + NAD+ = an aldehyde + NADH + H+". The challenge is that general entities such as "primary alcohol" or "an aldehyde" do not necessarily exist in other reactions, where specific compounds are given instead, and that most often there is no direct, i.e. automated, possibility to link these entities. The Metacyc database [194] has overcome this issue partially with a hierarchical system, however, the underlying system is not made for metagenomics data, is not particularly user friendly in these regards, and therefore it is not often used for this purpose. As an example, it is possible to import all metabolic reactions, but it is not possible to utilize the associated taxonomic information in any way. If there was a possibility to at least integrate this information into the interface, or to mark pathways to be specific for only certain organisms in this dataset, it would make the functional mining and understanding of a metagenome considerably easier. In the KEGG database [227] specific reactions are associated to the EC numbers, which solves the issue partially, as it also does not require the user, which in many cases will be biologist, to be fully proficient in every detail of the underlying (bio)chemistry to resolve this problem.

Taking metabolic models from single organisms to mixed consortia adds another layer of complexity, where overly generic descriptions of enzymatic reactions pose important limitations to the modeller. More specifically, while in single organism models these issues can potentially (but not always) be resolved by manual curation, and it can be decided if “primary alcohol” refers to ethanol, propanol or other compounds, this turns out to be impossible for metagenomics data, where specific reactions can occur hundreds of times in a single metagenome, and where different organisms can perform similar, but not identical reactions, involving different compounds belonging to the same EC category. At this stage not even the “underground reactions” (side reactions besides the main reaction as defined in the EC definition, with minor turnover rates) [477] are taken into account, which could potentially lead to a totally different structure of the resulting metabolic network.

The second issue is that not all EC numbers can be predicted with sufficient accuracy. This applies to many cases, where enzymes show a high degree of sequence similarity. Examples include enzymes from CAZy class 13, which are often falsely identified as alpha-amylase, or where probably not enough/all representatives might have been found. This includes the case of the enzyme FolQ, EC 3.6.1.67, dihydroneopterin triphosphate diphosphatase, which is frequently not predicted in many organisms, despite the fact that it must be present, given the growth conditions of these organisms [478]. In some cases the predicted models are also not constructed in a proper way, or the underlying data is not sufficient. This leads to predictions which are insufficient for further usage (as in e.g. transporters, to which no substrates can be assigned).

Furthermore, to a quarter of all EC numbers not even a single associated enzyme exists (so called “orphan reactions”) [479]. Therefore, a reconstruction based on predicted EC numbers from the genes within the dataset will have gaps, which will eventually lead to models, which cannot be run [473]. In order to overcome this issue as much as possible, automatic gap-filling algorithms have been developed [480-482], allowing to initiate the metabolic reconstruction of increasingly complex microbial ecosystems. Currently available algorithms perform sufficiently well especially in cases where only small parts of a pathway are missing, and thus gap-filling in microbiomes with strongly understudied metabolic areas remain challenging.

The third issue results from a combination of the technical setup and the underlying biology. In chapter 5, while studying the rumen metabolism, the central research question focused on methane metabolism, and how this is affected by dietary interventions. Methane metabolism itself is a complicated network [483, 484], where at least seven different cofactors specific for this metabolism are involved (six mentioned in [484], as well as methanofuran). The technical issue in this case is that GSMMs do not properly simulate cofactors, and that there is no consensus on how to integrate them [485]. In GSMMs it is assumed that there is a steady state of metabolites, and that X is consumed when it is turned over to Y. Obviously, this is not the case for cofactors, which are recycled to their initial state at the end of such reactions. In many cases cofactors are just ignored, since it is not possible to simulate them properly, with the further reason that their abundance is so small (due to the fact that they persist), that there is most likely no considerable impact on the simulation. While this can probably be true under the right experimental conditions, we have the interesting case in our metatranscriptomic data that interdependence of organisms can be a driving factor for the structure of a community [79, 119], and therefore the ability or inability of the different organisms to produce cofactors needed by other organisms can be critical. More specifically, in the ruminal ecosystem studied in chapter 5, two different methanogens (*Methanobrevibacter smithii* and *Methanobrevibacter ruminantium*) co-exist, of which the latter is deficient for coenzyme M [446], a necessary cofactor at the final step of methanogenesis [484]. Interestingly, *M. ruminantium* is less impacted by the intervention in chapter 5 than *M. smithii*, which might or might not be attributable to also this fact (possibly via indirect connections to other cofactor producing organisms which were differently affected, or maybe not at all connected to this phenomenon). Therefore, in all simulations of microbial community metabolism, a qualitative assessment of the cofactor production should be included, e.g. by including the production of a nearly negligible amount of the relevant cofactors as done in e.g. [486] or [487].

Besides the above-mentioned, more technical, issue related to the lack of inclusion of cofactors in models, an additional challenge arises with insufficient biochemical knowledge regarding the mechanistic role of specific cofactors. Staying with the example of cofactors involved in ruminal methanogenesis, the underlying biology is only clear for five of the mentioned seven different cofactors [484]. The biosynthesis of methanofuran was only recently elucidated [488], whereas the pathway for the biosynthesis of methanophenazine is still not fully clear [489, 490]. Overall, these issues would have led to a big lack of underlying accuracy of the resulting simulations.

The fourth issue is again related to the insufficient match of the used technology (GSMM) to the underlying biology. In chapter 4, the degradation of different carbohydrate polymers by human faecal inocula was investigated, with a specific focus on possible mutualism in the degradation and on the mode of degradation. The issue in this case is that GSMMs cannot simulate polymers, especially not concerning their structure and size distribution. In different publications, this is overcome by introducing dummy reactions in the form of either "Polymer -> X Glucose" (e.g. [491]), or by defining polymers of a specific size (e.g. [492]) and defining the necessary reactions for the degradation. While for sure again the workaround of manual curation is possible, this could also only been done in this chapter with a huge effort, due to the complicated structure of the polymer. The structure of the polymer dictates which type of bond (alpha-1,4 or alpha-1,6) is or is not exposed at the surface of the branched polymer. Furthermore, the different enzymes involved in polymer breakdown can only access specific bonds in specific configurations, but in varying order, and with varying breakdown products, depending on the structure. Already for a small polymer this leads to an excessive amount of possible reactions (of which only the minority has a precise EC number).

The last complicating factor which was encountered in the meta-omics projects was the lack of genomic data. While meta-transcriptomic data can often replace genomic data with the added benefit of providing information regarding the expression of genes, the meta-transcriptomic data lacks a key feature of the metagenomic data: Connectivity. In meta-genomic data big contigs can often be assembled, allowing to assign genes to a specific organism, although it should be noted that, depending on complexity of the communities and sequencing depth, metagenomics-derived genome binning also has its challenges (unpublished results). With only transcriptomic data, generation of larger contigs is feasible only to a limited extent, and thus, the only way to associate genes to each other, which are not on the same or overlapping transcripts, is via their taxonomic assignment. This is, however, often imprecise, and even if a species can be assigned, it is often the case that multiple strains of the same species occur in the same biome. This can again be illustrated with data obtained from the rumen (chapter 5), where some of the reactions involved in methane metabolism could not clearly be associated to one of the two main methanogens present (supplementary file 6 of chapter 5). This basically leads to merged pseudo-organisms, which have a bigger metabolic potential than any real organism. While one could argue that these organisms potentially inhabit the same niche and have the same interdependencies, which would obliterate the problem, this is clearly not the case for some niches in the gut, as they are inhabited by multiple strains of the same species, which are dependent on each other [79, 118]. Therefore a lumped pseudo-organism for such a species (and even worse for higher order classifications like genus or family) would not be an accurate representation of the actual community architecture.

It should not be overlooked that metabolic modelling can be successful, also in this thesis. In chapter 3 a GSMM for *Romboutsia ilealis* was developed. While the resulting model is in big parts for sure inaccurate, and it is also only a single organism model, it does give correct results (at least qualitative) for the measures of interest, such as the production of short chain fatty acids (SCFAs).

Nevertheless, the various challenges discussed above add to the well-known issues with GSMMs [473] (e.g. being focused on optimizing only a single goal like growth, or disregarding non-metabolic processes), and would make a simulation of any of the studied biomes rather inaccurate. Therefore the results of other studies in this field (e.g. the ones mentioned earlier) need to be carefully evaluated to determine their actual usefulness and accuracy.

It's a biome, let's sequence it

Another issue is not necessarily related to this thesis, but is noticeable in microbiome literature. This is poorly thought-through study set-ups, which lack rigorous design and focus beyond the wish to sequence a biome. While the microbiome field is not yet anymore the ultimate hype topic as it was after the revelation that the microbiome contributes to obesity [28], the topic is still trending enough that more research groups are trying to get on the hype train. This seems to lead in some cases to the selection of an exotic microbiome (a purely subjective example being the study of the vaginal microbiota in wild baboons [493]), with nothing more than the outcome that two subgroups within that microbiome are different or exhibit a pattern. While fundamental research is important in all areas (including in these exotic microbiomes), it is not supposed to stop purely with a descriptive result. In many cases the outcomes are not pursued further, and end with the sequence-based description of a given biome, which does not lead to further mechanistic insights. This issue is tightly coupled to the issue of finding significant results, which are testable. Many studies will report that a number of organisms are differentially abundant when e.g. comparing two study populations, or the effect of a certain intervention. Due to the very nature of the microbiome, with its hundreds to thousands of different organisms, this will in most cases be a significant number even when corrected for multiple testing. This first poses the problem which organisms to select for downstream hypothesis testing, since many are probably just correlated to some factor and not the cause of the observed differences. The second problem is that their impact cannot always easily be tested. The existing *in vitro* fermentation systems, like e.g. SHIME-2 [263], offer not only many opportunities, but also have specific limitations, which makes them unfit for some purposes. Testing for e.g. colonization resistance [494] or the impact on host metabolism [77, 78] are not easily testable in *in vitro* systems that lack a host component. Working with mouse or other animal models poses not only ethical issues, but also logistic and monetary ones. The work with human volunteers is even more complicated, due to the involved bureaucracy (lengthiness of the process to obtain ethical consent and recruiting volunteers), compliance of the study subjects to the given tasks (e.g. diet), or challenges to obtain the material of interest (e.g. duodenum samples), although they are also performed [495].

While in some cases study designs are getting more sophisticated, with the collection of a large number of samples including comprehensive sets of associated metadata (e.g. [229]), this yields the question if the field of microbiome research will not at some point be degraded into a purely association based type of research, as e.g. the field of genome wide association studies (GWAS) has become. GWAS is based on collecting a large amount of genomic data, and finding genomic correlations to a specific phenotype [496]. While there might be potential, depending on the result, to investigate the underlying mechanism, the main goal of the field is to find associations, and only a limited number of researchers actually asks the question of "correlation or causation". To this end, recent developments of more sensitive *in vitro* systems that aim at integrating aspects of host- and microbiome functioning such as organoids [497] or gut-on-a-chip [264, 498, 499] have the potential to prevent that the main outcome of further microbiome studies is a bar chart with multiple asterisks, and more and more researchers seem to realize this too [500].

Reproducibility and best practices in microbiome research

The crisis in reproducibility of research has in recent years received increasing attention [501-503]. In some fields like psychology this seems to be a dominant problem [504]. While countermeasures are being taken in biology, and in the microbiome field, they are not yet sufficient.

The one step in which this field, together with other sequencing related fields (WGS, GWAS, etc.), is an example for other research fields, is open data. Many publishers require that if sequencing has been performed in the course of a given research, that this data is also made publicly available in a way that the data is findable, accessible, interoperable and reusable (FAIR) [505]. In most cases there are no sensible reasons to not do so, since the researcher already benefited from it and often no further research is planned on exactly this data. This leverages the hurdle for attempting to reproduce published results and increases the barrier for fraud. Sadly, not everyone adheres to this as was summarized in a recent editorial based on frequently encountered issues [506], where the data was not made available, and obtaining it posed such a major hurdle that the researchers gave up in the end. While this currently only applies to a minority of studies and datasets, an efficient way needs to be found to shame such behaviour, and to prevent a further loss of trust in science.

This problem was also encountered in the course of a side-project that I was involved in. Prior to the publication of the genome of the organohalide respiring *Desulfitobacterium hafniense* strain PCE-S [507], it was attempted to verify results previously described for the closely related *D. hafiense* strain Y-51 [508], due to the fact that in their genome assembly (full chromosome) the potential technical contamination phiX phage was integrated. Since this genome was used for reference based scaffolding, this also resulted in the integration of the phage genome in our genome. Verification could have been achieved by investigating the mapping of the reads to the chromosome and investigate the boundaries of the integrated phage. Unfortunately, because the authors did not publicly deposit their reads, this was not possible, and they did also not react to an attempt to contact them. The EBI also contacted me later, because someone else noticed the integrated phage in our genome, and wanted to point out an error. The only option I had was to explain the situation, and that right now I had to assume that the observed phage integration is actual biology, and not a technical issue. Without access to raw data, there is no way to ensure this without going back to the laboratory, which is not necessarily always possible.

On multiple other occasions during the course of the research described in this thesis, it was also attempted to get hold of published 16S rRNA datasets, just to find out that the authors did not follow best practices and did not upload their data, which then did not allow us to investigate if our target organism (*Romboutsia ilealis,* chapter 3) is present in human intestinal samples. It was noticeable that in some cases this might have stemmed from time pressure, forgetfulness and bad organization. This is apparent when e.g. a bioproject and sample accession numbers are cited, but no actual runs (the actual data) are attached to these. This potentially happened because someone wanted to submit a manuscript at a defined deadline, but forgot to upload their data prior to that. While this is not an issue, most people are not aware that after uploading data to the ftp of one of the institutions belonging to the international nucleotide sequencing database collaboration (INSDC, including among others the European, US American and Japanese repositories EBI, NCBI and DDBJ), it will take until the next day until it is possible to submit the data as a run. Therefore it happens that people want to submit a manuscript, but get told that it will take until the next day until the accession numbers are available. Then they opt to cite the readily available bioproject or sample accession IDs instead, and then potentially forget about uploading the actual data. While this might seem to be rather constructed, it has happened to me as well. Luckily I always remembered to upload the runs the next day. It can now be argued whether this is bad planning from my side, by not uploading the data when it was easily possible, or bad planning from the authors side, who waited with a critical step until less than 24 hours before the manuscript submission. Either way, it makes it understandable and applaudable that some journals, e.g. the journal "Genome Announcements", have as policy that specific identifiers need to be cited, and will reject a manuscript otherwise (as experienced during the submission process of [509], although due to different reasons).

But even if all the data is available, there are still multiple issues which can impair the reproducibility of research. In many occasions not enough details are provided. Authors forget version numbers and parameters of the tools they used, which can make any attempts to reproduce the same results a futile exercise. Sometimes even complicated custom tools are developed and not published, although services like e.g. Github, https://github.com/, in combination with Zenodo, https://zenodo.org/, would easily allow to do this. The author of this thesis is also guilty of this. I developed at least three complicated tools during the time of this PhD. None of these tools are currently publicly available, although other researchers might benefit from them. The main issue in this case is time. While developing software is not always complicated, developing good software is. Some of my software is not in a state to be easily used by others, and the time to make the required changes to be user friendly is just not there. This is not even considering that e.g. the script used to produce the supplementary figures in chapter 3 is based on older code of mine, and the code itself does not follow any coding conventions due to the early, incremental and unplanned development. Sanitizing this code to make it not an embarrassment for any computer scientist is nearly impossible. Taken these two factors together, there was no chance for these scripts to be made public. I am aware that I am not following best practices, but I am not able to see how to change this, without having an impact on the direct scientific output, due to the time it would take up.

The next point after this issue is that not all reproduction attempts are feasible. In this thesis, e.g. in chapter 4, the metatranscriptome assembly lasted multiple weeks on a server with 320GB of RAM. During my current postdoctoral research, another assembly required 1.5TB of RAM. While research groups I worked in were lucky enough to have these resources, others will not, and therefore will not be able to reproduce it. And even if the resources are available, some of the used tools (like the genome assemblers) contain inherently stochastic processes, like e.g. picking the seed read for the beginning of the assembly, which will make the exact reproduction impossible. While for sure intermediate data could be provided, and in chapter 5 we also attempted to do so in response to a request by the editor during the first submission of the manuscript, there is no dedicated infrastructure for this kind of data available. The current INSDC repositories [510] do not really offer this, and it partially would also require substantial investments in more hard- and software-resources to do, especially for the big data like e.g. the meta-transcriptome assemblies generated in the framework of the research described in this thesis. Some data can be distributed as supplementary material during the publication process, but this obviously has not been standardized or structured, and even there e.g. bigger count tables could pose a problem.

Even if some kind of standardization is performed, it is necessary to make it also meaningful. In the course of my PhD research, one publication was submitted to the journal “Standards in Genomic Sciences” [507], and another one is in preparation [511]. In this journal extended genome announcements are published, and the editors set as a goal to make them comparable by introducing standards. This includes e.g. fixed tables and fixed figures, which help the standardization process, since in theory they can easily be compared. But this is only in theory the case. The journal e.g. requires a table of how many proteins in the genome of interest could be assigned to COG categories. In theory this allows then to judge if the genetic makeup of the different organisms are comparable or not. But the journal does not give any recommendations how to obtain the COGs. They could be obtained by the procedure as described in the original paper [192], by single directional blast, bidirectional blast [512], and with any kind of e-value or bitscore cutoff, making the whole attempt of having a standardized table less meaningful. On the other hand, the choice for standardization or allowing methodological flexibility is ultimately important for the spirit of science. The choice of tools, parameters and used setups are in many cases discussable. Most often there will be cases where the standard approach will fail due to the underlying biological problem. Standardization is therefore not necessarily impossible, but sometimes just not desirable for good scientific practice. It might even supress other scientific ideas. What e.g. if the journals required specific tools to be used (e.g. genome assemblers in the case of “Standards in Genomic Sciences”)? The development of better approaches in this area might be considerable hindered by standardization measures. Science is not a "one size fits all" business, which makes standardization an extremely complicated topic.

If these problems (except the last one discussed) were all solved, this would only affect the computational side of the research. Reproducing results on the laboratory side would not be affected by these efforts. Laboratory work was never in the scope of this work, and therefore can only be shallowly reflected upon, but progress needs also to be made on this side to ensure good scientific conduct.

What to do if science is only partially bad?

Issues with reproducibility, up to the level of scientific fraud, are not necessary for a publication to be deemed "bad science". But what should be done about it?

For example, during the literature research for chapter 5, the paper by Ross *et al*. [513] was investigated. In this study, the authors also studied rumen microbiota. They had a dual approach, by sequencing deeply with Illumina technology, and doing shallow sequencing with 454 technology. Afterwards the authors annotated the longer 454 reads, and mapped the Illumina reads to the 454 reads to do the statistical analysis (and to another dataset from the JGI). I am speculating that this was done due to insufficient computational capabilities to perform a transcriptome assembly, and therefore this strategy was chosen, although it does not seem directly logical. This is not the point to be mainly criticized. The criticisable problem in this publication is shown in table 1. The data in this table shows the mapping rate of the Illumina data to the 454 data, which is at the end ~6% of the data. Hence, the main outcome of this paper is based on less than 6% of the data, since not all the data did get analysed. The research is also not fraudulent, since all the details are given in the publication, and it is up to the readers and reviewers to judge it. But this cannot be considered good science, because obviously the majority of available information was discarded in this process. Therefore it was opted to not cite this publication in chapter 5. But what should be done about it? Citing it and pointing out the issue will increase the impact of the publication due to higher citation count, without having a negative impact, since not all researchers will read all the citing publications. Not citing will not change the issue. Using other channels, like e.g. blogs, Twitter, etc., can work, but only in limited cases when the backlash due to bad science turns out to be overwhelming (e.g. [514], although this case is more bad PR than bad science), but this will not be effective in most cases. Some publishers have comment sections on their websites (e.g. Frontiers or bioRxiv), which allow researchers to point out these problems, but also there the impact is limited.

Another case, which is more problematic, is the publication from Barkin *et al*. [515], over which I came across during my postdoctoral work. In this publication, the authors screen for a specific antibiotic resistance by performing a PCR for a specific marker gene. The issue here is that the marker gene has been proven to confer antibiotic resistance in *Bacteroides fragilis* [516], but the authors intended to screen for *Clostridium difficile*, where this association has never been made before, but where the gene is also ubiquitously present, despite the fact that the antibiotic resistance is not widespread. Additionally the authors only screened stool samples, which have been found positive for *Clostridium difficile*. The authors would have realized, if they had performed actual testing for antibiotic resistance, that they would not have been able to isolate any resistant *C. difficile* from these samples, and that the gene itself could have originated from other organisms in these samples too. This publication is full of oversight from the authors and the reviewers, and is speculative at best. It is not fraudulent, but nevertheless a problematic case, which should not stay published as it is. But what could be done? Should the publisher be contacted? If so, there is no established way of doing this, and it might prove to be very inefficient. Contacting the authors will most likely also be futile. Writing a commentary could be an option, but not all journals offer this option either.

Overall, this shows that science has considerable issues, which are not directly related to fraudulent activity, but which also have not been fully tackled yet.

Challenges in evaluating meta-omics data

The biggest technical challenge in meta-omics data is the sheer amount of data. To give a very simple example, data obtained by differential transcriptome analysis of a single bacterial species can be sufficiently well evaluated in an excel table with the help of some other tools like Pathway Tools [243], since the amount of differentially expressed genes might go into the hundreds, but will rarely exceed 2000. In contrast, this becomes a struggle if there are hundreds of organisms in the sample, which might lead to tens of thousands of differentially expressed genes, as it is the case even for relatively simple microbial communities such as those active during *in vitro* fermentation experiments (chapter 4, Table S2). Given that the outcome of different experiments can vary extensively, there is also rarely ever a universal solution, and the approach itself needs to be evaluated after every intermediate result.

A simple start can be to evaluate the metabolism, since sufficient databases such as KEGG [227] and Metacyc [517] exist, and the network layout of metabolism itself helps greatly with understanding also bigger datasets. But further steps need to be taken. Interpreting the entire community metabolism as a single entity will obscure differences between organisms. This can be seen e.g. in chapter 5, Figure 3, where the differences in the vitamin B12 metabolism between members of the *Bacteroides* and *Clostridiales* group led to actually interesting differences in SCFA metabolism, which explained the observed measurements in succinate and propionate. Staying with this example, it can also be shown that only explaining metabolism is often not sufficient. In the diagram other interactions are included, which cannot directly be seen on the metabolic maps. In this case the interaction of vitamins (B12) and cofactors (iron), as well as their transport and storage (iron) are critical in explaining the different SCFA measurements. While in this case these partial results were simply obtained by manually inspecting the excel table with sharp eyes, in other instances this will neither be sufficient nor feasible. There are also no readily available tools, which would automatically allow a researcher to do this, and there might also in the near future not be any, due to the complexity of the data. Using substitutes, like the evaluation of over- and under-representation of GO terms [197] does greatly help in these scenarios, however, the final connections still often need to be established manually, either again by visual inspection, or by summarizing data in custom ways (e.g. counting specific functions per organism).

Challenges for the wet-lab researcher

This also makes the requests by wet lab scientists a nearly futile challenge. It is desirable (and often necessary) that a scientist can evaluate his or her own data, and in other fields, like physics, this is also often given. But in biology, many scientists do not have the necessary computational skills to perform these steps. Therefore currently the job of a computational biologist exists, in contrast to a computational physicist, which is often not a standalone job description.

With no standard solutions being available (also not for the computational biologist), the scientist him/herself must be able to improvise. This skill for improvisation is actually one of the most necessary skills for a scientist. This becomes obvious, if in a laboratory setting standard protocols for e.g. DNA extraction are not working, and further steps (centrifugation, other buffers, etc.) must be taken. If the standard protocols would always work, a scientist would not be necessary in this setting, and could easily be replaced by a robot or by cheaper, untrained personnel. Due to the fact that science is highly variable, this is not the case, and it is necessary to have an understanding of the underlying mechanisms, to make improvisation possible.

But without computational skills, improvisation is not possible in big data, and a researcher will not be able to fully understand his/her own data. While this has been clear now since the start of this millennium [518], the education systems still struggle with it [519], although it should be noted that due to the rising awareness [520] this situation is hopefully likely to change in the near future.

Future perspectives

Talking about the future - the field of microbiome research has been moving forward, but which direction is it going to take? There probably will be a trend towards more comprehensive studies, with more in-depth data and increasing efforts to provide causal findings rather than mere associations. This is not necessarily motivated by the field itself, since there is currently an expansion of the "just sequence it all" approach (as mentioned earlier). Luckily, there is increasing awareness in the field that large scale data-driven explorative research is needed for hypothesis finding, but should not be the endpoint. After the initial large scale surveying projects (the Human Microbiome Project [19] and MetaHit [27]), other similar projects with even more data are already being established [229]. While projects of this size are obviously only possible for very large research groups and collaborative programmes, the collection of more diverse data, being it metadata, physiological, immunological or chemical data, should be part of every future project. Furthermore, as also pointed out earlier, more sophisticated follow-up experiments are necessary. While currently many publications end with correlating values, more projects will hopefully change this, by investigating deeper the causative reasons of their findings (despite the associated challenges).

Another direction is probably the usage and development of more standardized pipelines. While this is not always beneficial (as pointed out earlier), the mentioned best practices problems will probably force the field in this direction. The availability and implementation of standardized pipelines will make reproducibility of results a negligible issue, due to logging functions and easily retrievable parameters. In turn, this will possibly also lead to scientific losses in certain datasets, since with standard approaches not every potentially interesting result can be retrieved. Re-investigations of bigger datasets by trained data scientists will therefore probably also become the norm, however, requires free access to such data in the public domain as discussed above.

We have been overwhelmed by this vast amount of data, generated by cheap sequencing technologies. If this continues, larger research groups will sequence even more, and maybe so much, that the field will be confronted with a big pile of non-analysed data. This can partially be seen in different publications, where e.g. metagenomics data had been sequenced, but only the standard approaches established for 16S rRNA gene amplicon analysis were used, without utilizing the additional information in this data. While these occasions are not prevalent and will for sure get less, it is possible that at some point the sequencing will be so fast and cheap that it will not be possible anymore to keep up with it. This point will probably appear only after the mentioned standardization measures, and will then require the real experts in machine learning. While players like Google and Microsoft are already dealing with computational biology [521, 522], they will probably be investing in this field even more in the coming years.

Concluding remarks

The research described in this thesis was divers, ranging from human, rat to cow gastrointestinal biomes, with several more side projects. There are two main messages to be taken home from this work.

The first one is that the microbiota is involved in nearly all processes that involve higher animals. It impacts digestion (chapter 1), skin health [523], oral health [524] and maybe mental health [525], as well as feed performance [34], gas output (chapter 5, [42]) and many more parameters. The second is that, while the field is now in its teen years, and approaching twenty years (chapter 2), there are still lots of unsolved problems, starting from various laboratory methods to computational approaches (this thesis, and e.g. [526]), to reproducibility issues and data distribution. Overall, this combination offers many challenges but also almost unlimited opportunities for future microbiome research.

References

1. **What is Systems Biology** [https://www.systemsbiology.org/about/what-is-systems-biology/]
2. Kitano H: **Systems Biology: a Brief Overview**. *Science* 2002, **295**:1662-1664.
3. Katagiri F: **Attacking complex problems with the power of systems biology**. *Plant physiology* 2003, **132**(2):417-419.
4. Keele KD: **Leonardo da Vinci as Physiologist**. *Postgrad Med J* 1952, **28**(324):521-528.
5. Coffey JC, O'Leary DP: **The mesentery: structure, function, and role in disease**. *The Lancet Gastroenterology & Hepatology* 2016, **1**(3):238-247.
6. Benias PC, Wells RG, Sackey-Aboagye B, Klavan H, Reidy J, Buonocore D, Miranda M, Kornacki S, Wayne M, Carr-Locke DL *et al*: **Structure and Distribution of an Unrecognized Interstitium in Human Tissues**. *Scientific reports* 2018, **8**(1):4947.
7. Kormondy EJ: **A Brief Introduction to the History of Ecology**. *The American Biology Teacher* 2012, **74**(7):441-443.
8. Bruggeman FJ, Westerhoff HV: **The nature of systems biology**. *Trends in microbiology* 2007, **15**(1):45-50.
9. Yoshida T, Jones LE, Ellner SP, Fussmann GF, Hairston Jr NG: **Rapid evolution drives ecological dynamics in a predator-prey system**. *Nature* 2003, **424**:303-306.
10. **Definition of Ecology** [http://www.caryinstitute.org/discover-ecology/definition-ecology]
11. Locey KJ, Lennon JT: **Scaling laws predict global microbial diversity**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **113**(21):5970-5975.
12. Staley JT, Konopka A: **measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats**. *Annual review of microbiology* 1985, **39**:321-346.
13. Boers SA, Hiltemann SD, Stubbs AP, Jansen R, Hays JP: **Development and evaluation of a culture-free microbiota profiling platform (MYcrobiota) for clinical diagnostics**. *Eur J Clin Microbiol Infect Dis* 2018, **37**(6):1081-1089.
14. Dworkin M: **Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist**. *FEMS microbiology reviews* 2012, **36**(2):364-379.
15. Chung K-T, Ferris DH: **Martinus Willem Beijerink (1851-1931) - pioneer of general microbiology**. *ASM News* 1996, **62**(10):539-543.
16. Castiglione F, Pappalardo F, Bianca C, Russo G, Motta S: **Modeling biology spanning different scales: an open challenge**. *BioMed research international* 2014, **2014**:902545.
17. Weinberg AM: **impact of large-scale science on the united states**. *Science* 1961, **134**(3473):161-164.
18. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The Sequence of the Human Genome**. *Science* 2001, **291**(5507):1304-1351.
19. Human Microbiome Project C: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.
20. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA**. *Genomics* 2016, **107**(1):1-8.
21. Pop M: **Genome assembly reborn: recent computational challenges**. *Briefings in bioinformatics* 2009, **10**(4):354-366.
22. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikh R *et al*: **assemblathon 2 - evaluating de novo methods of genome assembly in three vertebrate species**. *GigaScience* 2013, **2**(10).
23. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M *et al*: **GAGE: A critical evaluation of genome assemblies and assembly algorithms**. *Genome research* 2012, **22**(3):557-567.

24. Sender R, Fuchs S, Milo R: **Revised Estimates for the Number of Human and Bacteria Cells in the Body**. *PLoS biology* 2016, **14**(8):e1002533.

25. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al*: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**(5667):66-74.

26. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al*: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**(5721):554-557.

27. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464**(7285):59-65.

28. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest**. *Nature* 2006, **444**(7122):1027-1031.

29. Schubert AM, Rogers MA, Ring C, Mogle J, Petrosino JP, Young VB, Aronoff DM, Schloss PD: **Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls**. *mBio* 2014, **5**(3):e01021-01014.

30. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR: **Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(34):13780-13785.

31. van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, Visser CE, Kuijper EJ, Bartelsman JF, Tijssen JG *et al*: **Duodenal infusion of donor feces for recurrent *Clostridium difficile***. *The New England journal of medicine* 2013, **368**(5):407-415.

32. Petrof EO, Khoruts A: **From stool transplants to next-generation microbiota therapeutics**. *Gastroenterology* 2014, **146**(6):1573-1582.

33. Song SJ, Lauber C, Costello EK, Lozupone CA, Humphrey G, Berg-Lyons D, Caporaso JG, Knights D, Clemente JC, Nakielny S *et al*: **Cohabiting family members share microbiota with one another and with their dogs**. *eLife* 2013, **2**:e00458.

34. Li F, Guan LL: **Metatranscriptomic Profiling Reveals Linkages between the Active Rumen Microbiome and Feed Efficiency in Beef Cattle**. *Applied and environmental microbiology* 2017, **83**(9):e00061-00017.

35. den Besten G, van Eunen K, Groen AK, Venema K, Reijngoud DJ, Bakker BM: **The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism**. *Journal of lipid research* 2013, **54**(9):2325-2340.

36. Hook SE, Wright AD, McBride BW: **Methanogens: methane producers of the rumen and mitigation strategies**. *Archaea* 2010, **2010**:945785.

37. Kittelmann S, Pinares-Patino CS, Seedorf H, Kirk MR, Ganesh S, McEwan JC, Janssen PH: **Two different bacterial community types are linked with the low-methane emission trait in sheep**. *PloS one* 2014, **9**(7):e103171.

38. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, Eijsink VG: **Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci**. *PloS one* 2012, **7**(6):e38571.

39. Taras D, Vahjen W, Simon O: **Probiotics in pigs — modulation of their intestinal distribution and of their impact on health and performance**. *Livestock Science* 2007, **108**(1-3):229-231.

40. Simmons CW, Reddy AP, D'Haeseleer P, Khudyakov J, Billis K, Pati A, Simmons BA, Singer SW, Thelen MP, VanderGheynst JS: **Metatranscriptomic analysis of lignocellulolytic microbial communities involved in high-solids decomposition of rice straw**. *Biotechnology for biofuels* 2014, **7**:495.

41. Singh KM, Reddy B, Patel D, Patel AK, Parmar N, Patel A, Patel JB, Joshi CG: **High potential source for biomass degradation enzyme discovery and environmental aspects revealed through metagenomics of Indian buffalo rumen**. *BioMed research international* 2014, **2014**:267189.

42. Güllert S, Fischer MA, Turaev D, Noebauer B, Ilmberger N, Wemheuer B, Alawi M, Rattei T, Daniel R, Schmitz RA *et al*: **Deep metagenome and metatranscriptome analyses of microbial communities affiliated with an industrial biogas fermenter, a cow rumen, and elephant feces reveal major differences in carbohydrate hydrolysis strategies**. *Biotechnology for biofuels* 2016, **9**:121.

43. Shiffman ME, Soo RM, Dennis PG, Morrison M, Tyson GW, Hugenholtz P: **Gene and genome-centric analyses of koala and wombat fecal microbiomes point to metabolic specialization for Eucalyptus digestion**. *PeerJ* 2017, **5**:e4075.

44. He S, Ivanova N, Kirton E, Allgaier M, Bergin C, Scheffrahn RH, Kyrpides NC, Warnecke F, Tringe SG, Hugenholtz P: **Comparative metagenomic and metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding higher termites**. *PloS one* 2013, **8**(4):e61126.

45. Yamamuro A, Kouzuma A, Abe T, Watanabe K: **Metagenomic analyses reveal the involvement of syntrophic consortia in methanol/electricity conversion in microbial fuel cells**. *PloS one* 2014, **9**(5):e98425.

46. Mardanov AV, Bulygina ES, Nedoluzhko AV, Kadnikov VV, Beletskii AV, Tsygankova SV, Tikhonov AN, Ravin NV, Prokhorchuk EB, Skryabin KG: **Molecular analysis of the intestinal microbiome composition of mammoth and woolly rhinoceros**. *Doklady Biochemistry and biophysics* 2012, **445**:203-206.

47. Tito RY, Knights D, Metcalf J, Obregon-Tito AJ, Cleeland L, Najar F, Roe B, Reinhard K, Sobolik K, Belknap S *et al*: **Insights from characterizing extinct human gut microbiomes**. *PloS one* 2012, **7**(12):e51146.

48. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, Bradshaw CJ, Townsend G, Soltysiak A, Alt KW *et al*: **Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions**. *Nature genetics* 2013, **45**(4):450-455, 455e451.

49. Checinska A, Probst AJ, Vaishampayan P, White JR, Kumar D, Stepanov VG, Fox GE, Nilsson HR, Pierson DL, Perry J *et al*: **Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities**. *Microbiome* 2015, **3**:50.

50. Mayer T, Blachowicz A, Probst AJ, Vaishampayan P, Checinska A, Swarmer T, de Leon P, Venkateswaran K: **Microbial succession in an inflated lunar/Mars analog habitat during a 30-day human occupation**. *Microbiome* 2016, **4**:22.

51. Saffary R, Nandakumar R, Spencer D, Robb Frank T, Davila Joseph M, Swartz M, Ofman L, Thomas Roger J, DiRuggiero J: **Microbial survival of space vacuum and extreme ultraviolet irradiation: strain isolation and analysis during a rocket flight**. *FEMS microbiology letters* 2006, **215**(1):163-168.

52. Jonsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P: **Tardigrades survive exposure to space in low Earth orbit**. *Curr Biol* 2008, **18**(17):R729-R731.

53. **The Grand Finale Toolkit** [https://saturn.jpl.nasa.gov/mission/grand-finale/overview/]

54. Anesio AM, Lutz S, Chrismas NAM, Benning LG: **The microbiome of glaciers and ice sheets**. *NPJ Biofilms Microbiomes* 2017, **3**:10.

55. Meirelles LA, McFrederick QS, Rodrigues A, Mantovani JD, de Melo Rodovalho C, Ferreira H, Bacci M, Jr., Mueller UG: **Bacterial microbiomes from vertically transmitted fungal inocula of the leaf-cutting ant *Atta texana***. *Environ Microbiol Rep* 2016, **8**(5):630-640.

56. Vavourakis CD, Ghai R, Rodriguez-Valera F, Sorokin DY, Tringe SG, Hugenholtz P, Muyzer G: **Metagenomic Insights into the Uncultured Diversity and Physiology of Microbes in Four Hypersaline Soda Lake Brines**. *Frontiers in microbiology* 2016, **7**:211.

57. Sanchez-Andrea I, Rodriguez N, Amils R, Sanz JL: **Microbial diversity in anaerobic sediments at Rio Tinto, a naturally acidic environment with a high heavy metal content**. *Applied and environmental microbiology* 2011, **77**(17):6085-6093.

58. Baker BJ, Banfield JF: **Microbial communities in acid mine drainage**. *FEMS microbiology ecology* 2003, **44**(2):139-152.

59. Cardinale M, Kaiser D, Lueders T, Schnell S, Egert M: **Microbiome analysis and confocal microscopy of used kitchen sponges reveal massive colonization by *Acinetobacter*, *Moraxella* and *Chryseobacterium* species**. *Scientific reports* 2017, **7**(1):5791.

60. Gibson GR, Cummings JH, MacFarlane GT: **Use of a Three-Stage Continuous Culture System to Study the Effect of Mucin on Dissimilatory Sulfate Reduction and Methanogenesis by Mixed Populations of Human Gut Bacteria**. *Applied and environmental microbiology* 1988, **54**(11):2750-2755.

61. Macy JM, Probst I: **The biology of gastrointestinal bacteroides**. *Annual Reviews in Microbiology* 1979, **33**:561-591.

62. van Gastelen S, Antunes-Fernandes EC, Hettinga KA, Klop G, Alferink SJ, Hendriks WH, Dijkstra J: **Enteric methane production, rumen volatile fatty acid concentrations, and milk fatty acid composition in lactating Holstein-Friesian cows fed grass silage- or corn silage-based diets**. *Journal of dairy science* 2015, **98**(3):1915-1927.

63. IPCC: **Climate change 2007: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change**. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press; 2007.

64. McMichael AJ, Powles JW, Butler CD, Uauy R: **Food, livestock production, energy, climate change, and health**. *The Lancet* 2007, **370**(9594):1253-1263.

65. Hadhazy A: **Think Twice: How the Gut's "Second Brain" Influences Mood and Well-Being**. In: *Scientific American.* Nature America, Inc.; 2010.

66. Brown H: **The Other Brain Also Deals With Many Woes**. In: *The New York Times.* The New York Times Company 2005.

67. Berasategui A, Shukla S, Salem H, Kaltenpoth M: **Potential applications of insect symbionts in biotechnology**. *Applied microbiology and biotechnology* 2016, **100**(4):1567-1577.

68. O'Hara AM, Shanahan F: **The gut flora as a forgotten organ**. *EMBO reports* 2006, **7**(7):688-693.

69. Marteau P, Pochart P, Dore J, Bera-Maillet C, Bernalier A, Corthier G: **Comparative Study of Bacterial Groups within the Human Cecal and Fecal Microbiota**. *Applied and environmental microbiology* 2001, **67**(10):4939-4942.

70. Zoetendal EG, Raes J, van den Bogert B, Arumugam M, Booijink CC, Troost FJ, Bork P, Wels M, de Vos WM, Kleerebezem M: **The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates**. *The ISME journal* 2012, **6**(7):1415-1426.

71. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E: **Microbial degradation of complex carbohydrates in the gut**. *Gut Microbes* 2012, **3**(4):289-306.

72. Gerritsen J, Smidt H, Rijkers GT, de Vos WM: **Intestinal microbiota in human health and disease: the impact of probiotics**. *Genes & nutrition* 2011, **6**(3):209-240.

73. Fofanova TY, Petrosino JF, Kellermayer R: **Microbiome-Epigenome Interactions and the Environmental Origins of Inflammatory Bowel Diseases**. *J Pediatr Gastroenterol Nutr* 2016, **62**(2):208-219.

74. Lopez J, Grinspan A: **Fecal Microbiota Transplantation for Inflammatory Bowel Disease**. *Gastroenterology & Heaptology* 2016, **12**(6):374-379.

75. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO *et al*: **Vaginal microbiome of reproductive-age women**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108 Suppl 1**:4680-4687.

76. van Passel MW, Kant R, Zoetendal EG, Plugge CM, Derrien M, Malfatti SA, Chain PS, Woyke T, Palva A, de Vos WM *et al*: **The genome of *Akkermansia muciniphila*, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes**. *PloS one* 2011, **6**(3):e16876.

77. Roediger WEW: **role of anaerobic bacteria in the metabolic welfare of the colonic mucosa in man**. *Gut microbes* 1980, **21**:793-798.

78. den Besten G, Lange K, Havinga R, van Dijk TH, Gerding A, van Eunen K, Muller M, Groen AK, Hooiveld GJ, Bakker BM *et al*: **Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids**. *Am J Physiol Gastrointest Liver Physiol* 2013, **305**(12):G900-910.

79. Turroni F, Milani C, Duranti S, Mancabelli L, Mangifesta M, Viappiani A, Lugli GA, Ferrario C, Gioiosa L, Ferrarini A *et al*: **Deciphering bifidobacterial-mediated metabolic interactions and their impact on gut microbiota by a multi-omics approach**. *The ISME journal* 2016, **10**(7):1656-1668.

80. Riviere A, Gagnon M, Weckx S, Roy D, De Vuyst L: **Mutual Cross-Feeding Interactions between *Bifidobacterium longum* subsp. *longum* NCC2705 and *Eubacterium rectale* ATCC 33656 Explain the Bifidogenic and Butyrogenic Effects of Arabinoxylan Oligosaccharides**. *Applied and environmental microbiology* 2015, **81**(22):7767-7781.

81. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, Dorrestein PC, Esquenazi E, Hunter RC, Cheng G *et al*: **Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(1):244-249.

82. Dwidar M, Monnappa AK, Mitchell RJ: **The dual probiotic and antibiotic nature of *Bdellovibrio bacteriovorus***. *BMB reports* 2012, **45**(2):71-78.

83. Atterbury RJ, Hobley L, Till R, Lambert C, Capeness MJ, Lerner TR, Fenton AK, Barrow P, Sockett RE: **Effects of orally administered *Bdellovibrio bacteriovorus* on the well-being and *Salmonella* colonization of young chicks**. *Applied and environmental microbiology* 2011, **77**(16):5794-5803.

84. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R: **Diversity, stability and resilience of the human gut microbiota**. *Nature* 2012, **489**(7415):220-230.

85. Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, Dunn JR, Farley MM, Holzbauer SM, Meek JI, Phipps EC *et al*: **Burden of *Clostridium difficile* infection in the United States**. *The New England journal of medicine* 2015, **372**(9):825-834.

86. Stein RR, Bucci V, Toussaint NC, Buffie CG, Ratsch G, Pamer EG, Sander C, Xavier JB: **Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota**. *PLoS computational biology* 2013, **9**(12):e1003388.

87. Voth DE, Ballard JD: ***Clostridium difficile* toxins: mechanism of action and role in disease**. *Clinical microbiology reviews* 2005, **18**(2):247-263.

88. Grundy SM: **A Constellation of Complicatons: The Metabolic Syndrome**. *Clinical Cornerstone* 2005, **7**(2/3):36-45.

89. O'Neill S, O'Driscoll L: **Metabolic syndrome: a closer look at the growing epidemic and its associated pathologies**. *Obes Rev* 2015, **16**(1):1-12.

90. Xia Q, Grant SF: **The genetics of human obesity**. *Annals of the New York Academy of Sciences* 2013, **1281**:178-190.

91. Swinburn BA, Caterson I, Seidell JC, James WPT: **Diet, nutrition and the prevention of excess weight gain and obesity**. *Public Health Nutrition* 2007, **7**(1a):123-146.

92. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S *et al*: **Richness of human gut microbiome correlates with metabolic markers**. *Nature* 2013, **500**(7464):541-546.

93. McNeil NI: **The contribution of the large intestine to energy supplies in man**. *The American journal of clinical nutrition* 1984, **39**:338-342.

94. Bergman EN: **Energy Contributions of Volatile Fatty Acids From the Gastrointestinal Tract in Various Species**. *Physiol Rev* 1990, **70**(2):567-590.

95. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI: **A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis**. *Science* 2003, **299**(5615):2074-2076.

96. Lammerts van Bueren A, Saraf A, Martens EC, Dijkhuizen L: **Differential Metabolism of Exopolysaccharides from Probiotic Lactobacilli by the Human Gut Symbiont Bacteroides thetaiotaomicron**. *Applied and environmental microbiology* 2015, **81**(12):3973-3983.

97. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI: **the effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice**. *Genetics and Diet* 2009, **1**(6):6ra14.

98. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, Griffin NW, Lombard V, Henrissat B, Bain JR *et al*: **Gut microbiota from twins discordant for obesity modulate metabolism in mice**. *Science* 2013, **341**(6150):1241214.

99. Minihane AM, Vinoy S, Russell WR, Baka A, Roche HM, Tuohy KM, Teeling JL, Blaak EE, Fenech M, Vauzour D *et al*: **Low-grade inflammation, diet composition and health: current research evidence and its translation**. *The British journal of nutrition* 2015, **114**(7):999-1012.

100. Chassaing B, Gewirtz AT: **Has provoking microbiota aggression driven the obesity epidemic?** *Bioessays* 2016, **38**(2):122-128.

101. Chassaing B, Gewirtz AT: **Gut microbiota, low-grade inflammation, and metabolic syndrome**. *Toxicologic pathology* 2014, **42**(1):49-53.

102. MacIver NJ, Jacobs SR, Wieman HL, Wofford JA, Coloff JL, Rathmell JC: **Glucose metabolism in lymphocytes is a regulated process with significant effects on immune cell function and survivial**. *J Leukoc Biol* 2008, **84**(4):949-957.

103. Grundy SM, Brewer HB, Jr., Cleeman JI, Smith SC, Jr., Lenfant C, American Heart A, National Heart L, Blood I: **Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition**. *Circulation* 2004, **109**(3):433-438.

104. Al Khodor S, Reichert B, Shatat IF: **The Microbiome and Blood Pressure: Can Microbes Regulate Our Blood Pressure?** *Front Pediatr* 2017, **5**:138.

105. Durgan DJ, Ganesh BP, Cope JL, Ajami NJ, Phillips SC, Petrosino JF, Hollister EB, Bryan RM, Jr.: **Role of the Gut Microbiome in Obstructive Sleep Apnea-Induced Hypertension**. *Hypertension* 2016, **67**(2):469-474.

106. Schiffrin EL: **Immune mechanisms in hypertension and vascular injury**. *Clin Sci (Lond)* 2014, **126**(4):267-274.

107. Pluznick JL, Protzko RJ, Gevorgyan H, Peterlin Z, Sipos A, Han J, Brunet I, Wan LX, Rey F, Wang T *et al*: **Olfactory receptor responding to gut microbiota-derived signals plays a role in renin secretion and blood pressure regulation**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(11):4410-4415.

108. Ghazalpour A, Cespedes I, Bennett BJ, Allayee H: **Expanding role of gut microbiota in lipid metabolism**. *Curr Opin Lipidol* 2016, **27**(2):141-147.

109. Joyce SA, MacSharry J, Casey PG, Kinsella M, Murphy EF, Shanahan F, Hill C, Gahan CG: **Regulation of host weight gain and lipid metabolism by bacterial bile acid modification in the gut**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(20):7421-7426.

110. Fu J, Bonder MJ, Cenit MC, Tigchelaar EF, Maatman A, Dekens JA, Brandsma E, Marczynska J, Imhann F, Weersma RK *et al*: **The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids**. *Circ Res* 2015, **117**(9):817-824.

111. Raza GS, Putaala H, Hibberd AA, Alhoniemi E, Tiihonen K, Makela KA, Herzig KH: **Polydextrose changes the gut microbiome and attenuates fasting triglyceride and cholesterol levels in Western diet fed mice**. *Scientific reports* 2017, **7**(1):5294.

112. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA *et al*: **Insights into the phylogeny and coding potential of microbial dark matter**. *Nature* 2013, **499**(7459):431-437.

113. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, Bittar F, Fournous G, Gimenez G, Maraninchi M *et al*: **Microbial culturomics: paradigm shift in the human gut microbiome study**. *Clin Microbiol Infect* 2012, **18**(12):1185-1193.

114. Moissl-Eichinger C, Huber H: **Archaeal symbionts and parasites**. *Current opinion in microbiology* 2011, **14**(3):364-370.

115. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI: **Viruses in the faecal microbiota of monozygotic twins and their mothers**. *Nature* 2010, **466**(7304):334-338.

116. Parfrey LW, Walters WA, Knight R: **Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions**. *Frontiers in microbiology* 2011, **2**:153.

117. Samuel BS, Gordon JI: **A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(26):10011-10016.

118. Turroni F, Ozcan E, Milani C, Mancabelli L, Viappiani A, van Sinderen D, Sela DA, Ventura M: **Glycan cross-feeding activities between bifidobacteria under in vitro conditions**. *Frontiers in microbiology* 2015, **6**:1030.

119. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A, Shah N, Wang C, Magrini V, Wilson RK *et al*: **Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(14):5859-5864.

120. de Vos WM: **Microbial biofilms and the human intestinal microbiome**. *npj Biofilms and Microbiomes* 2015, **1**:15005.

121. Elias S, Banin E: **Multi-species biofilms: living with friendly neighbors**. *FEMS microbiology reviews* 2012, **36**(5):990-1004.

122. Parsek MR, Greenberg EP: **Sociomicrobiology: the connections between quorum sensing and biofilms**. *Trends in microbiology* 2005, **13**(1):27-33.

123. Marchesi JR, Ravel J: **The vocabulary of microbiome research: a proposal**. *Microbiome* 2015, **3**:31.

124. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R *et al*: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences**. *Nature biotechnology* 2013, **31**(9):814-821.

125. Asshauer KP, Wemheuer B, Daniel R, Meinicke P: **Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data**. *Bioinformatics* 2015, **31**(17):2882-2884.

126. Iwai S, Weinmaier T, Schmidt BL, Albertson DG, Poloso NJ, Dabbagh K, DeSantis TZ: **Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes**. *PloS one* 2016, **11**(11):e0166104.

127. Rooijers K, Kolmeder C, Juste C, Dore J, de Been M, Boeren S, Galan P, Beauvallet C, De Vos WM, Schaap PJ: **In iterative workflow for mining the human intestinal metaproteome**. *BMC genomics* 2011, **12**(6).

128. Walker A, Pfitzner B, Neschen S, Kahle M, Harir M, Lucio M, Moritz F, Tziotis D, Witting M, Rothballer M *et al*: **Distinct signatures of host-microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet**. *The ISME journal* 2014, **8**(12):2380-2396.
129. El Aidy S, Derrien M, Merrifield CA, Levenez F, Dore J, Boekschoten MV, Dekker J, Holmes E, Zoetendal EG, van Baarlen P *et al*: **Gut bacteria-host metabolic interplay during conventionalisation of the mouse germfree colon**. *The ISME journal* 2013, **7**(4):743-755.
130. Wilmes P, Heintz-Buschart A, Bond PL: **A decade of metaproteomics: where we stand and what the future holds**. *Proteomics* 2015, **15**(20):3409-3417.
131. Nicholson JK, Lindon JC: **Metabonomics**. *Nature* 2008, **455**:1054-1056.
132. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies**. *Nature reviews Genetics* 2016, **17**(6):333-351.
133. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-Time DNA Sequencing from Single Polymerase Molecules**. *Science* 2009, **323**:133-138.
134. Deamer D, Akeson M, Branton D: **Three decades of nanopore sequencing**. *Nature biotechnology* 2016, **34**(5):518-524.
135. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VG, McHardy AC, Nederbragt AJ, Pope PB: **Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data**. *Scientific reports* 2016, **6**:25373.
136. Marshall CW, Ross DE, Fichot EB, Norman RS, May HD: **Electrosynthesis of commodity chemicals by an autotrophic microbial community**. *Applied and environmental microbiology* 2012, **78**(23):8412-8420.
137. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, Gies EA, Cheng JF, Copeland A, Klenk HP *et al*: **High-resolution phylogenetic microbial community profiling**. *The ISME journal* 2016, **10**(8):2020-2032.
138. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK: **Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system**. *PeerJ* 2016, **4**:e1869.
139. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J: **Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification**. *BMC microbiology* 2016, **16**(1):274.
140. Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J, Kaplan LA: **Improved performance of the PacBio SMRT technology for 16S rDNA sequencing**. *Journal of microbiological methods* 2014, **104**:59-60.
141. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM *et al*: **Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis**. *Genome medicine* 2015, **7**:99.
142. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO: **Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies**. *Nucleic acids research* 2013, **41**(1):e1.
143. Stackebrandt E, Goebel BM: **taxonomic note: a place for DNA-DNA reassociation and 16s rRNA sequence analysis in the prsent species definition in bacteriology**. *International journal of systematic and evolutionary microbiology* 1994, **44**(4):846-849.
144. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E *et al*: **report of the ad hoc committee on reconciliation of approaches to bacterial systematics**. *International journal of systematic and evolutionary microbiology* 1987, **37**(4):463-464.
145. van de Peer Y, Chapelle S, De Wachter R: **a quantitative map of nucleotide substitution rates in bacterial rRNA**. *Nucleic acids research* 1996, **24**(17):3381–3391.
146. Hermes GD, Zoetendal EG, Smidt H: **Molecular ecological tools to decipher the role of our microbial mass in obesity**. *Beneficial microbes* 2015, **6**(1):61-81.

147. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW: **Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions**. *Nucleic acids research* 2010, **38**(22):e200.

148. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding C, Fungal Barcoding Consortium Author L: **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(16):6241-6246.

149. McMurdie PJ, Holmes S: **waste not, want not: why rarefying microbiome data is inadmissable**. *PLoS computational biology* 2014, **10**(4):e1003531.

150. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools**. *Nucleic acids research* 2013, **41**(Database issue):D590-596.

151. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM: **Ribosomal Database Project: data and tools for high throughput rRNA analysis**. *Nucleic acids research* 2014, **42**(Database issue):D633-642.

152. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**. *Applied and environmental microbiology* 2006, **72**(7):5069-5072.

153. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H *et al*: **Sequence-specific error profile of Illumina sequencers**. *Nucleic acids research* 2011, **39**(13):e90.

154. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez Pena A, Goodrich JK, Gordon JI *et al*: **QIIME allows analysis of high-throughput community sequencing data**. *Nature methods* 2010, **7**(5):335-336.

155. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ *et al*: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities**. *Applied and environmental microbiology* 2009, **75**(23):7537-7541.

156. McMurdie PJ, Holmes S: **phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data**. *PloS one* 2013, **8**(4):e61217.

157. Albanese D, Fontana P, De Filippo C, Cavalieri D, Donati C: **MICCA: a complete and accurate software for taxonomic profiling of metagenomic data**. *Scientific reports* 2015, **5**:9743.

158. Ramiro-Garcia J, Hermes GDA, Giatsis C, Sipkema D, Zoetendal EG, Schaap PJ, Smidt H: **NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes**. *F1000Research* 2016, **5**:1791.

159. Faith D: **Conservation evaluation and phylogenetic diversity**. *Biological Conservation* 1992, **61**:1-10.

160. Shannon CE: **A Mathematical Theory of Communication**. *The Bell System Technical Journal* 1948, **27**:379–423, 623–656.

161. Simpson EH: **Measurement of Diversity**. *Nature* 1949, **163**:688.

162. Chiarucci A, Bacaro G, Scheiner SM: **Old and new challenges in using species diversity for assessing biodiversity**. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2011, **366**(1576):2426-2437.

163. Heip CHR, Herman PMJ, Soetart K: **Indices of diversity and evenness**. *Oceanis* 1998, **24**(4):61-87.

164. Chao A: **Nonparametric Estimation of the Number of Classes in a Population**. *Scandinavian Journal of Statistics* 1984, **11**:265-270.

165. Chao A, Lee S-M: **Estimating the Number of Classes via Sample Coverage**. *Journal of the American Statistical Association* 1992, **84**(417):210-217.

166. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM: **Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity**. *Applied and environmental microbiology* 2001, **67**(10):4399-4406.

167. Jaccard P: **The distribution of the flora in the alpine zone**. *New Phytologist* 1912, **XI**(2):37-50.

168. Bray JR, Curtis JT: **An Ordination of the Upland Forest Communities of Southern Wisconsin**. *Ecological Monographs* 1957, **27**(4):326-349.

169. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities**. *Applied and environmental microbiology* 2005, **71**(12):8228-8235.

170. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP *et al*: **A core gut microbiome in obese and lean twins**. *Nature* 2009, **457**(7228):480-484.

171. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(31):11070-11075.

172. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP *et al*: **Human gut microbiome viewed across age and geography**. *Nature* 2012, **486**(7402):222-227.

173. Walker AW, Ince J, Duncan SH, Webster LM, Holtrop G, Ze X, Brown D, Stares MD, Scott P, Bergerat A *et al*: **Dominant and diet-responsive groups of bacteria within the human colonic microbiota**. *The ISME journal* 2011, **5**(2):220-230.

174. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R *et al*: **Linking long-term dietary patterns with gut microbial enterotypes**. *Science* 2011, **334**(6052):105-108.

175. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X *et al*: **A survey of best practices for RNA-seq data analysis**. *Genome biology* 2016, **17**:13.

176. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM: **An extensive evaluation of read trimming effects on Illumina NGS data analysis**. *PloS one* 2013, **8**(12):e85024.

177. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

178. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nature methods* 2015, **12**(1):59-60.

179. Leimena MM, Ramiro-Garcia J, Davids M, Van den Bogert B, Smidt H, Smid EJ, Boekhorst J, Zoetendal EG, Schaap PJ, Kleerebezem M: **A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets**. *BMC genomics* 2013, **14**:530.

180. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: **Metagenomic microbial community profiling using unique clade-specific marker genes**. *Nature methods* 2012, **9**(8):811-814.

181. Sharon I, Banfield JF: **Genomes from Metagenomics**. *Science* 2013, **342**(1057).

182. Davids M, Hugenholtz F, Martins Dos Santos V, Smidt H, Kleerebezem M, Schaap PJ: **Functional Profiling of Unfamiliar Microbial Communities Using a Validated De Novo Assembly Metatranscriptome Pipeline**. *PloS one* 2016, **11**(1):e0146423.

183. Namiki T, Hachiya T, Tanaka H, Sakakibara Y: **MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads**. *Nucleic acids research* 2012, **40**(20):e155.

184. Wu Y-W, Tsang Y-H, Tringe SG, Simmons BA, Singer SW: **maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm**. *Microbiome* 2014, **2**:26.

185. Yang B, Peng Y, Leung HC-M, Yiu S-M, Chen J-C, Chin FY-L: **unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers**. *BMC bioinformatics* 2010, **11**(Suppl 2):S5.

186. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**. *Genome research* 2015, **25**(7):1043-1055.

187. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome research* 2007, **17**(3):377-386.

188. Wood DE, Salzberg SL: **kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome biology* 2014, **15**:46.

189. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments**. *Nature methods* 2007, **4**(1):63-72.

190. Lindgreen S, Adair KL, Gardner PP: **An evaluation of the accuracy and speed of metagenome analysis tools**. *Scientific reports* 2016, **6**:19233.

191. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC bioinformatics* 2010, **11**:119.

192. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution**. *Nucleic acids research* 2000, **28**(1):33-36.

193. Claudel-Renard C, Chevalet C, Faraut T, Khan D: **enzyme-specific profiles for genome annotation - PRIAM**. *Nucleic Acids Reseach* 2003, **31**(22):6633-6639.

194. Karp PD, Paley S, Altman T: **Data mining in the MetaCyc family of pathway databases**. In: *Data mining for systems biology: Methods and protocols.* vol. 939. New York: Springer Science+Business Media; 2013: 183-200.

195. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic acids research* 2014, **42**(Database issue):D490-495.

196. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation**. *Nucleic acids research* 2012, **40**(Web Server issue):W445-451.

197. The Gene Ontology Consortium: **gene ontology: tool for the unification of biology**. *Nature genetics* 2000, **25**:95-98.

198. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S *et al*: **InterPro in 2011: new developments in the family and domain prediction database**. *Nucleic acids research* 2012, **40**(Database issue):D306-312.

199. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al*: **A large-scale evaluation of computational protein function prediction**. *Nature methods* 2013, **10**(3):221-227.

200. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF: **Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla**. *mBio* 2013, **4**(5):e00708-00713.

201. Probst AJ, Weinmaier T, Raymann K, Perras A, Emerson JB, Rattei T, Wanner G, Klingl A, Berg IA, Yoshinaga M *et al*: **Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface**. *Nature communications* 2014, **5**:5497.

202. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, Verberkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH *et al*: **fermentation, hydrogen, and sulfur metabolism in mulitple uncultivated bacterial phyla**. *Science* 2012, **337**:1661-1665.

203. **Go Enrichment Analysis** [http://geneontology.org/page/go-enrichment-analysis]

204. Prosser JI: **Replicate or lie**. *Environmental microbiology* 2010, **12**(7):1806-1810.

205. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I *et al*: **IMG/M: a data management and analysis system for metagenomes**. *Nucleic acids research* 2008, **36**(Database issue):D534-538.

206. Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, Pesseat S, Potter S, Scheremetjew M, Sterk P *et al*: **EBI metagenomics in 2016--an expanding and evolving resource for the analysis and archiving of metagenomic data**. *Nucleic acids research* 2016, **44**(D1):D595-603.

207. Shi Y, Tyson GW, Eppley JM, DeLong EF: **Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean**. *The ISME journal* 2011, **5**(6):999-1013.

208. Nocker A, Richter-Heitmann T, Montijn R, Schuren F, Kort R: **Discrimination between live and dead cellsin bacterial communities from environmental water samples analyzed by 454 pyrosequencing**. *Int Microbiol* 2010, **13**(2):59-65.

209. Kopylova E, Noe L, Touzet H: **SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data**. *Bioinformatics* 2012, **28**(24):3211-3217.

210. Celaj A, Markle J, Danska J, Parkinson J: **Comparison of assembly algorithms for improving rate of metatranacriptomic functional annotation**. *Microbiome* 2014, **2**:39.

211. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

212. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome biology* 2014, **15**:12.

213. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA *et al*: **Diet rapidly and reproducibly alters the human gut microbiome**. *Nature* 2014, **505**(7484):559-563.

214. Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, Niazi F, Affourtit J, Egholm M, Henrissat B, Knight R *et al*: **Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(16):7503-7508.

215. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL *et al*: **Shotgun metaproteomics of the human distal gut microbiota**. *The ISME journal* 2009, **3**(2):179-189.

216. Fisher CK, Mehta P: **Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression**. *PloS one* 2014, **9**(7):e102451.

217. Cui H, Zhang Y: **Alignment-free supervised classification of metagenomes by recursive SVM**. *BMC genomics* 2013, **14**:641.

218. Knights D, Costello EK, Knight R: **Supervised classification of human microbiota**. *FEMS microbiology reviews* 2011, **35**(2):343-359.

219. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, Pei Z, Blaser MJ, Aliferis CF, Alekseyenko AV: **a comprehensive evaluation of multicategory classification methods for microbiomic data**. *Microbiome* 2013, **1**:11.

220. Parloff R: **Why deep learning is suddenly changing your life**. In: *Fortune.* Time Inc.; 2016.

221. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM *et al*: **Enterotypes of the human gut microbiome**. *Nature* 2011, **473**(7346):174-180.

222. Jain AK: **data clustering: 50 years beyond k-means**. *Pattern Recognition Letters* 2010, **31**.

223. Desgraupes B: **clusterCrit: Clustering Indices**. In., R package version 1.2.7 edn; 2016.

224. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, Knight R: **Rethinking "enterotypes"**. *Cell host & microbe* 2014, **16**(4):433-437.

225. Schubert E, Koos A, Emrich T, Züfle A, Schmid KA, Zimek A: **A Framework for Clustering Uncertain Data**. *Proceedings of the VLDB Endowment* 2015, **18**(12):1976-1979.

226. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **the WEKA data mining software - an update**. *SIGKDD Explorations* 2003, **11**(1):10-18.
227. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic acids research* 2012, **40**(Database issue):D109-114.
228. May A, Brand BW, El-Kebir M, Klau GW, Zaura E, Crielaard W, Heringa J, Abeln S: **metaModules identifies key functional subnetworks in microbiome-related disease**. *Bioinformatics* 2015, **32**:11.
229. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D *et al*: **Population-level analysis of gut microbiome variation**. *Science* 2016, **352**(6285):560-564.
230. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT *et al*: **Human genetics shape the gut microbiome**. *Cell* 2014, **159**(4):789-799.
231. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M *et al*: **Personalized Nutrition by Prediction of Glycemic Responses**. *Cell* 2015, **163**(5):1079-1094.
232. Ondov BD, Bergman NH, Philippy AM: **interactive metagenomic visualization in the web browser**. *BMC bioinformatics* 2011, **12**:385.
233. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D *et al*: **Visualization of omics data for systems biology**. *Nature methods* 2010, **7**(3 Suppl):S56-68.
234. Heer J, Bostock M, Ogievetsky V: **A tour through the visualization zoo**. *Communications of the ACM* 2010, **53**(6):59.
235. Kreft JU: **Conflicts of interest in biofilms**. *Biofilms* 2004, **1**(4):265-276.
236. Shou W, Ram S, Vilar JM: **Synthetic cooperation in engineered yeast populations**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(6):1877-1882.
237. Hansen AK, Moran NA: **Aphid genome expression reveals host-symbiont cooperation in the production of amino acids**. *Proceedings of the National Academy of Sciences* 2011, **108**(7):2849-2854.
238. Van Leuven JT, Meister RC, Simon C, McCutcheon JP: **Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one**. *Cell* 2014, **158**(6):1270-1280.
239. Morris JJ, Lenksi RE, Zinser ER: **The Black Queen Hypothesis: Evolution of Dependencies Through Adaptive Gene Loss**. *mBio* 2012, **3**(2):e00036-00012.
240. Shetty SA, 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, **41**(2):182-199.
241. Borenstein E: **Computational systems biology and in silico modeling of the human microbiome**. *Briefings in bioinformatics* 2012, **13**(6):769-780.
242. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL: **High-throughput generation, optimization and analysis of genome-scale metabolic models**. *Nature biotechnology* 2010, **28**(9):977-982.
243. Karp PD, Paley S, Romero P: **The Pathway tools software**. *Bioinformatics* 2002, **18**(Suppl. 1 2002):S225–S232.
244. **KBase - predictive biology** [http://kbase.us]
245. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nature protocols* 2010, **5**(1):93-121.

246. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO: **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(6):1777-1782.

247. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO: **A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011**. *Molecular systems biology* 2011, **7**:535.

248. van Heck RG, Ganter M, Martins Dos Santos VA, Stelling J: **Efficient Reconstruction of Predictive Consensus Metabolic Network Models**. *PLoS computational biology* 2016, **12**(8):e1005085.

249. Heinken A, Sahoo S, Fleming RM, Thiele I: **Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut**. *Gut microbes* 2013, **4**(1):28-40.

250. El-Semman IE, Karlsson FH, Shoaie S, Nookaew I, Soliman TH, Nielsen J: **genome-scale metabolic reconstructions of *Bifidobacterium adolescentis* L2-32 and *Facebaclibacterium prausnitzii* A2-165 and their interaction**. *BMC systems biology* 2014, **8**:41.

251. Levy R, Borenstein E: **Reverse Ecology: From Systems to Environments and Back**. *Adv Exp Med Biol* 2012, **751**:329-345.

252. Feist AM, Palsson BO: **The biomass objective function**. *Current opinion in microbiology* 2010, **13**(3):344-349.

253. Khandelwal RA, Olivier BG, Roling WF, Teusink B, Bruggeman FJ: **Community flux balance analysis for microbial consortia at balanced growth**. *PloS one* 2013, **8**(5):e64567.

254. Zomorrodi AR, Islam MM, Maranas CD: **d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities**. *ACS Synth Biol* 2014, **3**(4):247-257.

255. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, Bonilla G, Kar A, Leiby N, Mehta P *et al*: **Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics**. *Cell reports* 2014, **7**(4):1104-1115.

256. Munoz-Tamayo R, Laroche B, Walter E, Dore J, Duncan SH, Flint HJ, Leclerc M: **Kinetic modelling of lactate utilization and butyrate production by key human colonic bacterial species**. *FEMS microbiology ecology* 2011, **76**(3):615-624.

257. Munoz-Tamayo R, Laroche B, Walter E, Dore J, Leclerc M: **Mathematical modelling of carbohydrate degradation by human colonic microbiota**. *Journal of theoretical biology* 2010, **266**(1):189-201.

258. Xavier JB, Foster KR: **Cooperation and conflict in microbial biofilms**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(3):876-881.

259. Bucci V, Bradde S, Biroli G, Xavier JB: **Social interaction, noise and antibiotic-mediated switches in the intestinal microbiota**. *PLoS computational biology* 2012, **8**(4):e1002497.

260. Marino S, Baxter NT, Huffnagle GB, Petrosino JF, Schloss PD: **Mathematical modeling of primary succession of murine intestinal microbiota**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(1):439-444.

261. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Jr., Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**(2):389-401.

262. Williams CF, Walton GE, Jiang L, Plummer S, Garaiova I, Gibson GR: **Comparative analysis of intestinal tract models**. *Annual review of food science and technology* 2015, **6**:329-350.

263. Van den Abbeele P, Belzer C, Goossens M, Kleerebezem M, De Vos WM, Thas O, De Weirdt R, Kerckhof FM, Van de Wiele T: **Butyrate-producing *Clostridium* cluster XIVa species specifically colonize mucins in an in vitro gut model**. *The ISME journal* 2013, **7**(5):949-961.

264. Kim HJ, Li H, Collins JJ, Ingber DE: **Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-**

chip. *Proceedings of the National Academy of Sciences of the United States of America* 2016, **113**(1):E7-15.

265. Williams SC: **Gnotobiotics**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(5):1661.

266. Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, Mazmanian SK: **Bacterial colonization factors control specificity and stability of the gut microbiota**. *Nature* 2013, **501**(7467):426-429.

267. Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, Fulton R, Latreille P, Kim K, Wilson RK, Gordon JI: **Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(25):10643-10648.

268. Laycock G, Sait L, Inman C, Lewis M, Smidt H, van Diemen P, Jorgensen F, Stevens M, Bailey M: **A defined intestinal colonization microbiota for gnotobiotic pigs**. *Veterinary immunology and immunopathology* 2012, **149**(3-4):216-224.

269. Sonnenburg JL, Chen CT, Gordon JI: **Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host**. *PLoS biology* 2006, **4**(12):e413.

270. Backhed F, Manchester JK, Semenkovich CF, Gordon JI: **Mechanisms underlying the resistance to diet-induced obesity in germ-free mice**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(3):979-984.

271. Kong LC, Tap J, Aron-Wisnewsky J, Pelloux V, Basdevant A, Bouillot JL, Zucker JD, Dore J, Clement K: **Gut microbiota after gastric bypass in human obesity: increased richness and associations of bacterial genera with adipose tissue genes**. *The American journal of clinical nutrition* 2013, **98**(1):16-24.

272. Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing R, Rittmann BE *et al*: **Human gut microbiota in obesity and after gastric bypass**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(7):2365-2370.

273. Tremaroli V, Karlsson F, Werling M, Stahlman M, Kovatcheva-Datchary P, Olbers T, Fandriks L, le Roux CW, Nielsen J, Backhed F: **Roux-en-Y Gastric Bypass and Vertical Banded Gastroplasty Induce Long-Term Changes on the Human Gut Microbiome Contributing to Fat Mass Regulation**. *Cell metabolism* 2015, **22**(2):228-238.

274. Graessler J, Qin Y, Zhong H, Zhang J, Licinio J, Wong ML, Xu A, Chavakis T, Bornstein AB, Ehrhart-Bornstein M *et al*: **Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters**. *The pharmacogenomics journal* 2013, **13**(6):514-522.

275. Jalanka J, Mattila E, Jouhten H, Hartman J, de Vos WM, Arkkila P, Satokari R: **Long-term effects on luminal and mucosal microbiota and commonly acquired taxa in faecal microbiota transplantation for recurrent *Clostridium difficile* infection**. *BMC medicine* 2016, **14**:155.

276. Borody TJ, Warren EF, Leis S, Surace R, Ashman O: **Treatment of Ulcerative Colitis Using Fecal Bacteriotherapy**. *J Clin Gastroenterol* 2003, **37**(1):42-47.

277. Bojanova DP, Bordenstein SR: **Fecal Transplants: What Is Being Transferred?** *PLoS biology* 2016, **14**(7):e1002503.

278. Ott SJ, Waetzig GH, Rehman A, Moltzau-Anderson J, Bharti R, Grasis JA, Cassidy L, Tholey A, Fickenscher H, Seegert D *et al*: **Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With *Clostridium difficile* Infection**. *Gastroenterology* 2017, **152**(4):799-811.

279. Smith MB, Kelly C, Alm EJ: **How to regulate faecal transplants**. *Nature* 2014, **506**:290-291.

280. de Vos WM: **Fame and future of faecal transplantations--developing next-generation therapies with synthetic microbiomes**. *Microbial biotechnology* 2013, **6**(4):316-325.

281. Roberfroid M: **Prebiotics: The Concept Revisisted**. *The Journal of nutrition* 2007, **137**:830S-837S.
282. Martin FP, Wang Y, Sprenger N, Yap IK, Lundstedt T, Lek P, Rezzi S, Ramadan Z, van Bladeren P, Fay LB *et al*: **Probiotic modulation of symbiotic gut microbial-host metabolic interactions in a humanized microbiome mouse model**. *Molecular systems biology* 2008, **4**:157.
283. Ventura M, O'Connell-Motherway M, Leahy S, Moreno-Munoz JA, Fitzgerald GF, van Sinderen D: **From bacterial genome to functionality; case bifidobacteria**. *International journal of food microbiology* 2007, **120**(1-2):2-12.
284. Veiga P, Pons N, Agrawal A, Oozeer R, Guyonnet D, Brazeilles R, Faurie JM, van Hylckama Vlieg JE, Houghton LA, Whorwell PJ *et al*: **Changes of the human gut microbiome induced by a fermented milk product**. *Scientific reports* 2014, **4**:6328.
285. Nami Y, Abdullah N, Haghshenas B, Radiah D, Rosli R, Khosroushahi AY: **Probiotic assessment of *Enterococcus durans* 6HL and *Lactococcus lactis* 2HL isolated from vaginal microflora**. *Journal of medical microbiology* 2014, **63**(Pt 8):1044-1051.
286. Sakata T, Kojima T, Fujieda M, Takahashi M, Michibata T: **Influences of probiotic bacteria on organic acid production by pig caecal bacteria in vitro**. *The Proceedings of the Nutrition Society* 2003, **62**(1):73-80.
287. Hosseini E, Grootaert C, Verstraete W, Van de Wiele T: **Propionate as a health-promoting microbial metabolite in the human gut**. *Nutrition reviews* 2011, **69**(5):245-258.
288. De Keersmaecker SC, Verhoeven TL, Desair J, Marchal K, Vanderleyden J, Nagy I: **Strong antimicrobial activity of *Lactobacillus rhamnosus* GG against *Salmonella typhimurium* is due to accumulation of lactic acid**. *FEMS microbiology letters* 2006, **259**(1):89-96.
289. Gilad O, Jacobsen S, Stuer-Lauridsen B, Pedersen MB, Garrigues C, Svensson B: **Combined transcriptome and proteome analysis of *Bifidobacterium animalis* subsp. *lactis* BB-12 grown on xylo-oligosaccharides and a model of their utilization**. *Applied and environmental microbiology* 2010, **76**(21):7285-7291.
290. Everard A, Belzer C, Geurts L, Ouwerkerk J, Druart C, Bindels LB, Guiot Y, Derien M, Muccioli GG, Delzenne NM *et al*: **Cross-talk betweeen *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(22):9066–9071.
291. Etxeberria U, Arias N, Boque N, Macarulla MT, Portillo MP, Martinez JA, Milagro FI: **Reshaping faecal gut microbiota composition by the intake of trans-resveratrol and quercetin in high-fat sucrose diet-fed rats**. *The Journal of nutritional biochemistry* 2015, **26**(6):651-660.
292. Selma MV, Espin JC, Tomas-Barberan FA: **Interaction between phenolics and gut microbiota: role in human health**. *Journal of agricultural and food chemistry* 2009, **57**(15):6485-6501.
293. Carrington D: **Reading the book of life**. In: *BBC News.* UK; 2000.
294. Weiss R, Gillis J: **Teams Finish Mapping Human DNA**. In: *The Washington Post.* WP Company LLC; 2000.
295. Drmanac R: **The advent of personal genome sequencing**. *Genetics in medicine : official journal of the American College of Medical Genetics* 2011, **13**(3):188-190.
296. Maher B: **The case of the missing heritability**. *Nature* 2008, **456**(6):18-21.
297. Offit K: **Personalized medicine: new genomics, old lessons**. *Human genetics* 2011, **130**(1):3-14.
298. Snyder M, Du J, Gerstein M: **Personal genome sequencing: current approaches and challenges**. *Genes & development* 2010, **24**(5):423-431.
299. Flores M, Glusman G, Brogaard K, Price ND, Hood L: **P4 medicine: how systems medicine will transform the healthcare sector and society**. *Future Medicine* 2013, **10**(6):565–576.

300. Hood L, Balling R, Auffray C: **Revolutionizing medicine in the 21st century through systems approaches**. *Biotechnology journal* 2012, **7**(8):992-1001.

301. Robbins MJ: **I Got My Personal Genome Mapped and It Was Bullshit**. In: *vicecom.* VICE Media LLC 2013.

302. Hanage WP: **Microbiology: Microbiome science needs a healthy dose of scepticism**. *Nature* 2014, **512**(7514):247-248.

303. Carroll AE: **Exciting Microbe Research? Temper That Giddy Feeling in Your Gut**. In: *The New York Times.* The New York Times Company 2017.

304. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI: **Host-bacterial mutualism in the human intestine**. *Science* 2005, **307**(5717):1915-1920.

305. Scott KP, Gratz SW, Sheridan PO, Flint HJ, Duncan SH: **The influence of diet on the gut microbiota**. *Pharmacological research* 2013, **69**(1):52-60.

306. Ouwerkerk JP, de Vos WM, Belzer C: **Glycobiome: Bacteria and mucus at the epithelial interface**. *Best Practice & Research Clinical Gastroenterology* 2013, **27**(1):25-38.

307. Elia M, Cummings JH: **Physiological aspects of energy metabolism and gastrointestinal effects of carbohydrates**. *Eur J Clin Nutr* 2007, **61 Suppl 1**:S40-74.

308. Lange K, Hugenholtz F, Schols H, Kleerebezem M, Smidt H, Müller M, Hooiveld GJEJ: **Comparison of the effects of five dietary fibers on mucosal transcriptional profiles and luminial micribiota composition and SCFA concentrations in murine colon**. *Molecular Nutrition & Food Research* 2015, **59**(8):1590-1602.

309. Gerritsen J, Timmerman HM, Fuentes S, van Minnen LP, Panneman H, Konstantinov SR, Rombouts FM, Gooszen HG, Akkermans LM, Smidt H *et al*: **Correlation between protection against sepsis by probiotic therapy and stimulation of a novel bacterial phylotype**. *Applied and environmental microbiology* 2011, **77**(21):7749-7756.

310. Quigley EMM: **Gut Bacteria in Health and Disease**. *Gastroenterology & Hepatology* 2013, **9**(9):560-569.

311. Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, Zhang Y, Shen J, Pang X, Zhang M *et al*: **Symbiotic gut microbes modulate human metabolic phenotypes**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(6):2117-2122.

312. Booijink CCGM, Zoetendal EG, Kleerebezem M, de Vos WM: **Microbial communities in the human small intestine: coupling diversity to metagenomics**. *Future Microbiology* 2007, **2**(3):285-295.

313. den Bogert Bv, Erkus O, Boekhorst J, Goffau Md, Smid EJ, Zoetendal EG, Kleerebezem M: **Diversity of human small intestinal *Streptococcus* and *Veillonella* populations**. *FEMS microbiology ecology* 2013, **85**(2):376-388.

314. Zhang Z, Geng J, Tang X, Fan H, Xu J, Wen X, Ma Z, Shi P: **Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota**. *The ISME journal* 2013, **8**(4):881-893.

315. Van den Bogert B, Boekhorst J, Herrmann R, Smid EJ, Zoetendal EG, Kleerebezem M: **Comparative Genomics Analysis of *Streptococcus* Isolates from the Human Small Intestine Reveals their Adaptation to a Highly Dynamic Ecosystem**. *PloS one* 2014, **8**(12):e83418.

316. Gerritsen J, Fuentes S, Grievink W, van Niftrik L, Tindall BJ, Timmerman HM, Rijkers GT, Smidt H: **Characterization of *Romboutsia ilealis* gen. nov., sp. nov., isolated from the gastro-intestinal tract of a rat, and proposal for the reclassification of five closely related members of the genus *Clostridium* into the genera *Romboutsia* gen. nov., *Intestinibacter* gen. nov., *Terrisporobacter* gen. nov. and *Asaccharospora* gen. nov**. *International journal of systematic and evolutionary microbiology* 2014, **64**(Pt 5):1600-1616.

317. Galperin MY, Brover V, Tolstoy I, Yutin N: **Phylogenomic analysis of the family *Peptostreptococcaceae* (*Clostridium* cluster XI) and proposal for reclassification of *Clostridium litorale* (Fendrich et al. 1991) and *Eubacterium acidaminophilum* (Zindel et al.

1989) as *Peptoclostridium litorale* gen. nov. comb. nov. and *Peptoclostridium acidaminophilum* comb. nov. *International journal of systematic and evolutionary microbiology* 2016, **66**(12):5506-5513.

318. Lowe TM, Eddy SR: **trnascan-SE - a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic acids research* 1997, **25**(5):955–964.

319. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic acids research* 2007, **35**(9):3100-3108.

320. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters**. *Bioinformatics* 2007, **23**(10):1282-1288.

321. UniProt C: **Activities at the Universal Protein Resource (UniProt)**. *Nucleic acids research* 2014, **42**(Database issue):D191-198.

322. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A: **Rfam 11.0: 10 years of RNA families**. *Nucleic acids research* 2013, **41**(Database issue):D226-232.

323. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P: **CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats**. *BMC bioinformatics* 2007, **8**:209.

324. Latendresse M, Krummenacker M, Trupp M, Karp PD: **Construction and completion of flux balance models from pathway databases**. *Bioinformatics* 2012, **28**(3):388-396.

325. Stams AJM, Van Dijk JB, Dijkema C, Plugge CM: **Growth of Syntrophic Propionate-Oxidizing Bacteria with Fumarate in the Absence of Methanogenic Bacteria**. *Applied and environmental microbiology* 1993, **59**(4):1114-1119.

326. Karasawa T, Ikoma S, Yamakawa K, Nakamura S: **A defined growth medium for *Clostridium difficile***. *Microbiology* 1995, **141**:371-375.

327. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357-359.

328. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nature biotechnology* 2013, **31**(1):46-53.

329. ter Braak C, Šmilauer P: **Canoco reference manual and user’s guide: software of ordination (version 5.0).** Ithaca USA : Microcomputer Power 2012.

330. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M *et al*: **The treatment-naïve microbiome in new-onset Crohn’s disease**. *Cell host & microbe* 2014, **15**(3):382-392.

331. Alipour M, Zaidi D, Valcheva R, Jovel J, Martínez I, Sergi C, Walter J, Mason AL, Wong GK-S, Dieleman LA *et al*: **Mucosal Barrier Depletion and Loss of Bacterial Diversity are Primary Abnormalities in Paediatric Ulcerative Colitis**. *Journal of Crohn's & Colitis* 2016, **10**(4):462-471.

332. Lee ZM, Bussema C, 3rd, Schmidt TM: **rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea**. *Nucleic acids research* 2009, **37**(Database issue):D489-493.

333. Klappenbach JA, Dunbar JM, Schmidt TM: **rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria**. *Applied and environmental microbiology* 2000, **66**(4):1328-1333.

334. Yano K, Wada T, Suzuki S, Tagami K, Matsumoto T, Shiwa Y, Ishige T, Kawaguchi Y, Masuda K, Akanuma G *et al*: **Multiple rRNA operons are essential for efficient cell growth and sporulation as well as outgrowth in *Bacillus subtilis***. *Microbiology* 2013, **159**(11):2225-2236.

335. Bensaadi-Merchermek N, Salvado J-C, Cagnon C, Karama S, Mouchès C: **Characterization of the unlinked 16S rDNA and 23S-5S rRNA operon of *Wolbachia pipientis*, a prokaryotic parasite of insect gonads**. *Gene* 1995, **165**(1):81-86.

336. Schwartz JJ, Gazumyan A, Schwartz I: **rRNA gene organization in the Lyme disease spirochete, *Borrelia burgdorferi***. *Journal of bacteriology* 1992, **174**(11):3757-3765.

337. Tatusov RL: **A Genomic Perspective on Protein Families**. *Science* 1997, **278**(5338):631-637.

338. Reichardt N, Duncan SH, Young P, Belenguer A, McWilliam Leitch C, Scott KP, Flint HJ, Louis P: **Phylogenetic distribution of three pathways for propionate production within the human gut microbiota**. *The ISME journal* 2014, **8**(6):1323-1335.

339. Dhillon A, Goswami S, Riley M, Teske A, Sogin M: **Domain Evolution and Functional Diversification of Sulfite Reductases**. *Astrobiology* 2005, **5**(1):18-29.

340. Czyzewski BK, Wang D-N: **Identification and characterization of a bacterial hydrosulfide ion channel**. *Nature* 2012, **483**(7390):494-497.

341. Mitsuhashi H, Nojima Y, Tanaka T, Ueki K, Maezawa A, Yano S, Naruse T: **Sulfite is released by human neutrophils in response to stimulation with lipopolysaccharide**. *Journal of Leukocyte Biology* 1998, **64**(5):595-599.

342. LeBlanc DJ, Mortlock RP: **Metabolism of d-Arabinose: a New Pathway in *Escherichia coli***. *Journal of bacteriology* 1971, **106**(1):90-96.

343. Almagro-Moreno S, Boyd EF: **Insights into the evolution of sialic acid catabolism among bacteria**. *BMC Evolutionary Biology* 2009, **9**:118.

344. Ng KM, Ferreyra JA, Higginbottom SK, Lynch JB, Kashyap PC, Gopinath S, Naidu N, Choudhury B, Weimer BC, Monack DM *et al*: **Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens**. *Nature* 2013, **502**(7469):96-99.

345. Derrien M, van Passel MWJ, van de Bovenkamp JHB, Schipper RG, de Vos WM, Dekker J: **Mucin-bacterial interactions in the human oral cavity and digestive tract**. *Gut microbes* 2010, **1**(4):254-268.

346. Roggentin P, Gutschker-Gdaniec G, Schauer R, Hobrecht R: **Correlative Properties for a Differentiation of Two *Clostridium sordellii* Phenotypes and their Distinction from *Clostridium bifermentans***. *Zentralblatt für Bakteriologie, Mikrobiologie und Hygiene Series A: Medical Microbiology, Infectious Diseases, Virology, Parasitology* 1985, **260**(3):319-328.

347. Mobley HL, Island MD, Hausinger RP: **Molecular biology of microbial ureases**. *Microbiological Reviews* 1995, **59**(3):451-480.

348. Fuller MF, Reeds PJ: **Nitrogen cycling in the gut**. *Annual review of nutrition* 1998, **18**:385-411.

349. Rutherford JC: **The Emerging Role of Urease as a General Microbial Virulence Factor**. *PLoS pathogens* 2014, **10**(5):e1004062.

350. Ridlon JM, Kang DJ, Hylemon PB: **Bile salt biotransformations by human intestinal bacteria**. *Journal of lipid research* 2006, **47**(2):241-259.

351. Commichau FM, Rothe FM, Herzberg C, Wagner E, Hellwig D, Lehnik-Habrink M, Hammer E, Völker U, Stülke J: **Novel Activities of Glycolytic Enzymes in *Bacillus subtilis*: Interactions with Essential Proteins Involved in mRNA Processing**. *Molecular & Cellular Proteomics : MCP* 2009, **8**(6):1350-1360.

352. Calusinska M, Happe T, Joris B, Wilmotte A: **The surprising diversity of clostridial hydrogenases: a comparative genomic perspective**. *Microbiology* 2010, **156**(Pt 6):1575-1588.

353. Kortman GAM, Raffatellu M, Swinkels DW, Tjalsma H: **Nutritional iron turned inside out: intestinal stress from a gut microbial perspective**. *FEMS microbiology reviews* 2014, **38**(6):1202-1234.

354. van Hijum SAFT, Kralj S, Ozimek LK, Dijkhuizen L, van Geel-Schutten IGH: **Structure-Function Relationships of Glucansucrase and Fructansucrase Enzymes from Lactic Acid Bacteria**. *Microbiology and Molecular Biology Reviews* 2006, **70**(1):157-176.

355. Sharma V, Prere M, Canal I, Firth AE, Atkins JF, Baranov PV, Fayet O: **Analysis of tetra- and hepta-nucleotides motifs promoting -1 ribosomal frameshifting in *Escherichia coli***. *Nucleic acids research* 2014, **42**(11):7210-7225.

356. Baldomà L, Aguilar J: **Metabolism of L-fucose and L-rhamnose in *Escherichia coli*: aerobic-anaerobic regulation of L-lactaldehyde dissimilation**. *Journal of bacteriology* 1988, **170**(1):416-421.

357. Hooper LV, Xu J, Falk PG, Midtvedt T, Gordon JI: **A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(17):9833-9838.

358. Scott KP, Martin JC, Campbell G, Mayer CD, Flint HJ: **Whole-genome transcription profiling reveals genes up-regulated by growth on fucose in the human gut bacterium "*Roseburia inulinivorans*"**. *Journal of bacteriology* 2006, **188**(12):4340-4349.

359. Becker DJ, Lowe JB: **Fucose: biosynthesis and biological function in mammals**. *Glycobiology* 2003, **13**(7):41R-53R.

360. Robbe C, Capon C, Coddeville B, Michalski J-C: **Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract**. *Biochemical Journal* 2004, **384**(Pt 2):307-316.

361. Robbe C, Capon C, Maes E, Rousset M, Zweibaum A, Zanetta JP, Michalski J-C: **Evidence of regio-specific glycosylation in human intestinal mucins: presence of an acidic gradient along the intestinal tract**. *The Journal of biological chemistry* 2003, **278 47**:46337-46348.

362. Stahl M, Friis LM, Nothaft H, Liu X, Li J, Szymanski CM, Stintzi A: **l-Fucose utilization provides *Campylobacter jejuni* with a competitive advantage**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(17):7194-7199.

363. Vimr ER: **Unified Theory of Bacterial Sialometabolism: How and Why Bacteria Metabolize Host Sialic Acids**. *ISRN Microbiology* 2013, **2013**:816713.

364. Vimr ER, Kalivoda KA, Deszo EL, Steenbergen SM: **Diversity of Microbial Sialic Acid Metabolism**. *Microbiology and Molecular Biology Reviews* 2004, **68**(1):132-153.

365. Pickard JM, Maurice CF, Kinnebrew MA, Abt MC, Schenten D, Golovkina T, Bogatyrev SR, Ismagilov RF, Pamer EG, Turnbaugh PJ *et al*: **Rapid fucosylation of intestinal epithelium sustains host-commensal symbiosis in sickness**. *Nature* 2014, **514**(7524):638-641.

366. Deutscher J: **The mechanisms of carbon catabolite repression in bacteria**. *Current opinion in microbiology* 2008, **11**(2):87-93.

367. Marshall BJ, Barrett LJ, Prakash C, McCallum RW, Guerrant RL: **Urea Protects *Helicobacter (Campylobacter) pylori* From the Bactericidal Effect of Acid**. *Gastroenterology* 1990, **99**:697-702.

368. Bergner H, Simon O, Zebrowska T, Münchmeyer R: **Studies on the secretion of amino acids and of urea into the gastrointestinal tract of pigs**. *Archiv für Tierernaehrung* 1986, **36**(6):479-490.

369. Dupuy B, Daube G, Popoff MR, Cole ST: ***Clostridium perfringens* urease genes are plasmid borne**. *Infection and immunity* 1997, **65**(6):2313-2320.

370. Crost EH, Tailford LE, Le Gall G, Fons M, Henrissat B, Juge N: **Utilisation of mucin glycans by the human gut symbiont *Ruminococcus gnavus* is strain-dependent**. *PloS one* 2013, **8**(10):e76341.

371. Dethlefsen L, McFall-Ngai M, Relman DA: **An ecological and evolutionary perspective on human-microbe mutualism and disease**. *Nature* 2007, **449**(7164):811-818.

372. Flint HJ, Duncan SH, Scott KP, Louis P: **Interactions and competition within the microbial community of the human colon: links between diet and health**. *Environmental microbiology* 2007, **9**(5):1101-1111.

373. Cani PD, Everard A: **Talking microbes: When gut bacteria interact with diet and host organs**. *Molecular Nutrition and Food Research* 2016, **60**(1):58-66.

374. Garrett WS, Gordon JI, Glimcher LH: **Homeostasis and inflammation in the intestine**. *Cell* 2010, **140**(6):859-870.

375. Becattini S, Taur Y, Pamer EG: **Antibiotic-Induced Changes in the Intestinal Microbiota and Disease**. *Trends Mol Med* 2016, **22**(6):458-478.

376. Bowman KA, Broussard EK, Surawicz CM: **Fecal microbiota transplantation: current clinical efficacy and future prospects**. *Clin Exp Gastroenterol* 2015, **8**:285-291.

377. Scott KP, Martin JC, Duncan SH, Flint HJ: **Prebiotic stimulation of human colonic butyrate-producing bacteria and bifidobacteria, in vitro**. *FEMS microbiology ecology* 2014, **87**(1):30-40.

378. Leemhuis H, Dobruchowska JM, Ebbelaar M, Faber F, Buwalda PL, van der Maarel MJ, Kamerling JP, Dijkhuizen L: **Isomalto/malto-polysaccharide, a novel soluble dietary fiber made via enzymatic conversion of starch**. *Journal of agricultural and food chemistry* 2014, **62**(49):12034-12044.

379. Aguirre M, Ramiro-Garcia J, Koenen ME, Venema K: **To pool or not to pool? Impact of the use of individual and pooled fecal samples for in vitro fermentation studies**. *Journal of microbiological methods* 2014, **107**:1-7.

380. Rösch C, Venema K, Gruppen H, Schols HA: **Characterisation and in vitro fermentation of resistant maltodextrins using human faecal inoculum and analysis of bacterial enzymes present**. *Bioactive Carbohydrates and Dietary Fibre* 2015, **6**:46-53.

381. Gu F, Borewicz K, Richter B, van der Zaal P, Smidt H, Buwalda PL, Schols H: **In vitro fermentation behaviour of isomalto/malto-polysaccharides using human faecal inoculum indicates prebiotic potential**. *Molecular Nutrition and Food Research* 2018, **62**(12):1800232.

382. Kang S, Denman SE, Morrison M, Yu Z, McSweeney CS: **An efficient RNA extraction method for estimating gut microbial diversity by polymerase chain reaction**. *Current microbiology* 2009, **58**(5):464-471.

383. Martin M: **Cutadapt removes adapter sequences from high-througput sequencing reads**. *EMBnetjournal* 2011, **17**:10-12.

384. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics* 2011, **27**(6):863-864.

385. Peng Y, Leung HC, Yiu SM, Chin FY: **IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth**. *Bioinformatics* 2012, **28**(11):1420-1428.

386. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM**. *Nucleic Acids Res* 2003, **31**(22):6633-6639.

387. Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, Sensen CW, Schomburg D: **BRENDA in 2015: exciting developments in its 25th year of existence**. *Nucleic acids research* 2015, **43**(Database issue):D439-446.

388. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

389. Liao Y, Smyth GK, Shi W: **The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote**. *Nucleic acids research* 2013, **41**(10).

390. Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D: **Using Tablet for visual exploration of second-generation sequencing data**. *Briefings in bioinformatics* 2013, **14**(2):193-202.

391. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

392. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4**. *Genome research* 2011, **21**(9):1552-1560.

393. R Core Team: **R: A language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing; 2012.

394. Sun J, Nishiyama T, Shimizu K, Kadota K: **TCC: an R package for comparing tag count data with robust normalization strategies**. *BMC bioinformatics* 2013, **14**:219.

395. van der Walt S, Colbert C, Varoquaux G: **The NumPy Array: A structure for Efficient Numerical Computation**. *Computing In Science & Engineering* 2011, **13**:22-30.

396. Schonlau M: **Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams**. *Computational Statistics* 2004, **19**:95-111.

397. Ester M, Kriegel H-P, Sander J, Xu X: **A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise**. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining: 1996*. 226-231.

398. Bairoch A: **The ENZYME database in 2000**. *Nucleic acids research* 2000, **28**(1):304-305.

399. 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**(10):5810-5817.

400. Abou Hachem M, S. Møller M, M. Andersen J, Fredslund F, Majumder A, Nakai H, Lo Leggio L, Goh Y-J, Barrangou R, R. Klaenhammer T *et al*: **A Snapshot into the Metabolism of Isomalto-oligosaccharides in Probiotic Bacteria**. *Journal of Applied Glycoscience* 2013, **60**(2):95-100.

401. Kuchtová A, Janeček Š: **Domain evolution in enzymes of the neopullulanase subfamily**. *Microbiology* 2016, **162**(12):2099-2115.

402. Ganzle MG, Follador R: **Metabolism of oligosaccharides and starch in lactobacilli: a review**. *Frontiers in microbiology* 2012, **3**:340.

403. Bailey RW, Clarke RTJ: **A bacterial dextranase**. *Biochemical Journal* 1959, **72**(1):49-54.

404. Khalikova E, Susi P, Korpela T: **Microbial Dextran-Hydrolyzing Enzymes: Fundamentals and Applications**. *Microbiology and Molecular Biology Reviews* 2005, **69**(2):306-325.

405. Gibson GR, Probert HM, Loo JV, Rastall RA, Roberfroid MB: **Dietary modulation of the human colonic microbiota: updating the concept of prebiotics**. *Nutrition research reviews* 2004, **17**(2):259-275.

406. Kohmoto T, Fukui F, Takaku H, Machida Y, Arai M, Mitsuoka T: **Effect of Isomalto-oligosaccharides on Human Fecal Flora**. *Bifidobacteria and Microflora* 1988, **7**(2):61-69.

407. Kaneko T, Yokoyama A, Suzuki M: **Digestibility Characteristics of Isomaltooligosaccharides in Comparison with Several Saccharides Using the Rat Jejunum Loop Method**. *Bioscience, Biotechnology, and Biochemistry* 1995, **59**(7):1190-1194.

408. Ketabi A, Dieleman LA, Ganzle MG: **Influence of isomalto-oligosaccharides on intestinal microbiota in rats**. *Journal of applied microbiology* 2011, **110**(5):1297-1306.

409. Cockburn DW, Koropatkin NM: **Polysaccharide Degradation by the Intestinal Microbiota and Its Influence on Human Health and Disease**. *Journal of Molecular Biology* 2016, **428**(16):3230-3252.

410. De Vos WM: **Metabolic engineering of sugar catabolism in lactic acid bacteria**. *Antonie van Leeuwenhoek* 1996, **70**:223-242.

411. Ze X, Duncan SH, Louis P, Flint HJ: ***Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon**. *The ISME journal* 2012, **6**(8):1535-1543.

412. Rogosa M: ***Acidaminococcus* gen. n., *Acidaminococcus fermentans* sp. n., Anaerobic Gram-negative Diplococci Using Amino Acids as the Sole Energy Source for Growth**. *Journal of bacteriology* 1969, **92**(2):756-766.

413. Baron EJ: **Bilophila wadsworthia: a Unique Gram-negative Anaerobic Rod**. *Anaerobe* 1997, **3**(2):83-86.

414. Moller MS, Fredslund F, Majumder A, Nakai H, Poulsen JC, Lo Leggio L, Svensson B, Abou Hachem M: **Enzymology and structure of the GH13_31 glucan 1,6-alpha-glucosidase that**

confers isomaltooligosaccharide utilization in the probiotic Lactobacillus acidophilus NCFM. *J Bacteriol* 2012, **194**(16):4249-4259.

415. Hu Y, Ketabi A, Buchko A, Ganzle MG: **Metabolism of isomalto-oligosaccharides by *Lactobacillus reuteri* and bifidobacteria**. *Letters in applied microbiology* 2013, **57**(2):108-114.

416. Liu S, Ren F, Zhao L, Jiang L, Hao Y, Jin J, Zhang M, Guo H, Lei X, Sun E *et al*: **Starch and starch hydrolysates are favorable carbon sources for bifidobacteria in the human gut**. *BMC microbiology* 2015, **15**:54.

417. Koropatkin NM, Cameron EA, Martens EC: **How glycan metabolism shapes the human gut microbiota**. *Nature reviews Microbiology* 2012, **10**(5):323-335.

418. Fischbach MA, Sonnenburg JL: **Eating for two: how metabolism establishes interspecies interactions in the gut**. *Cell host & microbe* 2011, **10**(4):336-347.

419. Macfarlane S, Macfarlane GT: **Regulation of short-chain fatty acid production**. *The Proceedings of the Nutrition Society* 2003, **62**(1):67-72.

420. Schleussner C-F, Lissner TK, Fischer EM, Wohland J, Perrette M, Golly A, Rogelj J, Childers K, Schewe J, Frieler K *et al*: **Differential climate impacts for policy-relevant limits to global warming: the case of 1.5 °C and 2 °C**. *Earth System Dynamics* 2016, **7**(2):327-351.

421. Murray RM, Bryant AM, Leng RA: **Rates of production of methane in the rumen and large intestine of sheep**. *British Journal of Nutrition* 2007, **36**(1):1-14.

422. Li RW, Connor EE, Li C, Baldwin Vi RL, Sparks ME: **Characterization of the rumen microbiota of pre-ruminant calves using metagenomic tools**. *Environ Microbiol* 2012, **14**(1):129-139.

423. Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark D, Chen F, Zhang T *et al*: **Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen**. *Science* 2011, **331**:463-467.

424. van Gastelen S, Antunes-Fernandes EC, Hettinga KA, Kop G, Alferink SJ, Hendriks WH, Dijkstra J: **Enteric methane production, rumen volatile fatty acid concentrations, and milk fatty acid composition in lactating holstein-friesian cows fed grass silage- or corn silage-based diets**. *Journal of dairy science* 2015, **98**(3):1915-1927.

425. Liu H, Vaddella V, Zhou D: **Effects of chestnut tannins and coconut oil on growth performance, methane emission, ruminal fermentation, and microbial populations in sheep**. *Journal of dairy science* 2011, **94**(12):6069-6077.

426. Pitta DW, Parmar N, Patel AK, Indugu N, Kumar S, Prajapathi KB, Patel AB, Reddy B, Joshi C: **Bacterial Diversity Dynamics Associated with Different Diets and Different Primer Pairs in the Rumen of Kankrej Cattle**. *PloS one* 2014, **9**(11):e111710.

427. Fernando SC, Purvis II HT, Najar FZ, Sukharnikov LO, Krehbiel CR, Nagaraja TG, Roe BA, DeSilva U: **Rumen Microbial Population Dynamics during Adaptation to a High-Grain Diet**. *Applied and environmental microbiology* 2010, **76**(22):7482-7490.

428. Shi W, Moon CD, Leahy SC, Kang D, Froula J, Kittelmann S, Fan C, Deutsch S, Gagic D, Seedorf H *et al*: **Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome**. *Genome research* 2014, **24**(9):1517-1525.

429. Poulsen M, Schwab C, Jensen BB, Engberg RM, Spang A, Canibe N, Hojberg O, Milinovich G, Fragner L, Schleper C *et al*: **Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen**. *Nature communications* 2013, **4**:1428.

430. Dai X, Tian Y, Li J, Su X, Wang X, Zhao S, Liu L, Luo Y, Liu D, Zheng H *et al*: **Metatranscriptomic analyses of plant cell wall polysaccharide degradation by microorganisms in cow rumen**. *Applied and environmental microbiology* 2014, **81**(4):1375-1386.

431. Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P *et al*: **Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(6):1948-1953.

432. Dassa B, Borovok I, Ruimy-Israeli V, Lamed R, Flint HJ, Duncan SH, Henrissat B, Coutinho P, Morrison M, Mosoni P *et al*: **Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains**. *PloS one* 2014, **9**(7):e99221.

433. Roehe R, Dewhurst RJ, Duthie CA, Rooke JA, McKain N, Ross DW, Hyslop JJ, Waterhouse A, Freeman TC, Watson M *et al*: **Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance**. *PLoS genetics* 2016, **12**(2):e1005846.

434. Jose VL, Appoothy T, More RP, Arun AS: **Metagenomic insights into the rumen microbial fibrolytic enzymes in Indian crossbred cattle fed finger millet straw**. *AMB Express* 2017, **7**:13.

435. van Zijderveld SM, Fonken B, Dijkstra J, Gerrits WJ, Perdok HB, Fokkink W, Newbold JR: **Effects of a combination of feed additives on methane production, diet digestibility, and animal performance in lactating dairy cows**. *Journal of dairy science* 2011, **94**(3):1445-1454.

436. Zoetendal EG, Booijink CCGM, Klaassens ES, Heilig HGHJ, Kleerebezem M, Smidt H, de Vos WM: **Isolation of RNA from bacterial samples of the human gastrointestinal tract**. *Nature protocols* 2006, **1**(2):954-959.

437. Murphy NR, Hellwig RJ: **Improved nucleic acid organic extraction through use of a unique gel barrier material**. *BioTechniques* 1996, **21**(5):934-936, 938-939.

438. Liao Y, Smyth GK, Shi W: **The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote**. *Nucleic acids research* 2013, **41**(10):e108.

439. Pedregoa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al*: **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research* 2011, **12**:2825-2830.

440. Hunter JD: **Matplotlib: A 2D graphics environment**. *Computing In Science & Engineering* 2007, **9**(3):90-95.

441. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS: **Microbial Cellulose Utilization: Fundamentals and Biotechnology**. *Microbiology and molecular biology reviews : MMBR* 2002, **66**(3):506-577.

442. Bayer EA, Lamed R, White BA, Flint HJ: **From cellulosomes to cellulosomics**. *The Chemical Record* 2008, **8**(6):364-377.

443. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA: **Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis**. *Nature reviews Microbiology* 2008, **6**(2):121-131.

444. van Kessel JAS, Russel JB: **The effect of pH on ruminal methanogenesis**. *FEMS microbiology ecology* 1996, **20**:205-210.

445. Kleen JL, Hooijer GA, Rehage J, Noordhuizen JPTM: **Subacute Ruminal Acidosis (SARA): a Review**. *J Vet Med A Physiol Pathol Clin Med* 2003, **50**(8):406-414.

446. Leahy SC, Kelly WJ, Altermann E, Ronimus RS, Yeoman CJ, Pacheco DM, Li D, Kong Z, McTavish S, Sang C *et al*: **The genome sequence of the rumen methanogen *Methanobrevibacter ruminantium* reveals new possibilities for controlling ruminant methane emissions**. *PloS one* 2010, **5**(1):e8926.

447. Wang H, Zheng H, Browne F, Roehe R, Dewhurst RJ, Engel F, Hemmje M, Lu X, Walsh P: **Integrated metagenomic analysis of the rumen microbiome of cattle reveals key biological mechanisms associated with methane traits**. *Methods* 2017, **124**:108-119.

448. Wallace RJ, Rooke JA, McKain N, Duthie CA, Hyslop JJ, Ross DW, Waterhouse A, Watson M, Roehe R: **The rumen microbial metagenome associated with high methane production in cattle**. *BMC genomics* 2015, **16**:839.

449. Parmar NR, Pandit PD, Purohit HJ, Nirmal Kumar JI, Joshi CG: **Influence of Diet Composition on Cattle Rumen Methanogenesis: A Comparative Metagenomic Analysis in Indian and Exotic Cattle**. *Indian J Microbiol* 2017, **57**(2):226-234.

450. Hudman JF, Gregg K: **Genetic Diversity among Strains of Bacteria from the Rumen**. *Current microbiology* 1989, **19**:313-318.

451. Sasson G, Kruger Ben-Shabat S, Seroussi E, Doron-Faigenboim A, Shterzer N, Yaacoby S, Berg Miller ME, White BA, Halperin E, Mizrahi I: **Heritable Bovine Rumen Bacteria Are Phylogenetically Related and Correlated with the Cow's Capacity To Harvest Energy from Its Feed**. *mBio* 2017, **8**(4):e00703-00717.

452. Finlay BJ, Esteban G, Clarke KJ, Williams AG, Embley TM, Hirt RP: **Some rumen ciliates have endosymbiotic methanogens**. *FEMS microbiology letters* 1994, **117**:157-162.

453. Mullins CR, Mamedova LK, Carpenter AJ, Ying Y, Allen MS, Yoon I, Bradford BJ: **Analysis of rumen microbial populations in lactating dairy cattle fed diets varying in carbohydrate profiles and *Saccharomyces cerevisiae* fermentation product**. *Journal of dairy science* 2013, **96**:5872-5881.

454. Jami E, Israel A, Kotser A, Mizrahi I: **Exploring the bovine rumen bacterial community from birth to adulthood**. *The ISME journal* 2013, **7**(6):1069-1079.

455. Carberry CA, Waters SM, Kenny DA, Creevey CJ: **Rumen Methanogenic Genotypes Differ in Abundance According to Host Residual Feed Intake Phenotype and Diet Type**. *Applied and environmental microbiology* 2014, **80**(2):586-594.

456. Seedorf H, Kittelmann S, Janssen PH: **Few highly abundant operational taxonomic units dominate within rumen methanogenic archaeal species in New Zealand sheep and cattle**. *Applied and environmental microbiology* 2015, **81**(3):986-995.

457. Li Z, Zhang Z, Xu C, Zhao J, Liu H, Fan Z, Yang F, Wright AD, Li G: **Bacteria and methanogens differ along the gastrointestinal tract of Chinese roe deer (*Capreolus pygargus*)**. *PloS one* 2014, **9**(12):e114513.

458. Janssen PH, Kirs M: **Structure of the archaeal community of the rumen**. *Applied and environmental microbiology* 2008, **74**(12):3619-3625.

459. Özcan S, Johnston M: **Function and Regulation of Yeast Hexose Transporters**. *Microbiology and molecular biology reviews : MMBR* 1999, **63**(3):554-569.

460. Sloothaak J, Odoni DI, de Graaff LH, Martins Dos Santos VA, Schaap PJ, Tamayo-Ramos JA: ***Aspergillus niger* membrane-associated proteome analysis for the identification of glucose transporters**. *Biotechnology for biofuels* 2015, **8**:150.

461. Peer A, Smith SP, Bayer EA, Lamed R, Borovok I: **Noncellulosomal cohesin- and dockerin-like modules in the three domains of life**. *FEMS microbiology letters* 2009, **291**:1-16.

462. Ze X, David YB, Laverde-Gomez JA, Dassa B, Sheridan PO, Duncan SH, Louis P, Henrissat B, Juge N, Koropatkin NM *et al*: **Unique Organization of Extracellular Amylases into Amylosomes in the Resistant Starch-Utilizing Human Colonic *Firmicutes* Bacterium *Ruminococcus bromii***. *mBio* 2015, **6**(5):e01058-01015.

463. Mackenzie AK, Pope PB, Pedersen HL, Gupta R, Morrison M, Willats WG, Eijsink VG: **Two SusD-like proteins encoded within a polysaccharide utilization locus of an uncultured ruminant Bacteroidetes phylotype bind strongly to cellulose**. *Applied and environmental microbiology* 2012, **78**(16):5935-5937.

464. Morgavi DP, Martin C, Jouany JP, Ranilla MJ: **Rumen protozoa and methanogenesis: not a simple cause-effect relationship**. *The British journal of nutrition* 2012, **107**(3):388-397.

465. Holmes DE, Giloteaux L, Orellana R, Williams KH, Robbins MJ, Lovley DR: **Methane production from protozoan endosymbionts following stimulation of microbial metabolism within subsurface sediments**. *Frontiers in microbiology* 2014, **5**:366.

466. Snelling TJ, Wallace RJ: **The rumen microbial metaproteome as revealed by SDS-PAGE**. *BMC microbiology* 2017, **17**:9.

467. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(21):9546-9551.

468. Bucci V, Xavier JB: **Towards predictive models of the human gut microbiome**. *Journal of Molecular Biology* 2014, **426**(23):3907-3916.

469. Henry CS, Bernstein HC, Weisenhorn P, Taylor RC, Lee JY, Zucker J, Song HS: **Microbial Community Metabolic Modeling: A Community Data-Driven Network Reconstruction**. *J Cell Physiol* 2016, **231**(11):2339-2345.

470. 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**(2):594.

471. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA: **Metabolic modeling of a mutualistic microbial community**. *Molecular systems biology* 2007, **3**:92.

472. Ponce-de-Leon M, Tamarit D, Calle-Espinosa J, Mori M, Latorre A, Montero F, Pereto J: **Determinism and Contingency Shape Metabolic Complementation in an Endosymbiotic Consortium**. *Frontiers in microbiology* 2017, **8**:2290.

473. van Heck RGA: **Metabolic modeling to understand and redesign microbial systems** Wageningen: Wageningen; 2017.

474. Machado D, Andrejev S, Tramontano M, Patil KR: **Fast automated reconstruction of genome-scale metabolic models for microbial species and communities**. *bioRxiv* 2018.

475. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B: **Accelerating the reconstruction of genome-scale metabolic networks**. *BMC bioinformatics* 2006, **7**:296.

476. McDonald AG, Tipton KF: **Fifty-five years of enzyme classification: advances and difficulties**. *FEBS J* 2014, **281**(2):583-592.

477. D'Ari R, Casadesus J: **Underground metabolism**. *BioEssays* 1998, **20**:181-186.

478. Magnusdottir S, Ravcheev D, de Crecy-Lagard V, Thiele I: **Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes**. *Frontiers in genetics* 2015, **6**:148.

479. Hanson AD, Pribat A, Waller JC, de Crecy-Lagard V: **'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it**. *The Biochemical journal* 2010, **425**(1):1-11.

480. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A: **Toward the automated generation of genome-scale metabolic networks in the SEED**. *BMC bioinformatics* 2007, **8**:139.

481. Durot M, Bourguignon PY, Schachter V: **Genome-scale models of bacterial metabolism: reconstruction and applications**. *FEMS microbiology reviews* 2009, **33**(1):164-190.

482. Smith AA, Belda E, Viari A, Medigue C, Vallenet D: **The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes**. *PLoS computational biology* 2012, **8**(5):e1002540.

483. Ferry JG: **Fundamentals of methanogenic pathways that are key to the biomethanation of complex biomass**. *Current opinion in biotechnology* 2011, **22**(3):351-357.

484. Ferry JG: **Biochemistry of Acetotrophic Methanogenesis**. 2010:357-367.

485. Xavier JC, Patil KR, Rocha I: **Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes**. *Metab Eng* 2017, **39**:200-208.

486. Wodke JA, Puchalka J, Lluch-Senar M, Marcos J, Yus E, Godinho M, Gutierrez-Gallego R, dos Santos VA, Serrano L, Klipp E *et al*: **Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling**. *Molecular systems biology* 2013, **9**:653.

487. Surthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CD: **A Genome-Scale Metabolic Reconstruction of *Mycoplasma genitalium*, i PS189**. *PLoS computational biology* 2009, **5**(2):e1000285.

488. Wang Y, Xu H, Jones MK, White RH: **Identification of the Final Two Genes Functioning in Methanofuran Biosynthesis in *Methanocaldococcus jannaschii***. *Journal of bacteriology* 2015, **197**(17):2850-2858.

489. Blankenfeldt W, Parsons JF: **The structural biology of phenazine biosynthesis**. *Curr Opin Struct Biol* 2014, **29**:26-33.

490. Guttenberger N, Blankenfeldt W, Breinbauer R: **Recent developments in the isolation, biological function, biosynthesis, and synthesis of phenazine natural products**. *Bioorg Med Chem* 2017, **25**(22):6149-6166.

491. Mendoza SN, Canon PM, Contreras A, Ribbeck M, Agosin E: **Genome-Scale Reconstruction of the Metabolic Network in *Oenococcus oeni* to Assess Wine Malolactic Fermentation**. *Frontiers in microbiology* 2017, **8**:534.

492. Tomas-Gamisans M, Ferrer P, Albiol J: **Integration and Validation of the Genome-Scale Metabolic Models of Pichia pastoris: A Comprehensive Update of Protein Glycosylation Pathways, Lipid and Energy Metabolism**. *PloS one* 2016, **11**(1):e0148031.

493. Miller EA, Livermore JA, Alberts SC, Tung J, Archie EA: **Ovarian cycling and reproductive state shape the vaginal microbiota in wild baboons**. *Microbiome* 2017, **5**(1):8.

494. Perez-Cobas AE, Moya A, Gosalbes MJ, Latorre A: **Colonization Resistance of the Gut Microbiota against *Clostridium difficile***. *Antibiotics (Basel)* 2015, **4**(3):337-357.

495. Roestenberg M, Mo A, Kremsner PG, Yazdanbakhsh M: **Controlled human infections: A report from the controlled human infection models workshop, Leiden University Medical Centre 4-6 May 2016**. *Vaccine* 2017, **35**(51):7070-7076.

496. Hardy J, Singleton A: **Genomewide Association Studies and Human Disease**. *New England Journal of Medicine* 2009, **360**(17):1759-1768.

497. Eisenstein M: **Organoids: the body builders**. *Nature methods* 2018, **15**(1):19-22.

498. Kim HJ, Ingber DE: **Gut-on-a-Chip microenvironment induces human intestinal cells to undergo villus differentiation**. *Integrative biology : quantitative biosciences from nano to macro* 2013, **5**(9):1130-1140.

499. Shah P, Fritz JV, Glaab E, Desai MS, Greenhalgh K, Frachet A, Niegowska M, Estes M, Jager C, Seguin-Devaux C *et al*: **A microfluidics-based in vitro model of the gastrointestinal human-microbe interface**. *Nature communications* 2016, **7**:11535.

500. Zoetendal EG, Smidt H: **Endothelial dysfunction: what is the role of the microbiota?** *Gut* 2018, **67**(2):201-202.

501. Baker M: **Is there a reproducibility crisis?** In: *Nature.* vol. 533; 2016: 452-455.

502. Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, Hirst T, Hemblade R, Bahor Z, Nunes-Fonseca C *et al*: **Risk of Bias in Reports of In Vivo Research: A Focus for Improvement**. *PLoS biology* 2015, **13**(10):e1002273.

503. Flier JS: **Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy**. *Mol Metab* 2017, **6**(1):2-9.

504. Baker M: **Over half of psychology studies fail reproducibility test**. In: *Nature.* Nature Inc.; 2015.

505. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE *et al*: **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data* 2016, **3**:160018.

506. Langille MGI, Ravel J, Fricke WF: **"Available upon request": not good enough for microbiome data!** *Microbiome* 2018, **6**(1):8.

507. Goris T, Hornung B, Kruse T, Reinhold A, Westermann M, Schaap PJ, Smidt H, Diekert G: **Draft genome sequence and characterization of *Desulfitobacterium hafniense* PCE-S**. *Standards in genomic sciences* 2015, **10**(1):15.

508. Nonaka H, Keresztes G, Shinoda Y, Ikenaga Y, Abe M, Naito K, Inatomi K, Furukawa K, Inui M, Yukawa H: **Complete Genome Sequence of the Dehalorespiring Bacterium *Desulfitobacterium hafniense* Y51 and Comparison with *Dehalococcoides ethenogenes* 195**. *Journal of bacteriology* 2006, **188**(6):2262-2274.

509. Palakawong Na Ayudthaya S, Hornung B, Ravikumar Varadarajan A, Plugge W, Plugge CM: **Draft Genome Sequence of *Actinomycessucciniciruminis* Strain Am4^{T}, Isolated from Cow Rumen Fluid**. *Genome announcements* 2017, **5**(29).

510. Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D: **Nucleotide Sequence Database Policies**. *Science* 2002, **298**(5597):1333.

511. Gerritsen J, Hornung B, Ritari J, Paulin L, Rijkers GT, Schaap PJ, De Vos WM, Smidt H: **Full genome sequence of *Romboutsia hominis* FRIFI^{T} and draft genome sequence of *Romboutsia lituseburensis* A25K^{T}**. *Standards in genomic sciences* 2018, **In preparation**.

512. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proceedings of the National Academy of Sciences of the United States of America* 1990, **96**:2896-2901.

513. Ross EM, Moate PJ, Marett L, Cocks BG, Hayes BJ: **Investigating the effect of two methane-mitigating diets on the rumen microbiome using massively parallel sequencing**. *Journal of dairy science* 2013, **96**(9):6030-6046.

514. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz Julia M, Reeves D, Gandara J, Chhangawala S *et al*: **Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics**. *Cell Systems* 2015, **CELS1**:1-15.

515. Barkin JA, Sussman DA, Fifadara N, Barkin JS: ***Clostridium difficile* Infection and Patient-Specific Antimicrobial Resistance Testing Reveals a High Metronidazole Resistance Rate**. *Dig Dis Sci* 2017, **62**(4):1035-1042.

516. Reysset G, Haggoud A, Sebald M: **Genetics of Resistance of *Bacteroides* Species to 5-Nitroimidazole**. *Clinical Infectious Disease* 1993, **16**((Suppl 4)):S401-403.

517. Karp PD, Caspi R: **A survey of metabolic databases emphasizing the MetaCyc family**. *Archives of toxicology* 2011, **85**(9):1015-1033.

518. Wickware P: **Next-generation biologists must straddle computation and biology**. *Nature* 2000, **404**:683.

519. Gammie A, Lorsch J, Singh S: **Catalyzing the Modernization of Graduate Education**. In: *NIGMS Feedback Loop Blog: A catalyst for interaction with the scientific community.* vol. 2018; 2015.

520. Dreyfuss E: **Want to make it as a biologst? Better learn to code**. In: *Wired.* New York: Conde Nast Publications; 2017.

521. **Biological Computation** [https://www.microsoft.com/en-us/research/group/biological-computation/]

522. **Google Genomics** [https://cloud.google.com/genomics/]

523. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC, Komarow HD, Program NCS *et al*: **Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis**. *Genome research* 2012, **22**(5):850-859.

524. Hajishengallis G, Liang S, Payne MA, Hashim A, Jotwani R, Eskan MA, McIntosh ML, Alsam A, Kirkwood KL, Lambris JD *et al*: **Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement**. *Cell host & microbe* 2011, **10**(5):497-506.

525. Strati F, Cavalieri D, Albanese D, De Felice C, Donati C, Hayek J, Jousson O, Leoncini S, Renzi D, Calabro A *et al*: **New evidences on the altered gut microbiota in autism spectrum disorders**. *Microbiome* 2017, **5**(1):24.

526. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung FE *et al*: **Towards standards for human fecal sample processing in metagenomic studies**. *Nature biotechnology* 2017, **35**(11):1069-1076.

Co-author affiliations

Vitor A.P. Martins dos Santos[1,2]

Peter J. Schaap[1]

Bernadet Renckens[5]

Ger T. Rijkers[7,8]

Klaudyna Borewicz[1]

Pieter H. van der Zaal[11]

Bartholomeus van den Bogert[3,12]

Caroline M. Plugge[3]

Hauke Smidt[3]

Jacoline Gerritsen[3,4]

Sacha A. F. T. van Hijum[5,6]

Willem M. de Vos[3,9]

Fangjie Gu[10]

Henk Schols[10]

Mark Davids[1]

[1] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, the Netherlands;

[2] LifeGlimmer GmbH, Markelstrasse 38, Berlin, Germany;

[3] Laboratory of Microbiology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, the Netherlands;

[4] Winclove Probiotics, Hulstweg 11, 1032 LB, Amsterdam, the Netherlands;

[5] Nijmegen Centre for Molecular Life Sciences, CMBI, Radboud UMC, P.O. Box 9101, 6500 HB, Nijmegen, the Netherlands;

[6] NIZO, Kernhemseweg 2, 6718 ZB, Ede, the Netherlands;

[7] Laboratory for Medical Microbiology and Immunology, St. Antonius Hospital, P.O. Box 2500, 3430 EM, Nieuwegein, the Netherlands;

[8] Department of Science, University College Roosevelt, P.O. Box 94, 4330 AB, Middelburg, the Netherlands;

[9] Departments of Microbiology and Immunology and Veterinary Biosciences, University of Helsinki, P.O. Box 66, 00014, University of Helsinki, Finland

[10]Laboratory of Food Chemistry, Wageningen University & Research, Bornse Weilanden 9, 6708 WG Wageningen, the Netherlands

[11]Biobased Chemistry and Technology, Wageningen University & Research, the Netherlands

[12] Top Institute Food and Nutrition (TIFN), P.O. Box 557, 6700 AN Wageningen, the Netherlands

Summary

The aim of this thesis was to elucidate how various microbial communities work, with a focus on next generation sequencing data.

The introduction in chapter 1 focuses on the history of biology, how the field of systems biology arose, and how the rise of nucleic acid sequencing has shaped a completely new field (among others), the microbiome research.

In chapter 2, an overview is given how the microbiota can be studied, in connection to metabolic syndrome and its sub-pathologies, including obesity, type II diabetes, elevated blood pressure, and dyslipidemia. We summarize which different methodologies (16S rRNA amplicon sequencing, metagenomics, metatranscriptomics) can be used to investigate the microbiome with different foci, and how as a next step the microbiome can be modelled, *in vitro* and *in silico*.

Chapter 3 describes the genome and transcriptome of the rat gut commensal *Romboutsia ilealis* CRIB^{T}. We characterized genomic properties, including those related to metabolism and sporulation abilities. The transcriptome allowed us to investigate the organism’s carbohydrate degradation abilities, including its potential regulation.

Chapter 4 is an investigation of an *in vitro* fermentation system, inoculated with human faecal material and the potential prebiotic Isomalto/malto-polysaccharides. The metatranscriptome of this system gave an insight into which genes are involved in the carbohydrate degradation, and which different types of organisms are involved and potentially need to cooperate for a full utilization of this carbohydrate.

In chapter 5, the cow rumen microbiota is investigated under different feeding regimes. The metatranscriptome of the cow rumen microbiota showed distinct patterns depending on the ratio of starch or cellulose enriched feed components, namely maize vs. grass silage. The increase in starch led to a decrease in methane emissions of the cow rumen microbiota, which was reflected in the metatranscriptomics data. Most notably, lower expression levels of genes encoding for proteins involved in methanogenic pathways of the rumen archaeon *Methanobrevibacter smithii* was observed.

The last chapter, the general discussion, mainly discusses the research described in this thesis with a focus on the relevant issues with modelling microbial communities, as well as overall scientific integrity in relationship with microbiome research.

List of publications

2015

- Tobias Goris, **Bastian Hornung**, Thomas Kruse, Anika Reinhold, Martin Westermann, Peter J. Schaap, Hauke Smidt, and Gabriele Diekert. „Draft genome sequence and characterization of *Desulfitobacterium hafniense* PCE-S". Standards in Genomics, 10:15, doi: https://dx.doi.org/10.1186%2F1944-3277-10-15

2017

- Susakul Palakawong Na Ayudthaya*, **Bastian Hornung***, Adithi Ravikumar Varadarajan, Wendy Plugge and Caroline M. Plugge. "Draft Genome Sequence of *Actinomyces succiniciruminis* Strain Am4^{T}*,* Isolated from Cow Rumen Fluid". Genome Announcements, 5(29): e01587-16, doi: https://dx.doi.org/10.1128%2FgenomeA.01587-16
- Jacoline Gerritsen*, **Bastian Hornung***, Bernadette Renckens, Sacha A.F.T. van Hijum, Vitor A.P. Martins dos Santos, Ger T. Rijkers, Peter J. Schaap, Willem M. de Vos and Hauke Smidt. "Genomic and functional analysis of *Romboutsia ilealis* CRIB^{T} reveals adaptation to the small intestine". PeerJ 5:e3698, doi: https://doi.org/10.7717/peerj.3698

2018

- **Bastian Hornung**, Vitor A.P. Martins dos Santos, Hauke Smidt and Peter J. Schaap and Hauke Smidt. "Studying microbial functionality within the gut ecosystem by systems biology". BMC Genes and Nutrition 13:5, doi: https://doi.org/10.1186/s12263-018-0594-6
- Siavash Atashgahi*, **Bastian Hornung***, Marcelle J. van der Waals, Ulises Nunes da Rocha, Floor Hugenholtz, Bart Nijsse, Douwe Molenaar, Rob van Spanning, Alfons J.M. Stams, Jan Gerritse and Hauke Smidt. "A benzene-degrading nitrate-reducing microbial consortium displays aerobic and anaerobic benzene degradation pathways". Scientific Reports 8:4490, doi: https://doi.org/10.1038/s41598-018-22617-x
- Loowee Chia, **Bastian Hornung**, Steven Aalvink, Peter J. Schaap, Willem M. de Vos, Jan Knol and Clara Belzer. "Deciphering the trophic interaction between Akkermansia muciniphila and the butyrogenic gut commensal Anaerostipes caccae using a metatranscriptomic approach." Antonie Van Leeuwenhoek 111(6):859-873, doi: https://doi.org/10.1007/s10482-018-1040-x
- Wiep Klaas Smits, Scott J. Weese, Adam P. Roberts, Celine Harmanus and **Bastian Hornung**. "A helicase-containing module defines a family of pCD630-like plasmids in Clostridium difficile." Anaerobe 39:78-84, doi: https://doi.org/10.1016/j.anaerobe.2017.12.005
- Amoe Baktash, Elisabeth M. Terveer, Romy D. Zwittink, **Bastian Hornung**, Jeroen Corver, Ed J. Kuijper and Wiep Klaas Smits. "Mechanistic Insights in the

Success of Fecal Microbiota Transplants for the Treatment of *Clostridium difficile* Infections". Frontiers in Microbiology 9:1242, doi: https://doi.org/10.3389/fmicb.2018.01242

- Jacoline Gerritsen, Aleksandr Umanetc, Ivelina Staneva, **Bastian Hornung**, Jarmo Ritari, Lars Paulin, Ger T. Rijkers, Hauke Smidt, Willem M. de Vos. "Romboutsia hominis sp. nov., the first human gut-derived representative of the genus Romboutsia, isolated from ileostoma effluent", International Journal of Systematic and Evolutionary Microbiology, doi: https://doi.org/10.1099/ijsem.0.003012
- **Bastian Hornung***, Bartholomeus van den Bogert*, Mark Davids, Vitor A.P. Martins dos Santos, Caroline M. Plugge, Peter J. Schaap and Hauke Smidt. "The Rumen Metatranscriptome Landscape Reflects Dietary Adaptation and Methanogenesis in Lactating Dairy Cows." bioRxiv, doi: https://doi.org/10.1101/275883

Submitted

- Jeroen Corver, Jeff Sen, **Bastian Hornung**, Bart Mertens, Eric Berssenbrugge, Celine Harmanus, Ingrid Sanders, Nitin Kumar, Trevor D. Lawley, Ed Kuijper, Paul Hensbergen and Simone Nicolardi. "Identification and validation of two peptide markers for the recognition of *Clostridioides difficile* MLST-1 and MLST-11 by MALDI-MS." Clinical Microbiology and Infection, submitted
- Jacoline Gerritsen*, **Bastian Hornung***, Jarmo Ritari, Lars Paulin, Ger T. Rijkers, Peter J. Schaap, Willem M. de Vos, Hauke Smidt. "Full genome sequence of *Romboutsia hominis* FRIFI^T and draft genome sequence of *Romboutsia lituseburensis* A25K^T", Standards in Genomic Sciences, submitted
- Jueeli Vaidya, **Bastian Hornung**, Hauke Smidt, Joan Edwards, Caroline Plugge. "Characterization of Propionibacterium ruminifibrarum sp.nov., isolated from cow rumen fibrous content" ", International Journal of Systematic and Evolutionary Microbiology, submitted

Manuscripts in preparation

- Klaudyna Borewicz*, **Bastian Hornung***, Fangjie Gu, Pieter H. van der Zaal, Henk A. Schols, Peter J. Schaap, Hauke Smidt. "Metatranscriptomics analysis indicates prebiotic effect of Isomalto/malto-polysaccharides on human colonic microbiota *in-vitro.*"
- **Bastian Hornung**, Romy Zwittink, Ed Kuijper. "Controls in microbiome research do not work"

*Contributed equally

Acknowledgements

Many things happened in the last 4+ years, and I am thankful for the good times here.

Ich möchte als erstes meiner Familie danken. Niemand hat eine Augenbraue verzogen (im negativen Sinne) als ich gesagt habe, dass ich in die Niederlande gehe. Und ich kann immer zurück gehen, ich weiß das, und es gut, das zu wissen.

The other most important „thanks" goes to my office mate Jasper. Man, without your help I wouldn't have continued with this work. You answered every stupid question I had, and gave your thoughts to every stupid (and sometimes hopefully not so stupid) thought I had.

I am grateful to my supervisors Hauke, Peter and Vitor that they trusted my abilities to become an independent scientist, and that they allowed me to travel this journey. I do not regret coming to this country, and would do this and this job again.

I want to thank you Maria, for answering all the random questions which I ever asked you. I hold your opinion in very high regard.

I need to thank my corridor mate Jael. You drove me crazy. It was hard, it was good, it was necessary. It changed me, and I don't know how life would have been without you.

My other corridor mates and their friends (Carlos, Pedro, Vicente, Karen, Evi, Violace, Phelia, Kenan, Lauriane, Matthias, Somsak, Yuwono, Andre, Marcos, Santiago, Allah, Irene, Chandra, Barrat), guys, I always thought living together was a necessity. And while it for sure is, I realized it can be fun too. We had good times together, while it was on a trip, or just sitting in the kitchen and chatting. I do not want to miss that time. We should have gone to Amsterdam and party though, instead of discussing about it every weekend until the last train was gone.

Additionally I want to thank Carlos and Santiago that they dragged me to the first Salsa classes I ever took, and motivated me to go the next time too. Without you, I'd have a big part of my social life missing.

Lauriane, I also want to thank you for listening to all the concerns which with I bothered you. Your opinion helped me a lot.

Ruben and Dorett, you were good friends in the time I needed them most. Without you it all would have gone differently.

Many other people, from the lab and from elsewhere, enriched my social life. We had long evenings in the de Zaaire, during barbecues (some involuntarily at my place), at the IC and on various trips. I will miss everyone. From the lab, in no particular order: Kalyanasundaram (to write it out, once), Nikolas, Benoit, Maarten, Sven, Erika, Rob, Emma, Irene, Aleks, Catalina, Lara, Monika, Sudarshan, Yuan, Cristina, Gerben, Johanna, Jueeli, Jeroen, Rebecca, Giulio, Teresita, Sabina, Angela, Alicia, Javier, Juanan, Vicente, Conall, Ivette, Nico (Puerto Rico). I remember wild parties, wet kisses (not mainly to me), long hugs, a barely evaded bar fight, long talks, nonsensical talks, and being lost in various places. I don't want to miss the cultural experiences, the drunk dancing experiences, also sober dancing, the pubquizes, the various parties and festivals elsewhere and here (especially on someone's rooftop). All of it was great (except when the alcohol level reached a new high, that was sometimes terrible. That mainly means the combo of Maarten, Benoit, Irene and Javier).

And Peer too, don't want to forget you again.

Catalina, thanks for being a good friend. I've spent a lot of time with you and I think you're one of the nicest people I have met.

Aleks, thanks for listening to everything stupid I had to say during and after parties. We also spent lots of time together, and I don't want to miss any of it.

Angela, I blame most of the good things, which happened in the time after you left, and also during it, on you and Alicia, Sudarshan and Aleks coming to Wageningen. I think without you the social life in MIB/SSB would have been very different.

Erika, you also have a special place in my heart. For a long time I have had very deep and nowhere going thoughts on friendship, how they change, how they develop, and how long it takes to have a good friend. The last question got answered. I considered you to be a very good friend already after two months after you arrived in Wageningen, and that changed a few things on how I thought.

Benoit, you big kid, you made many things just fun by being you, and that was great.

Maarten, you know, I actually really like you. Unless you're drunk. Please drink less.

Javier and Yuan… you squatted at my place way longer than I wanted. It was a good time.

Lara, thanks for being a good friend and sharing your thoughts with me.

Sven, als mein bester deutscher Freund bist du natürlich auch sehr wichtig. Du hast fast alles wichtige und unwichtige in meinem Leben festgehalten, und wegen dir ist es möglich dass alles nochmal zu erleben. Ich mag dich aber auch wenn du deine Kamera nicht dabei hast.

Kal, you're one of the weirdest persons I know, and because of you we talked about lots of weird things, and for some of them you also were the subject (e.g. the single fruit lunch diet). I hope you stay as you are.

Rob, I'm not unhappy that you are not here, because you are probably happy wherever you are right now. I'd still like to have a drink with you again.

Irene, you have an as dirty mind as me. And I loved it.

I am not writing a personal sentence for everyone. I hope nobody is hurt. You are still all special to me.

Some people I mainly know not from parties and long evenings in bars. But I had good times with you either as office mate (Bart, Niels, Jesse) during lunch discussions (Ioannis K, Ioannis M, Nico (my personal most favourite PhD student), Maria, Emmy, Daan, Hugo, Indra, Romy, Dennis, Anna, Carrie, Milad, Michael, Wen, Brendan, Shreyans, Melanie) and other events, from the lab and outside the lab, at housewarming parties (this mainly means Lennart), or in general as colleagues (Linde, Stamatis, Nhung, Nong, Tjerko, Agniezska).

Milad, you are the person who I know who can talk the most about the most random and unusual things. That was always fun, I never had a boring time with you.

I want to thank the three other PhD students who were with me on the same project (Niru, Neeraj, Balaji), and who had to share the same headaches as me and with whom I could share some of my problems.

Sjon, the party animal, also needs to be mentioned. Been with you to more big things than with any of the others, I think.

Also thanks to Wim and Caroline, who had to deal with all the small problems I caused and all the small requests I had.

Thanks to my co-authors (past and future). I hear often from people that they have problems with their co-authors, either about authorship/order, that people don't do their work, etc. I never had any real problems with you (Klaudyna, Jacoline, Tom, Loowee, Martijn, Joan, Jueeli, Irene, Thomas, Siavash, Susakul), working was most often pleasant with you (but sometimes complicated).

Some friends of friends also became valuable. I had good times (parties, dancing, discussions) with people who I got introduced to by my friends, or who turned out to be common friends, this time is also valuable to me. Thanks Gerli, Pamela, Agi, Claudia, Carlotta, Rui, Andrijana, and the Spanish people (Martha, Paula, Natalia). Camilla, you I got to know independently of our common friends, somehow random, and I don't want to miss you either.

And there are lots of people who I met elsewhere, during parties, during dancing, during sports, during courses, just random (but also from the lab). Many things would have been different without you. I want to thank: The Ecuadorians (Gabriel, Giovanni, Jhon, Luis), the Caribbeans (Kevin, Omar, Nathania, Tasneema, Jerry, Perly, Adrian) and "associated" people (Maarten, Michelle (Poei), Michelle (the Italian), Lucas, Francisca, Marianna), Adil, Sander, Marcel, Alex, David, Vicente, Moses (the Peruvian), Edgar, Lalo, Diana, Silvia, Juan, Gee, Sam, Pietro, Iskra, Franka, Daniel, the "ISOW people" and people who I got to know via the ISOW (Nele, Eleonora, Susana, Thomas, Linda, Angel, Nina, Gabor, Engin, Felix), the students from the Lab (Prokopis, Christos, Enrique, Max, Zac, Rodolfo, Valentine, Tianhe), the Salsa teachers (Jose, Lucia, Santos, Perci, Sofia, Imke, Pentcho), my dance partners (Celine, Geertje, Monique, Pam, Anna, Caroline, Gabi, Mimi, Sophie, Celia, Amber, Tanvi, Sharita, Sylke, Maddy, Amanda) and the guys who I got to know via this (Niels, Karl, Jordy), the gym people (Mark, Reinier, Manus, Moses (the Mexican), Phillip, Valerie, Shauna, Vincenzo, Hischam) and the Microbial Ecology department from NIOO, who I all got to know via various ways, but mostly not through work (Victor, Mattias, Afnan, Kay, Ruth, Noriko, Anna, Kesia, Vittorio). It was great to go anywhere in Wageningen and to run into you. No matter if it was at a party, in a dance class or at the gym, it was nice to have friends everywhere and to have a chat at all different kinds of occasions (even if it was between deadlifting sets).

Especially mentioned should be Gabriel, Giovanni and Jhon, who's company gave to many bystanders at many parties the impression I was a Latino, or at least from Spain.

You guys, and also Kevin, I could reliably meet at every party. Kevin, on some weekends I saw you more often than any of my best friends. That was weird, but I connect many good times with you too, like e.g. three days in a row partying at the IC.

It was also good that my too expensive apartment was used extensively as shelter for homeless people, or people who got locked out of their homes (or actually for planned temporary stay). In sort of chronological order: Phillipo, Carlos, Milad (uncountable times), Yuan, Jael, Javier, Natalia, Benoit x2, Giulio, Lara, Maarten.

At this place it should also be mentioned that I'm grateful to be in exchange being able to squat at some of your places, while either staying in Wageningen (Benoit + Yuan, Sven, Kal), or while being on holidays (Sabina, Dorett, Juanan, Lara).

I also want to thank the International Club, Café de Zaaier, Café Loburg, the Doctor Pub, Het Gat, The Spot, the H41, Villa Bloem, the Junushoff, the water house and roughly every second house in Droevendaal for having awesome evenings there and that we could party hard at so many occasions. Things like the Greek birthday, the Psytrance party or like every cramped party at the IC, these I keep in very good memory.

The ISOW and ISN also deserve some thanks for organizing so many enjoyable things, be it trips to different locations or nice parties.

Overall, when I am thinking back about the past years, I connect Wageningen and everyone I met there with good times. Thanks everyone.

Unacknowledgements

For some things I am not thankful though.

I do not want to thank the staff at BMC Microbiome, where one of my manuscripts was waiting for three months to have the editor have a look at it, and another three months to get rejected. For the same manuscript, I'm also not thanking the staff of BMC Genomics. While the manuscript was rejected very fast after only 2.5 months, nobody managed to have a look at my appeal to this decision for another three months.

Furthermore I'll also not thank the editor, who rejected my first ever submitted manuscript within only 30 minutes after submission. This was warranted, but it felt bad. I am a tad bit thankful though, because it was probably the only time ever that something of at least medium importance in this PhD came to a decision in a time frame which was less than one month.

SENSE

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

DIPLOMA

For specialised PhD training

The Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Bastian Volker Helmut Hornung

born on 24 August 1985 in Fulda, Germany

has successfully fulfilled all requirements of the
Educational Programme of SENSE.

Wageningen, 1 November 2018

On behalf of the SENSE board	the SENSE Director of Education
Prof. dr. Huub Rijnaarts	Dr. Ad van Dommelen

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The SENSE Research School declares that **Bastian Volker Helmut Hornung** has successfully fulfilled all requirements of the Educational PhD Programme of SENSE with a work load of 36.5 EC, including the following activities:

SENSE PhD Courses

- Environmental research in context (2013)
- Research in context activity: 'Co-producing accessible video presentation of content and possible impacts of the emerging scientific field of synthetic biology' (2014)
- SENSE writing week (2014)

Other PhD and Advanced MSc Courses

- Microme Workshop on Microbial Metabolism, European Bioinformatics Institute (2013)
- Genetics and physiology of food-associated microorganisms, VLAG graduate school, Wageningen (2013)
- IPOP Statistics course, Wageningen University (2014)
- Decision science, Wageningen University (2014)
- Datasharing, Dutch Techcentre for Life Sciences (2014)
- IPOP Network reconstruction course, Wageningen University (2015)
- Human microbiome in health and desease, TNO/Micropia (2015)
- Workshop on data ownership, Dutch Techcentre for Life sciences (2015)
- Scientific publishing, Wageningen University

Management and Didactic Skills Training

- Supervising BSc student with thesis entitled 'Development of an automated assembly pipeline for the production of high quality assemblies of prokaryotic genomes' (2015)
- Teaching in the MSc course 'Systems@Work' (2013-2014)
- Assisting in the MSc course 'Bioformation technology' (2015)

Oral Presentations

- *Flash presentation: Genomic and functional analysis of Romboutsia Ilealis CRIBT reveals adaptation to the small intestine*. SB@NL2014 Systems Biology Symposium, 15-16 December 2014, Maastricht, The Netherlands
- *Characterizing and understanding the rumen microbiota.* Dutch Bioinformatics & Systems Biology conference (BioSB), 20-21 May 2016, Lunteren, The Netherlands

SENSE Coordinator PhD Education

Dr. Peter Vermeulen

The research described in this thesis was financially supported by the
IPOP program Systems Biology of Wageningen University & Research.

Printed by Digiforce and ProefschriftMaken, Wageningen.