TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

ZIEL – Institute for Food & Health

Core Facility Microbiome

Improved detection methods to assess microbial communities using high-throughput sequencing

Isabel Sophie Abellan Schneyder

Vollständiger Abdruck der von der promotionsführenden Einrichtung TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat)

genehmigten Dissertation.

Vorsitzender: apl. Prof. Dr. Michael Pfaffl

Prüfer der Dissertation:

1. Priv.-Doz. Dr. Klaus Neuhaus
2. Prof. Dr. Lindsay Hall

Die Dissertation wurde am 29.06.2021 bei der Technischen Universität München eingereicht und durch die promotionsführende Einrichtung TUM School of Life Sciences der Technischen Universität München am 18.10.2021 angenommen.

List of content

Zusammenfassung

Zur Analyse mikrobieller Populationen wird heutzutage, neben Kultivierungstechniken, eine Sequenzierung der Gene der 16S beziehungsweise der 18S rRNS verwendet. Dadurch können Bakterien, Archaeen und eukaryotischen Mikroorganismen (z.B. Hefen und Pilze) nach Abgleich mit Referenzdatenbanken bestimmt werden. Meist erfolgt eine solche Analyse durch die Sequenzierung von kurzen Sequenzabschnitten mit einer Größe von 300-600 bp. Durch die technischen Entwicklungen der letzten Jahre ist es nun auch möglich, längere Sequenzfragmente zu generieren. Diese können entweder durch das Sequenzieren auf einem sogenannten *long-read*-Sequenzierer entstehen oder es werden synthetische Volllängenfragmente *de novo* nach der Sequenzierung auf einem *short-read*-Sequenzierer assembliert. In der vorliegenden Studie wurden mittels verschiedener Methoden entweder kurze, lange oder synthetische Volllängen-Sequenzfragmente erzeugt. Die dabei angewandten Methoden wurden verbessert oder teilweise neu entworfen und getestet. Dafür wurden künstliche mikrobielle Gemeinschaften (sogenannte *mock communities*), humane Fäzesproben und Kuhmilchproben herangezogen. Für die Sequenzierung kurzer Fragmente haben sich nach der Testung verschiedenerer Oligonukleotide, Analyseverfahren und Referenzdatenbanken die Verwendung von Oligonukleotidsequenzen, welche die variablen Regionen V3-V4 des 16S rRNS Gens amplifizieren als am besten herausgestellt. Die Verwendung von Silva oder RDP als Referenzdatenbanken sowie das Clustern der Sequenzen mit Hilfe von Programmen, die über sogenannte *denoising* Schritte verfügen, zeigten die besten Ergebnisse.

Lange oder synthetische Volllängen-Sequenzfragmente des 16S/18S rRNS Gens sind von Vorteil, wenn eine Klassifizierung der Mikroorganismen bis hin zum Spezieslevel angestrebt wird. Mittels synthetischer Volllängensequenzierung (LoopSeq) konnte gezeigt werden, dass die Spezies-Klassifizierung im Vergleich zu den Ergebnissen, die durch die Sequenzierung kurzer Fragmente entstanden, verbessert wurde. Des Weiteren konnte gezeigt werden, dass diese Methode zur Identifizierung von gegebenenfalls pathogenen Mikroorganismen, zum Beispiel zur Identifizierung von Mastitis-Erregern in Milch herangezogen werden kann. Gleichzeitig wurde eine eigene Methode zur Sequenzierung von synthetischen Volllängen entworfen und evaluiert. Durch die hier vorgestellte Implementierung eines *digital droplet PCR* (ddPCR) Schrittes nach der Herstellung einer Genbibliothek für die Sequenzierung können nun deutlich verringerte Mengen an genomischer DNS (gDNS) für die Erzeugung der Bibliothek verwendet werden. Trotz niedriger initialer gDNS Mengen von unter 1 ng kann durch den ddPCR Schritt eine erfolgreiche Sequenzierung gewährleistet werden. Dies ist besonders für bestimmte Umweltproben interessant, die nur sehr geringe Mengen an Mikroorganismen aufweisen, wie zum Beispiel Wasser- oder Milchproben.

Abstract

The analysis of microbial populations is primarily performed by either culture-dependent technologies or by sequencing approaches. If sequencing is applied, mostly the 16S and 18S rRNA genes are targeted to identify bacteria, archaea, and eukaryotic microorganisms such as yeasts and fungi. For determining the different microbial genera and species, the produced sequencing reads are compared to reference databases. Sequencing of the microbiota is usually performed by short amplicon sequencing targeting a small proportion of the 16S rRNA gene. Usually, amplicons of 300-600 bp in size are produced and sequenced.

Nevertheless, several new technologies arose in the last couple of years that allow the sequencing of longer fragments. Either long-read sequencing approaches enable sequencing of up to several thousand base pairs, or so-called synthetic long-read methods are applied. For the latter, samples are fragmented and then sequenced on a short read sequencer, followed by a *de novo* assembly, which allows obtaining the full-length sequence.

In this thesis, different sequencing approaches (including short, long, and synthetic long-read sequencing approaches) were tested, created, evaluated, and compared. For this, mock communities of known composition, human fecal samples, and bovine milk samples were used as targets. Short amplicon sequencing approaches were tested for their best performance concerning primers, pipelines, and parameters. This evaluation showed that targeting the variable regions (V-regions) V3-V4 of the 16S rRNA gene allowed the most accurate analysis of human fecal samples and mock communities. In further testing, it was found that RDP or Silva performed best as reference databases. Further, clustering approaches that include denoising steps such as those producing amplicon sequence variants (ASVs), or zero-radius operational taxonomic units (zOTUs) performed better than the standard OTU clustering approach. This study uses both long-read and synthetic full-length approaches to estimate whether the species-level classification could be improved compared to short amplicon sequencing. Indeed, I could show that by applying a synthetic full-length approach (LoopSeq), species-level classification was enhanced compared to short amplicon sequencing. Moreover, as a proof of concept, I showed that it is possible to identify putative mastitis pathogens at species-level in milk samples tested. Besides that an in-house synthetic full-length sequencing approach was established, which was further evaluated and tested here.

By introducing a digital droplet PCR (ddPCR) step after library preparation, I could show that dramatically decreased input amounts, compared to traditional methods, of genomic DNA (gDNA, < 1 ng) are sufficient for successful library preparation and sequencing. Thus, when using ddPCR, 16S rRNA gene sequencing becomes accessible to low biomass samples (e.g., water or milk samples) as they often fail in preparing a sequencing library due to insufficient initial gDNA amounts.

1. Introduction

1.1 Microorganisms, microbiota, and microbiomes

Microorganisms are generally defined as organisms that can be seen only through a microscope. This group includes bacteria, protozoa, algae, and fungi. It is debated if viruses should be included (Wessner, 2010). Nevertheless, microorganisms are present everywhere on earth, and sterility in any biological sample is extremely rare (Madigan *et al.*, 2014, p. 31). Different predictions about the actual number and variety of species of microorganisms on our earth exist. Through the years, estimations varied from 10^5 - 10^7 prokaryotic species (Whitman *et al.*, 1998) to 1 trillion (10^{12}) microbial species (Locey and Lennon, 2016). On earth, the total estimated number of microbial cells is ~2.5×10^{30}. It is supposed that 66% of all microorganisms can be found in marine subsurface, 26% on terrestrial subsurface, 4.8% on soil surfaces, 2.2% in the oceans, and 1% in all other habitats. Those other habitats include freshwater and salt lakes, domesticated animals, sea ice, termites, humans, and domesticated birds (Madigan *et al.*, 2014, pp. 32-34).

All microbial taxa that are associated with a distinct environment are considered microbiota (of this habitat). However, the term microbiome includes, besides the catalog of the species of microbes present, also their genetic material (Cho and Blaser, 2012, Marchesi and Ravel, 2015, Ursell *et al.*, 2012, Whiteside *et al.*, 2015).

1.1.1 A brief history of microbiome research

Robert Hooke published in 1665 his book "*Micrographia"* which was the first illustrated book on microscopy. The book contained 60 observations of objects that he investigated microscopically. Therefore, Hooke is assumed to be the first person to describe and publish microorganisms in detail (Gest, 2004). In the 1670s, Antonie van Leeuwenhoek was the first to report little animals or *animalcules*, which were later recognized as bacteria and protists (Lane, 2015). Moreover, he was the first to describe the human-associated microbiota as he reported and drew in 1680 bacteria that he had extracted from a human mouth (Egerton, 2006). Furthermore, he analyzed his own stool under the microscope and found several "*animalcules*" (Farré-Maduell and Casals-Pascual, 2019). In the 1840s, some bacteria of the gastrointestinal tract were described. For instance, John Goodsir described a microorganism that was obtained from the stomach of a patient. He further proposed that these microorganisms were the source of the patient's illness condition. But this assumption was refused amongst others by Friedrich Theodor von Frerichs, a German pathologist who also described microorganisms in the digestive tract but considered those to be harmless (Farré-Maduell and Casals-Pascual, 2019). In 1853, Joseph Leidy published his work "*A Flora and Fauna within living animals" (Leidy, 1853)*, which is often referred to be a starting point of microbiota research (Trautwein, 2020).

The late 19th century and the early years of the 20th century represent the golden age of microbiology, where significant advancements in microbiology and bacteriology were made (Blevins and Bronze, 2010, Friedmann, 2014, Moxon, 1997). This time was essential for microbiota and microbiome research as well since many fundamental working techniques were developed, and hypotheses were made and confirmed.

Robert Koch founded, with the so-called *Koch's postulates* and his research work, the field of medical bacteriology. Of importance for all further microbial studies were his developments and achievements in the cultivation of microorganisms. He developed culturing strategies and techniques to study and observe changes in bacterial cultures over time (Blevins and Bronze, 2010). Another influential researcher of that time was Louis Pasteur, who demonstrated that living organisms do not originate from non-living matter. Moreover, he set several fundamentals in microbiology. He invented techniques such as sterilization, aseptically working, and developed several vaccines (Smith, 2012, Walden, 2003).

In 1885, Theodor Escherich published a study based on the isolation of one of today's most studied bacterial species. He named it *Bacterium coli* commune, which is today known as *Escherichia coli*. Escherich isolated the strain from neonatal stool samples and described it as common among neonates (Martinson and Walk, 2020). Only a year later, Escherich published his habilitation thesis entitled "*The intestinal bacteria of neonates and their relationship to the physiology of digestion"* (Hacker and Blum-Oehler, 2007). There, he described the bacterial composition of the infant's intestinal tract. With this work, Escherich started to become one of the leading bacteriologists of his time (Farré-Maduell and Casals-Pascual, 2019, Hacker and Blum-Oehler, 2007).

Henry Tissier, who worked at the Pasteur Institute in France, isolated bacteria from healthy babies' stool. The bacteria that dominated healthy infants' stool were Y-shaped and are now known to be *Bifidobacteria*. He proposed administering those bacteria to babies suffering from diarrhea, as he observed that those babies lacked the Y-shaped bacteria (Leahy *et al.*, 2005, Siezen and Wilson, 2010). Also working at the Pasteur Institute, Ilya Mechnikov proposed in 1908 that undesirable gut bacteria could cause premature aging of tissues and organs, which could be prevented or delayed through the admission of beneficial microorganisms. With that, he laid the fundamentals for future research on probiotics (Farré-Maduell and Casals-Pascual, 2019, Gordon, 2016, Podolsky, 2012).

In 1909 Arthur Kendall published his study on "*Some observations on the study of the intestinal bacteria"* (Kendall, 1909). There, he described his comments on stool microbiota and the effect of diet on the intestinal microbiota of monkeys (Aziz, 2009).

In 1916, the German Alfred Nissle published a study where he showed that specific *Escherichia coli* (*E. coli*) isolates could inhibit the growth of co-cultured *Salmonella* and other enteropathogens. Nissle named this phenomenon *antagonistic activity*. He investigated further

how and if antagonistic *E. coli* strains could be used to treat intestinal diseases such as diarrhea (Nissle, 1916, Sonnenborn, 2016). One of the nowadays most commonly and frequently used probiotic strains is the *E. coli* Nissle 1917 strain (Wassenaar, 2016). This strain was shown to have an intestinal anti-inflammatory effect, but the mechanism of action is still not fully understood (Scaldaferri *et al.*, 2016). Nevertheless, with this finding, Nissle contributed to the history of studying the human microbiota, but he also showed the potential for further investments in an industry promoted to support human health.

Until the mid-1900s, it was only possible to grow bacteria aerobically on defined media on Petri dishes or in liquid culture. In 1944, Robert E. Hungate published a method that allowed him to cultivate an anaerobic *Clostridium* strain from the bovine rumen (Hungate, 1944). The complete experimental procedure to culture anaerobic bacteria was published in 1950 and the culturing approach to adapt for anaerobic methanogens in 1969 (Hungate, 1950, Hungate, 1969).

The mid-20th century is known as the golden age of antibiotic discoveries, as one-half of the drugs commonly used today were discovered in this period (Davies, 2006). At the same time, bacterial infections with *Clostridium difficile* were treated for the first time through fecal microbiota transplantation. Eiseman *et al.* (1958) treated four patients that were previously shown to not respond to any other treatment tested, with fecal samples of healthy donors. Those microbiota transplantations aimed to reinstall a normal function of the gut microbiota. Today, this method is especially favored for patients that show severe and complicated forms of *Clostridium difficile* infections and showed no lasting response to treatments with antibiotics (Baunwall *et al.*, 2020, Khoruts and Sadowsky, 2016).

In order to move from observation to experimental settings, mouse models became an important tool in microbiome research. Indeed, large mouse experimental studies assessing the role of the gut microbiota in physiology and disease are nowadays performed all over the world. This became only feasible after establishing techniques that allow generating germ-free mice. In 1965, Schaedler *et al*. described the first transfer of bacterial culture to germ-free mice. Thus, the effects of distinct bacterial species or fecal microbiota can be studied in detail, and functional microbiota-host relationships can be explored (Park and Im, 2020, Schaedler *et al.*, 1965).

During the late 1970s and the beginning of the 1980s, first publications aiming to find a molecular characterization method to distinguish bacterial genera arose. The ribosomal RNA (rRNA) was shown to be usable for bacterial phylogeny (Woese and Fox, 1977). Moreover, Woese & Fox (1977) were the first describing the three domains of life, archaea, bacteria, and eukarya, which they could differentiate through the use of the rRNA as a phylogenetic and evolutionary marker (Zhulin, 2016). In the mid-1980s, the 16S rRNA gene was first described to be used for the taxonomical differentiation of bacteria. Woese *et al.* (1985) described that

they could differentiate ten major groups of *Eubacteria* by using the 16S rRNA gene. Shortly afterward, Lane *et al.* (1985) published the first method for efficient 16S gene sequencing.
Concerning sequencing, several developments were necessary before reaching the high-throughput sequencing we are used to today. The first widely used sequencing method had been developed by Sanger in 1977 (Sanger *et al.*, 1977). In addition, the polymerase chain reaction (PCR), necessary to obtain enough DNA from template molecules, had been described mid-1980s (Mullis *et al.*, 1986, Mullis and Faloona, 1987). In any case, high-throughput sequencing of the 16S rRNA gene became only possible with the so-called next-generation sequencers. In 2005, 454 Life Sciences, a US-based biotechnology company, released the first next-generation sequencer (Margulies *et al.*, 2005). Shortly after, the Solexa Genome Analyzer (2006) and the SOLiD sequencer (2007) followed. Finally, the human microbiome project (HMP), which started in 2007, aimed to develop research resources and analyzed if changes in the microbiome have an important impact on health and disease (NIH HMP Working Group *et al.*, 2009). Thus, this project played a fundamental role in and for all subsequent microbiome studies. Lately, new sequencing techniques such as long-read sequencing (Branton *et al.*, 2008, Eid *et al.*, 2009), whole-genome (Gilissen *et al.*, 2014), metagenome assembly (Ghurye *et al.*, 2016), and shallow shotgun (Hillmann *et al.*, 2018) sequencing gained interest by the community and each technique will certainly further impact microbiome research.

1.1.2 The human gut microbiome

The intestinal microbiota of a human is composed of approximately 4×10^{13} microorganisms and is, therefore, similar to the number of human body cells (Sender *et al.*, 2016). The human microbiome is, per definition, the combined genetic material of a complex interacting entity composed of bacteria, archaea, viruses, and eukaryotic microbes that live inside the intestinal tract, on the skin, and in other places of the body. Nevertheless, definitions vary as phages, viruses, plasmids, and mobile elements are not always considered as parts of the microbiome (Berg *et al.*, 2020, Gill *et al.*, 2006).
The most densely colonized organ of the human body is the gastrointestinal tract (GI), which harbors a diverse and dynamic community of different microorganisms (Coleman and Haller, 2018). The adult human gut microbiota consists of hundreds to thousands of different bacterial species, mostly belonging to the phyla of *Bacteroidetes*, *Firmicutes*, *Actinobacteria,* and *Proteobacteria* with major inter-individual variations (Donaldson *et al.*, 2016, Thursby and Juge, 2017). Parts of those differences might be explained by different environmental impact factors such as location, physiological status, medication, or yet unknown impact factors (Turnbaugh *et al.*, 2007).

Generally, high microbiota diversity and temporal stability have been associated with health. A relative lack of diversity, on the other hand, has been linked to diseases like inflammatory bowel disease (IBD), obesity, and diabetes (Lloyd-Price *et al.*, 2016). Concerning the temporal and intra-individual stability, it was shown that samples, which originated from the same individual are more similar to each other than samples obtained from different individuals. Thus, it is expected that healthy humans have, to some degree, a stable gut microbiota (Caporaso *et al.*, 2011, Lozupone *et al.*, 2012, Turnbaugh *et al.*, 2009). Faith *et al.* (2013) could show that of the about 200 strains which they found in each of their 37 human-associated stool microbiota samples, on average, 60% of the strains stayed stable over their sampling period of five years. A more recent study by Lloyd-Price *et al.* (2017) confirmed the previous findings and stated that the gut microbiota of each individual is highly personalized.

Dysbiosis describes a change and imbalance in the gut microbial composition (Petersen and Round, 2014). Such a state is associated with many diseases like IBD, Crohn's disease, ulcerative colitis, colorectal cancer, diabetes, asthma, allergies, or even autism (De Almeida *et al.*, 2019, Mahnic *et al.*, 2020, Parracho *et al.*, 2005, Petersen and Round, 2014). Moreover, Petersen & Round (2014) defined three categories into which dysbiosis can be classified. The first is marked by the loss of beneficial microorganisms, the second by the expansion of potentially harmful microorganisms or pathobionts, and the third by a general loss of microbial diversity. The general therapeutic goal after dysbiosis was determined as to regain a normal or balanced microbiota. Thus, if dysbiosis is suspected, it is aimed to improve the gut microbiota diversity, e.g., by fecal microbiota transplantation (Mosca *et al.*, 2016).

Even though several studies try to aim to define a healthy and normal gut microbiome, no general accordance with parameters or measures has been described. McBurney *et al.* (2019) stated that a valid definition of a healthy microbiome could only be derived if we would overcome the lack of validated biomarkers. One needs to be able to precisely define and measure microbiome-host interactions. This shows that technical improvements and standardization of research approaches are necessary to reliably study functional impacts on the microbiome and its effects on human health and disease.

1.1.3 The bovine milk microbiome

Not long ago, many scientists believed that milk inside the bovine udder, which is considered the intramammary microbiota, would be sterile (Rainard, 2017). Nevertheless, recent studies rather suggest that intramammary microbiota is a low biomass sample (i.e., low numbers of microorganisms) and, thus, sterility is to be doubted (Oikonomou *et al.*, 2014, Oikonomou *et al.*, 2012). Nowadays, it is assumed that the mammary gland of adult cows represents an open and functional system that is directly connected to the environment (Taponen *et al.*, 2019). As most of the recent bovine milk studies used expressed milk, i.e., samples from milk that was

already outside of the mammary gland, the amount of microorganisms inside the mammary gland still needs to be investigated (Oikonomou *et al.*, 2020). Nevertheless, Metzger *et al.* (2018) could show that milk, which was taken from the bovine cistern, was not sterile. Here, future studies that can confirm these primary findings are needed.

As studies vary in layout and location, it is still difficult to find or define a shared healthy bovine milk microbiota. Nevertheless, *Staphylococcus*, *Streptococcus*, *Pseudomonas*, *Bifidobacterium*, *Propionibacterium*, *Bacteroides*, *Corynebacterium*, and *Enterococcus* were found to be commonly present taxa in the bovine milk microbiota (Addis *et al.*, 2016, Derakhshani *et al.*, 2018, Oikonomou *et al.*, 2020, Oikonomou *et al.*, 2014).

Generally, the investigation of the bovine milk microbiota is complex. Standard cultivation-based techniques are, at times, 20-30%, according to Taponen *et al.* (2019), culture-negative, even though samples were collected from cows diagnosed with clinical mastitis. Metagenomic studies, however, detect several microorganisms within those samples (Taponen *et al.*, 2019). Thus, it must be investigated in detail whether divergence is due to the analysis technique or based on contamination.

The benefits of cultivation-based techniques are the identification of unknown taxa and a good standardization, whereas disadvantages are very time-consuming analysis and an undetermined number of microorganisms due to those that are not culturable in standard conditions. The modern 16S rRNA gene sequencing approaches allow for an easy and convenient PCR-based approach, which can be used in high-throughput, provides higher resolution of analysis, and is thus more time-efficient. Nevertheless, training and bioinformatical knowledge are needed. Moreover, artifacts that arose through methodological biases must be studied and can sometimes not be fully evaluated (Addis *et al.*, 2016, Breitenwieser *et al.*, 2020, Dahlberg *et al.*, 2019, Dahlberg *et al.*, 2020). Besides the investigation method, several biasing factors influencing the bovine milk microbiota have already been described. As mentioned above, the way how and in which format raw milk is collected and potentially transferred or stored is important (Dahlberg *et al.*, 2020, Hiitiö *et al.*, 2016, Kable *et al.*, 2016, Kennang Ouamba *et al.*, 2020). Housing, seasonal changes, and the health status of the animals should be kept in mind (Doyle *et al.*, 2017, Du *et al.*, 2020, Kuehn *et al.*, 2013, Li *et al.*, 2018, Lima *et al.*, 2018, Metzger *et al.*, 2018, Oikonomou *et al.*, 2014, Parmar *et al.*, 2020).

Mastitis is an inflammation of the mammary gland and a disease, which economically impacts the dairy industry worldwide (Abebe *et al.*, 2016). It is the most common disease in dairy cows and can lead to a decrease in life span, herd productivity, and milk production of dairy cows (Francoz *et al.*, 2017). Mastitis is identified either by symptoms such as changes in milk secretion, udder swelling, reddening, pain, a rise in body temperature or disorder, by an increased somatic cell count (SCC), which estimate the number of immune cells in the milk or

by traditional plate-culture techniques targeting bacteria (Brennecke *et al.*, 2021, Halasa and Kirkeby, 2020). Besides those measures, a limited number of other (sub-)clinical detection methods were developed. Examples include the identification of mastitis-related microorganisms using qPCR, MALDI-TOF, the California mastitis test, or diagnostic methods such as the UdderCheck, which is based on the detection of enzymatic activity, or the Milk Checker, which measures electrical conductivity (Martins *et al.*, 2019). Nonetheless, 16S rRNA gene sequencing became a reliable and fast method over the years, potentially allowing a time-efficient long-term screening perspective. Hence, besides the time-consuming, often inefficient, and limited plate-counting approach, the techniques mentioned above suffer in classifying the complete microbiota at the species level, while this would be of great interest. Previously, it was shown that only some species are suspected to be mastitis-causing or mastitis-related strains (Dufour *et al.*, 2019), namely *Streptococcus uberis*, *Streptococcus agalactiae, Streptococcus dysgalactiae, Staphylococcus aureus, Corynebacterium bovis, Escherichia coli, and Klebsiella pneumoniae* (Bolte *et al.*, 2020, Cobirka *et al.*, 2020, Dalanezi *et al.*, 2020, Dufour *et al.*, 2019). Thus, a method allowing to detect and ideally quantify putative mastitis pathogens at species-level would be ideal.

1.2 The 16S rRNA gene as a taxonomical marker

In the 1970s, the rRNA genes were found to be perhaps useful as an evolutionary marker gene (Woese and Fox, 1977). The rRNAs are primary components of the ribosome, which consist of a small subunit (SSU) and a large subunit (LSU), forming an RNA/protein complex for protein production from mRNA (Fox, 2010). In bacteria, the SSU is made up of 21 ribosomal proteins and the 16S rRNA. The LSU, however, is composed of 34 ribosomal proteins, the 23S and 5S rRNA (Berg *et al.*, 2013, p. 1071). In contrast, the archaeal ribosome is composed of 50-70 proteins, depending on the species, the 16S, 23S and 5S rRNA (Londei, 2010). In eukaryotes, 80 proteins (in yeast 79) and the 18S, 23S, 5S and 5.8S rRNAs, respectively, form the ribosome (Wilson and Doudna Cate, 2012).

At least one 16S rRNA gene copy can be found in all prokaryotes. This, and other factors such as structural and functional conservation and a sufficient size make the 16S rRNA gene a use- and powerful molecular marker gene for deep taxonomical designation (Ludwig and Schleifer, 1994, Wang and Qian, 2009). The bacterial 16S rRNA has an approximate size of 1,550 bp and is composed of nine so-called variable regions (V-regions), interspaced by conserved regions. The V-regions show noticeable sequence diversity when comparing different bacteria. Thus, those can be used as a molecular marker to differentiate between different bacterial species. Conserved regions, which were shown to be evolutionary conserved, flank the V-regions and enable PCR amplification (Figure 1.2.1). Nonetheless, V-regions show different

extents of variability, and thus, none of the V-regions allow for differentiation of all bacteria (Baker *et al.*, 2003, Chakravorty *et al.*, 2007, Clarridge, 2004). Further, the conservation of the conserved regions is not absolute and differences between bacteria exist.

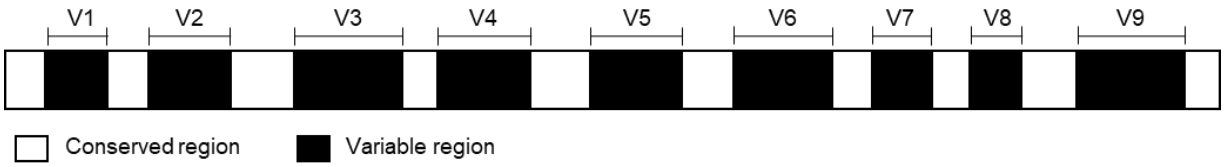

Figure 1.2.1: Structure of the 16S rRNA gene. The 16S rRNA gene is composed of conserved regions (white) and variable regions (black).

Even though it was shown that none of the V-regions could differentiate between all bacteria, some regions allow for a better taxonomic resolution than others (Bukin *et al.*, 2019). Moreover, it has to be considered that taxonomic resolutions of bacteria seem to differ between different V-regions, depending on the evolutionary speed of divergence for each V-region in each bacterial group (Pinna *et al.*, 2019, Yang *et al.*, 2016).

1.3 Amplicon sequencing strategies

1.3.1 The use of the 16S rRNA genes in amplicon-based studies

Initially, 16S rRNA gene sequencing was performed using Sanger sequencing, which uses dideoxynucleotides to block chain extension and labeled precursors for detection. Even though Sanger sequencing was greatly improved after its initial description (Sanger *et al.*, 1977), it remained costly and time-consuming (Heather and Chain, 2016). In the late 1990s, a new sequencing technique, pyrosequencing, was described. Fluorescently labeled nucleotides were used to perform a sequencing by synthesis reaction (Ahmadian *et al.*, 2006, Ronaghi *et al.*, 1998). The first commercially available sequencer was based on this pyrosequencing technique and was released in 2005. This changed the sequencing research dramatically, as it allowed for parallel sequencing of several sequencing reactions. Thereby, the sequencing output was drastically increased compared to Sanger sequencing (Heather and Chain, 2016, Margulies *et al.*, 2005). Shortly after the release of this machine, the 454, other nowadays so-called second-generation sequencers followed. Namely, those were Solexa/Illumina's Genome Analyzer and Applied Biosystems SOLiD technology, the latter refers to sequencing by oligonucleotide ligation and detection (Voelkerding *et al.*, 2009).

The Illumina MiSeq sequencer is, until today, one of the most purchased and used sequencing machines. It was released in 2011 and allows to sequence up to 600 bp in a 2x 300 bp paired-end mode on the MiSeq model, generating up to 15 Gb of data in a runtime of about 56 hours. With an error rate of ~0.1% (Ardui *et al.*, 2018), the MiSeq is very precise and, thus, a useful tool in microbiota analysis. As the read-length limit of Illumina MiSeq and other second-

generation sequencers is between 400-600 bp, 16S rRNA gene analyses are mostly performed on one, two, or three adjunct V-regions (Bukin *et al.*, 2019).

For microbiota analysis, so-called amplicon approaches are used. Amplicon refers to DNA products of a polymerase chain reaction (PCR). The general amplicon sequencing approach for microbiota samples works as described in the following. First, samples are collected and either stabilized in stabilizer fluids or flash frozen. Samples can be stored, the colder, the better (ideally at -80°C). Next, genomic DNA (gDNA) is extracted from samples. Fixed amounts of gDNA are used in a first step PCR, which amplifies selected V-regions. The primers for that first step PCR are composed of the sequence that anneals to the constant regions and of sequences referring to an overhang. This overhang is then targeted in a second step PCR, where barcoding can be applied. Thus, every sample gets a unique combination of forward and reverse barcoding primer, which later allows to identify and categorize the sample. After this second step PCR, products are checked for sufficient quality, i.e., no side product or impurities should be visible by using agarose gel electrophoresis. Samples are cleaned up using, for instance, magnetic beads, the concentration is measured, and samples are normalized. After the pooling, the so-called library is denatured, diluted, and loaded on the cartridge for sequencing (Figure 1.3.1).

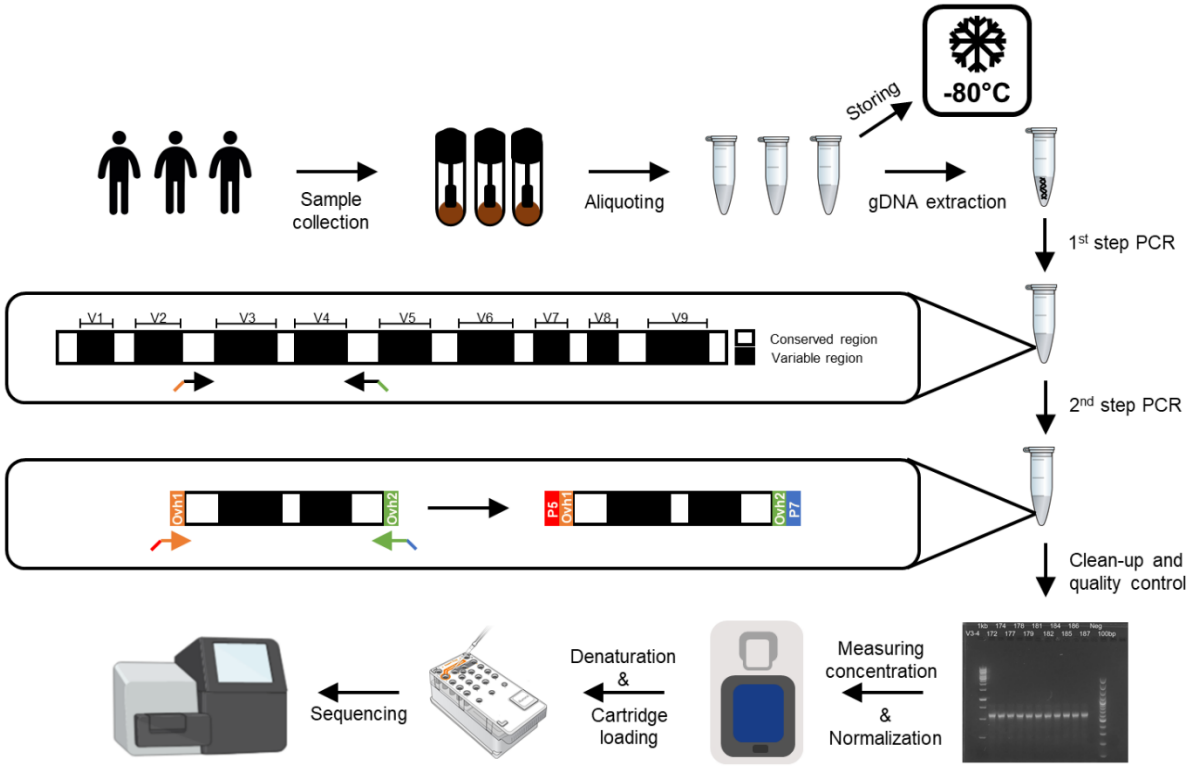

Figure 1.3.1: Scheme of 16S rRNA gene sequencing. First, samples must be collected, aliquoted, and then either stored or gDNA has to be extracted. 1st-step PCR targets the selected V-regions, whereas 2nd step PCR anneals barcodes to each product, allowing for multiplexing. Samples are cleaned up, and quality is controlled through the illustration of gel electrophoresis. The concentration of the amplicons is measured using a Qubit fluorometer, and normalization of the samples in equimolar ratios is performed. The final pool is denaturized and loaded on the sequencing cartridge before sequencing is performed.

Of note, both PCR steps can be conducted in one, but a two-step PCR allows for decreased detection levels (i.e., using less gDNA) and avoids PCR by-products since the cycle number is limited in each step (Berry *et al.*, 2011).

Even though standardization of short amplicon 16S rRNA gene sequencing protocols was improved over the last couple of years, experimental variance and differences in the study performance still exist. The human microbiome project (HMP), which was a research initiative from 2007 – 2016 to improve the understanding of the human microbiota in health and diseases, set several standards and illuminated possible pitfalls in microbiota or microbiome-related research (NIH HMP Working Group *et al.*, 2009, Turnbaugh *et al.*, 2007). Some of those pitfalls evaluated within the HMP are, for example, effects on the bacterial community due to storage conditions (Lauber *et al.*, 2010), differences in DNA extraction protocols (Sinha *et al.*, 2015), design and use of different 16S rRNA gene primer (Nossa *et al.*, 2010), sequencing machinery (Jumpstart Consortium Human Microbiome Project Data Generation Working, 2012), exact taxonomical classification, and improved data analysis pipelines (Edgar *et al.*, 2011, Schloss *et al.*, 2011). Until today, effects and preventive actions to improve the reliability of 16S rRNA gene sequencing studies are intensively researched and assessed. Nevertheless, until today variations in study design, performance and evaluation limit the comparability between microbiota studies and therefore, always need to be protocolled.

1.3.2 Factors Limiting 16S rRNA gene amplicon analysis

Unquestionably, 16S rRNA gene sequencing improved drastically within the last 30 years due to improvements in study design, sequencing strategy and capacity, and improved analysis pipelines. Nonetheless, several biasing factors must still be evaluated before studies are designed and subsequently performed. In this section, the following biasing factors are described in detail: sampling and sample storage, extraction method, primer choice and V-region issues, amplicon generation and sequencing, sequencing machinery, and bioinformatical analysis and reference databases.

Sample collection must be performed under clean and / or sterile conditions to prohibit sample or cross-contamination. The containers and equipment, such as gloves, spades, spoons, forceps, or liquid samplers, should be packaged accordingly (Gołębiewski and Tretyn, 2020). The detailed sample procedure is dependent on sampling content and study-specific concerns. Nevertheless, several important considerations were already previously discussed. This includes the sampling site, the impact of biomass (Bender *et al.*, 2018, Karstens *et al.*, 2019), the homogenization (Fidler *et al.*, 2020, Gorzelak *et al.*, 2015, Hsieh *et al.*, 2016), and transportation to the laboratory or processing site (Choo *et al.*, 2015, Dominianni *et al.*, 2014, Pollock *et al.*, 2018). For the sampling site, it could be shown in human fecal samples that within-sample difference accounts for minor shifts in taxonomical profiles but is negligible

compared to other technical and biological considerations such as primer choice, sequencing machinery, or reference database (Voigt *et al.*, 2015, Wu *et al.*, 2010a). To ensure the stability of microbial communities within one sample and prevent changes in the microbial composition, e.g., during the transportation of samples to the laboratory at ambient or only cold conditions, preservation buffers and stabilizers are widely used for microbiome research (Menke *et al.*, 2017). Commercial stabilizers such as the DNA stabilizer (STRATEC Biomedical AG, Birkenfeld, Germany), OMNIgene.GUT (DNA Genotek Inc., Ontario, Canada) or RNAlater (Thermo Fisher Scientific Inc., Waltham, USA) are high in cost and thus not always applicable. Some studies, e.g., Menke *et al.* (2017), Chongming *et al.* (2021), Han *et al.* (2018), or Chen *et al.* (2019), already showed that self-made stabilizer can show similar stabilizing effects compared to commercial products but are by far cheaper and relatively easy to produce. Nonetheless, it was reported that stabilizing agents could influence the resulting taxonomic composition of a sample (Gorzelak *et al.*, 2015, Lim *et al.*, 2020, Song *et al.*, 2016). Extensive freeze-thaw cycles were shown to influence the microbial composition as well (Cardona *et al.*, 2012, Fouhy *et al.*, 2015a, Gorzelak *et al.*, 2015). Thus, those should be kept to a minimum by aliquoting the samples before freezing. Freezing and storage temperatures were also previously evaluated as a biasing factor (Guo *et al.*, 2016, Tal *et al.*, 2017, Tedjo *et al.*, 2015). It could be shown that freezing, as soon and as cold as possible, performs best. Moreover, samples should be ideally extracted as soon as possible after arrival at the laboratory/testing facility (Gorzelak *et al.*, 2015, Nel Van Zyl *et al.*, 2020, Zhang *et al.*, 2017). Summed up, samples should be collected under clean and, if applicable sterile conditions. Stabilizing agents allow a sample transfer at ambient temperature. At the testing facility samples should be homogenized, aliquoted and stored at very cold temperatures (ideally -80°C or below).

Several aspects can impact the efficiency of the DNA extraction method. Thus, several studies investigated in detailed suitable and unsuitable protocols for DNA extraction. The investigated biases were, for example, uneven extraction efficiencies due to different lysis capabilities, the impact of inhibitors on the extraction efficiency, or the resulting DNA quality and quantity.

Generally, protocols or commercial DNA extraction kits perform either or a combination of mechanical, enzymatic, or chemical lysis. It was shown that mechanical lysis or a combination of enzymatic and mechanical lysis increased the resulting DNA yield and allowed for better detection of gram-positive bacteria, most likely due to a more powerful cell wall disruption compared to other lysis protocols (Costea *et al.*, 2017, Markusková *et al.*, 2021, Teng *et al.*, 2018, Videnska *et al.*, 2019). The remaining sample debris and organic compounds such as humic acid, bile salts, and polysaccharides are the main sources of inhibitors after DNA extraction (Pollock *et al.*, 2018). Thus, an efficient DNA extraction protocol should include steps or measures that guarantee the elimination of such compounds. If samples or environments of low biomass are targeted (e.g., milk samples, oral samples, water samples, etc.), special

precautions and testing must be performed beforehand to guarantee that samples do not get contaminated by either contaminant DNA (DNA originating from sampling or laboratory environments including DNA extraction systems) or by cross-contamination (Eisenhofer *et al.*, 2019).

The choice of appropriate primer and/ or V-region depends on several factors such as technical properties, e.g., read-length that can be sequenced using a defined methodology or machinery, the targeted environment, or if comparability with previous studies is desirable. Generally, no accordance with one sequencing strategy, primer pair, or V-region was found for short amplicon sequencing. The most frequently used V-regions depend on the environment targeted but are often V1-V2/V3, V3-V4/V5, or V4 (Abellan-Schneyder *et al.*, 2021a, Bharti and Grimm, 2019, Fouhy *et al.*, 2016). The choice of primer or V-region affects the resulting taxonomic profiles and taxonomic resolution (Bukin *et al.*, 2019, Pollock *et al.*, 2018). Moreover, different V-regions differ in their ability to identify some bacteria and different primer pairs have distinct affinities for DNA binding and thus, introduce biases during PCR (Kim *et al.*, 2011, Knight *et al.*, 2018). Other biasing sources during the amplicon generation step are the gDNA input amount, the number of cycles used for PCR, the PCR process, or the used DNA polymerase.

The use of different initial gDNA concentrations for amplicon generation can impact the resulting taxonomical profiles. Thus initial concentrations should always be stated (Multinu *et al.*, 2018). Sze & Schloss (2019) recommended using DNA polymerases with the highest possible fidelity and using as few PCR cycles as possible to reduce possible biases. Rausch *et al.* (2019) found that one- or two-step PCR procedures slightly differed from each other when analyzing the resulting taxonomical profiles. But significant and strong differences in taxonomic profiles were due to differences in selected V-region rather than amplification method.

After the amplicon generation, short 16S rRNA gene sequencing studies can be biased by using different sequencing platforms and the downstream analysis pipelines. D'Amore *et al.* (2016) benchmarked the use of different experimental strategies and sequencing platforms for 16S rRNA gene amplicon sequencing. They found that the taxonomical composition of samples was biased by the platform, sequenced region, and primer choice. Thus, they recommended to always perform a small trial using mock communities of known composition in the early stage of a sequencing project to assess possible biasing factors and to identify the most suitable protocol for the given research question. Today, bioinformatical analysis pipelines can be divided into two groups. The first group includes pipelines such as Mothur (Schloss *et al.*, 2009), Qiime (Caporaso *et al.*, 2010), or USEARCH-UPARSE (Edgar, 2013), which rely on operational taxonomic units (OTUs). The second group includes more recent pipelines, for example, DADA2 (Callahan *et al.*, 2016), Qiime2 (Bolyen *et al.*, 2019), or USEARCH-UNOISE3 (Edgar, 2018), which build-up on so-called amplicon sequence variants

(ASVs) or zero-radius OTUs (zOTUs). During data processing, sequences are clustered, either at a 97% (OTUs) or 99% (ASVs / zOTUs) sequence similarity threshold. The 97% identity approach carries the risk that multiple species, which show only very little sequence variation, are falsely grouped into one OTU. The newer pipelines correct for sequencing errors by different denoising approaches and require higher sequence identities, thus the risk of grouping different species into one OTU is reduced but the risk of falsely identifying new species due to sequencing errors increase (Almeida *et al.*, 2018, Callahan *et al.*, 2017, Nearing *et al.*, 2018). Prodan *et al.* (2020) compared the six above-mentioned bioinformatic pipelines for the analysis of amplicon data. They found that differences in sensitivity and specificity were large for the six tested pipelines. Thus, pipelines should be chosen wisely and after testing. The taxonomic assignment, which is performed after sequence clustering, is done using databases of known 16S rRNA gene sequences. Such databases are, for example, GreenGenes (DeSantis *et al.*, 2006), the Ribosomal Database Project (Cole *et al.*, 2014), EzBioCloud (Yoon *et al.*, 2017) or Silva (Quast *et al.*, 2013). Sierra *et al.* (2020) showed that the use of different reference databases could lead to variations and differences, especially at genus-level classification, in taxonomic compositions of given samples. Lately, some studies have described that taxonomic resolution could be improved by using environmental-specific databases (Dueholm *et al.*, 2020, Escapa *et al.*, 2020, Myer *et al.*, 2020). Meola *et al.* (2019) could, for example, show that by using their environmental-specific database, taxonomic accuracy could be greatly increased for the tested samples. Summed up, a variety of factors can influence the taxonomic classification of given samples, and thus, documentation and consistency are essential to minimize this effect when studies shall be compared.

1.4 DNA- vs. RNA-based studies

Even though most microbiota and microbiome studies rely on sequencing the 16S rRNA gene, direct 16S rRNA sequencing using the RNA as starting material is also possible. Directly sequenced rRNA samples will emphasize metabolic active bacteria of a sample, whereas DNA will only show the microbial rRNA gene composition of strains within a sample, whether active or not. The amount of rRNA correlates with growth rate, and the decrease of rRNA content is associated with a decreased growth rate (Blazewicz *et al.*, 2013). But several limitations of rRNA being used as an indicator for microbial activity in mixed microbial samples are also known. Some of those are, for example: that the relationship of activity and rRNA is not always constant, between different bacteria, the rRNA concentration and growth rate differ, or the relationship between non-growth activities, such as cell motility, coping oxidative stress, or conjugation, was not investigated sufficiently and thus, might falsify the relationship of rRNA

and activity (Blazewicz *et al.*, 2013). Standard DNA amplicons are biased by the number of ribosomal operons bearing a bacterial species. In general, copy numbers vary from one up to more than 15 copies per genome (Klappenbach *et al.*, 2001, Vetrovsky and Baldrian, 2013). Primer bias, due to unequal amplification or primer binding and amplification efficiencies, also plays a significant role in such analysis and is often referred to as problematic. The benefits and drawbacks of the DNA and RNA-based approaches are summarized in Figure 1.4.1.

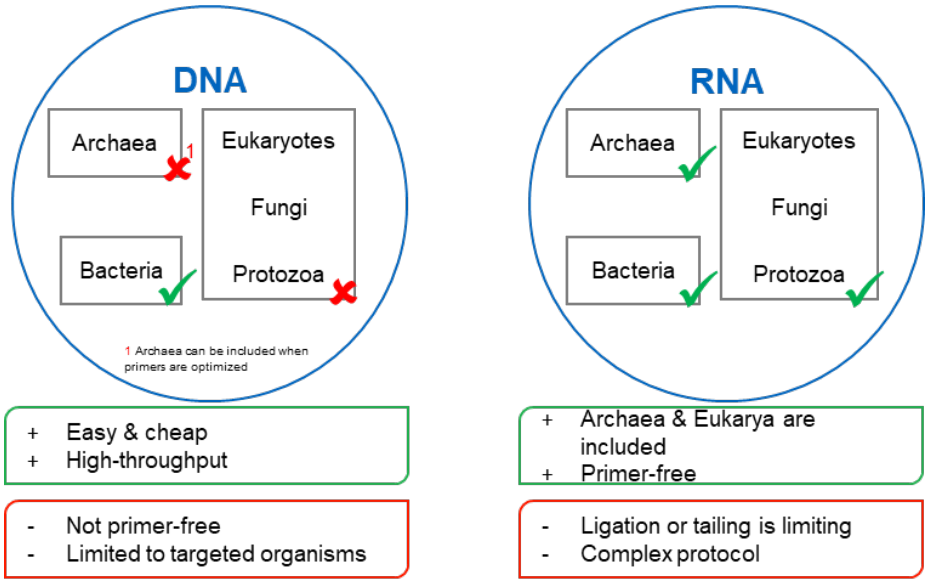

Figure 1.4.1: Benefits and drawbacks of RNA- and DNA-based sequencing approaches. Primer-free RNA approaches allow us to study not only bacteria but also archaea and eukaryotic microorganisms. But RNA based methods are more difficult, time-consuming, and often not tested for high throughput. DNA-based methods are widely accessible, easy, and tested for high throughput at low costs. Nevertheless, those methods are limited due to the introduced primer bias, which defines the targeted microorganism (e.g., standard 16S rRNA gene primer targets only bacteria).

Gremion *et al.* (2003) described that the combination of both DNA and RNA-based amplicon sequencing could be used to not only describe the microorganisms present in an environment but also to find those being metabolic active. Moreover, they found that frequently and in the high amount identified bacterial genera of the DNA-based approach were not found in the RNA-based analysis. Therefore, it can be assumed that those might not represent active constituents of the analyzed samples. Besides those mentioned above, only a few methods using direct 16S rRNA-derived amplicons for sequencing are published. Rosselli *et al.* (2016) described a primer-free sequencing approach, where the rRNA is enzymatically fragmented and subsequently sequenced. This method allows the detection of different taxonomical abundances and assesses physiological active genera. The authors suggest using this method as a complementary and not alternative approach to standard 16S rRNA gene sequencing, as due to its different approach, the results are not directly comparable. Karst *et al.* (2018) published a primer-independent full-length sequencing approach, which allows studying not only bacteria and archaea (16S) but also eukaryotic microorganisms (18S). This approach relies on the extraction of SSU rRNA, polyadenylation, and subsequent reverse transcription

using a poly-T primer. Thus, a primer bias is not given here, but the method is by far more tedious and not tested for high throughput. Interestingly, the authors could show that with their primer-free approach they were able to find a large proportion of novel microbial diversity.
Nevertheless, it must be noted here that the RNA-based approach also poses several drawbacks. The most prominent one is that RNA is less stable than DNA and is prone to degradation (Gołębiewski and Tretyn, 2020). Moreover, RNA is differentially expressed under changing environmental conditions and is therefore unsuitable for quantitative estimation of microbial diversity within an environment. Reverse transcription and difficulties in RNA extraction are further biasing steps. Still, they might be minor when direct rRNA analysis allows for novel findings in relationships between microbial diversity, ecosystem functions, and interactions (Liu *et al.*, 2019a).

1.5 Full-length SSU rRNA sequencing approaches

1.5.1 Third generation sequencing approaches

Until today, most microbiota or microbiome-related studies use short amplicon 16S rRNA gene strategies and thus second-generation sequencers such as the Illumina MiSeq. As reported above, one of the main drawbacks of this approach is the comparably short read-length of a maximum 2×300 bp. Thus, only parts and not the full-length 16S rRNA gene can be sequenced. Newest methods, using a so-called third-generation sequencer, overcome this read-length limitation.
The main differences between second- and third-generation sequencing approaches are differences in speed and read length. Third-generation sequencers allow sequencing of up to several thousand base pairs in real-time (Schadt *et al.*, 2010, van Dijk *et al.*, 2018). Today, the most prominent third-generation sequencers are Oxford Nanopore Technologies (ONT) MinION and the Pacific Biosciences (PacBio) Sequel Series (Ameur *et al.*, 2019, Branton *et al.*, 2008, Eid *et al.*, 2009). PacBio's sequencer relay on single-molecule real-time (SMRT) sequencing strategies. Therefore, so-called SMRTbells, which are circular templates, are created by ligating adapters to either extracted DNA or RNA. The SMRTbells are then loaded on a so-called SMRT cell. There, the SMRTbells diffuse into tens of thousands of single sequencing units, which are called zero-mode waveguides (ZMWs). In each of those ZMWs, a single DNA polymerase forms a complex with the template and gets immobilized on the bottom of the ZMW. Labeled nucleotides are added to the SMRT cells, and if a labeled dNTP is incorporated by the polymerase, a light pulse is produced, which is different for each base. Moreover, by tracking the speed of the polymerase, kinetic data can be generated, which enables the detection of modified bases (Eid *et al.*, 2009, Heather and Chain, 2016, Rhoads

and Au, 2015). The other, nowadays frequently used third-generation sequencing technology is the one offered by ONT. There, a single strand of either RNA or DNA is translocated through a nanopore. The pocket-size sequencing device, the so-called MinION, has at least 800 nanopores and allows detecting a change in the ionic current when DNA or RNA sequences pass through the pore, which is facilitated by an ATP-consuming adapter protein (Jain *et al.*, 2016). Due to the different steric demands of the different bases, the different changes in current can be used to decode the sequence. This technology does not need fluorescent-labeled dNTPs because the bases are identified by their three-dimensional properties, and therefore, it is also possible to determine base modifications (Laver *et al.*, 2015, Xu and Seki, 2020).

Several studies showed that long-read sequencing strategies could improve the taxonomic resolution of 16S rRNA gene sequencing, as the whole 16S rRNA gene and not only parts, could be sequenced (Benítez-Páez *et al.*, 2016, Callahan *et al.*, 2019, Cuscó *et al.*, 2019, Earl *et al.*, 2018, Matsuo *et al.*, 2021). Nevertheless, some studies suffer from still high overall sequencing error rates, especially if sequenced with ONTs MinION or with high costs and overall lower throughput. Further, surprisingly, the bioinformatic base calling is time consuming and should not be underestimated in total processing time (own observation). Regardless, Johnson *et al.* (2019) could show that by sequencing the full-length 16S rRNA gene, the taxonomic resolution was improved, and the ability to distinguish between different bacterial species and even strains was given.

1.5.2 Synthetic full-length approaches

Besides the above-mentioned real full-length sequencing approaches, relying on long-read sequencing machineries, alternative approaches exist. So-called synthetic long-reads are based on sequencing strategies of specific barcoding and fragmentation, which allow a later *de novo* assembly of short reads to their initial full-length (Goodwin *et al.*, 2016). Thus, sequencing can be performed on the more accurate second-generation sequencing types of machinery but allow for better taxonomic resolution due to the later assembled longer reads. This technology is benefitted by the higher accuracy and availability of second-generation sequencers but needs higher coverage and is thus by far more expensive (McCoy *et al.*, 2014). In the context of microbiota or microbiome research, only a limited number of synthetic-long read strategies were published or used. Burke & Darling (2016) described a method for full-length 16S rRNA gene sequencing using a synthetic approach. They amplified near full-length 16S rRNA genes with the commonly used 27F and 1391R primers. Then they fragmented the produced amplicons using a transposase. By incorporating a unique 10N-tag into their initial primer sequence, they were able to *de novo* assemble the sequences back to the full length after sequencing the smaller fragments on an Illumina machinery. Thus, Burke & Darling

(2016) provided the first proof of principle for generating near full-length 16S rRNA gene sequences in a cost-effective and potentially high-throughput manner using a precise second-generation sequencing device. Based on this method, Karst *et al.* (2018) described an improved strategy, which also allows for primer-free full-length 16S rRNA sequencing. Therefore, RNA is used as starting material, polyadenylated, reverse transcribed using a poly-T primer and then tagged by using individual 15N-tags at the 5'- and 3'-site. Full-length molecules, which are too long for direct sequencing, are fragmented by tagmentation to enable sequencing on an Illumina machinery and *in-silico de novo* assembly is enabled by extracting corresponding 15N-tags for forward and reverse reads (Karst *et al.*, 2018). Deutscher *et al.* (2018) used the methodology of Burke & Darling (2016) and could show that wild flies from different locations were accompanied by different *Asaia* strains. Such an analysis would not have been possible using only short amplicon reads. Nevertheless, Deutscher *et al.* (2018) stated that the full-length method does not only bear advantages. Insufficient sequencing depth is a major problem. This could, on the one hand, lead to an underrepresentation of sequences with the same molecular tag, which can therefore not be *de novo* assembled back to the full length. On the other hand, molecular tagging could be imprecise, leading to the fact that the same molecular tag is not observed more than once. Nevertheless, this can be compensated by only accepting tags as real tags when they are detected more than once (Deutscher *et al., 2018)*.

Lately, a new synthetic full-length approach based on a commercially available kit-based system was described. Loop Genomic, a 2015 founded company, provides novel technologies for long-read DNA sequencing. They offer several microbiome kits, including a 16S and a 16S & 18S long-read kit. Those build-up on the reconstruction of full-length molecules after sequencing small fragments on a short-read sequencer. Therefore, a unique molecular identifier (UMI) is attached to each initial sequence. The UMI is afterward distributed intramolecularly and allows thereby a later reconstruction to the full-length as all sequences originating from the same initial strand can be identified. A detailed protocol of this technique and the general workflow is unpublished. Nonetheless, it is known that the technique builds up on previously described methods by Stapleton *et al.* (2016) and Hong *et al.* (2014), which include enzymatic steps as well as self-circularization of the constructed amplicons (Callahan *et al.*, 2021, Chung *et al.*, 2020). Jeong *et al.* (2021) could show that the full-length 16S approach, using the LoopSeq kit, improves the taxonomical resolution of different bacteria compared to short-amplicon sequencing.

1.6 Non-SSU rRNA gene-based sequencing strategies

Besides the amplicon-based sequencing strategies, metagenomic sequencing is used to not only study the communities' microbiota but to gain both taxonomic and functional analysis of the analyzed sample (Fricker *et al.*, 2019). Metagenomics is defined as the analysis of the sum of all genetic material present in a distinct environment (Thomas *et al.*, 2012). Besides metagenomic analysis, also RNA-based sequencing approaches can be used to assess specific microbiome-related research questions. Metatranscriptomic analyses are RNA sequencing analyses, which allow insights into the expressed transcripts of a given sample. Thus, allowing to study and to compare gene expression under different environmental conditions and allowing insight on pathway activities (Hatch *et al.*, 2019, Shakya *et al.*, 2019). So, RNAseq allows to gain insight into genes that are actively expressed within a microbiome (Bashiardes *et al.*, 2016). Fremin *et al.* (2020) showed that an adaption of the common ribosome profiling protocol could be used to investigate thousands of translated mRNAs of mixed cultures simultaneously in microbiome samples. However, finding the source genome for each footprint is impossible.

1.6.1 Shotgun metagenomic sequencing

The term shotgun sequencing refers to the preparation method for DNA, which gets later sequenced. The first part of a shotgun metagenomic sequencing protocol is to extract the DNA of the community. Then, the DNA is fragmented into smaller parts by using either nebulization, sonication, or adaptive focused acoustics technology such as the Covaris technology. Afterward, those smaller fragments are prepared for sequencing, including steps for purification, ligation, or amplification of sequencing primers, normalization, and pooling of the samples (Madigan *et al.*, 2014, pp. 210-211, Poptsova *et al.*, 2014). After sequencing, the resulting reads can either be used to assemble microbial genomes or can be used for assembly-free approaches such as taxonomic profiling of functional analysis (Pérez-Cobas *et al.*, 2020). The advantage of shotgun metagenomic sequencing compared to 16S rRNA gene sequencing in assigning taxonomy is resolution. While 16S rRNA gene sequencing can only resolve species-level classification in the case of full-length sequencing or genus level classification when short amplicon sequencing is applied, metagenomic sequencing allows species and even strain level classification. However, this does not rely solely on 16S rRNA genes but also on other genes present and ascribed to a species (Pérez-Cobas *et al.*, 2020, Scholz *et al.*, 2016). Moreover, shotgun metagenomic sequencing is not only limited to bacteria and archaea due to the presence of the 16S rRNA gene but allows for detecting besides bacteria and archaea, also viruses, fungi, and perhaps parasites from a given sample (Dulanto Chiang and Dekker, 2019). Nevertheless, those advantages come with higher sequencing

costs, complex and time-consuming bioinformatical needs, and the fact that host-derived contaminations are present (Liu *et al.*, 2021). Therefore, shotgun sequencing should only be applied when besides the taxonomical assignment, functional analysis or the resolution of bacterial genomes is favored. Nevertheless, the challenge will be to develop and perform bioinformatic analysis that allows easy, precise, and time-efficient analysis of the targeted samples. Besides, researcher must be trained intensively to analyze the complex data sets and tools, databases and programs used. The software or tools for metagenomic shotgun sequencing need to be understandable and useable as they have a major impact on the resulting outcome (Couto *et al.*, 2018, Quince *et al.*, 2017).

1.6.2 Shallow shotgun metagenomic sequencing

As described before, shotgun metagenomic sequencing provides good functional and taxonomical resolution of mixed communities within a sample. In contrast, 16S rRNA gene sequencing only allows taxonomic resolution to genus or sometimes species level but is more cost-effective and easier to analyze. Nevertheless, there is a need for a high-resolution functional and taxonomic analysis that would be more affordable when compared to shotgun metagenomic sequencing (Hillmann *et al.*, 2018). This is the case for shallow shotgun metagenomic sequencing (SSMS), where samples are not sequenced as deep as for shotgun metagenomic sequencing, and therefore, sequencing costs are reduced. Thus, SSMS is more cost-effective while allowing for species-level classification and functional analysis of the targeted sample. Nevertheless, caution must be taken as Santiago-Rodriguez *et al.* (2020) could show that sequencing depth (coverage) impacts the subsequent taxonomical analysis and sequencing depth is dependent on the research question due to different needs, e.g., bacterial species-level classification did only need 0.5-0.75 Gb data, while for virus profiles sequencing depths of up to 5 Gb were necessary.

1.6.3 Metatranscriptomic and metatranslatomic analysis

Metatranscriptomic analysis uses RNA to investigate changes in the transcriptome, i.e., the mRNA content of a given sample, which can be used to investigate how microbial communities react to stress, time, or environmental changes (Shakya *et al.*, 2019). For metatranscriptomic analysis, mRNA needs to be extracted and rRNA depleted. Then, primers are ligated to the mRNA, and the mRNA is reversed transcribed, and prepared for sequencing, including library purification, normalization, and pooling. Schirmer *et al.* (2018) showed that metatranscriptomic paired with metagenomics could allow for the profiling of disease-related changes in microbiomes and provided new information on microbiome interactions. Li *et al.* (2019) could show that their metatranscriptomic analysis of the rumen microbiome allowed to better investigate rumen microorganisms and their association with host performances when

compared to metagenomics. Summed-up, the main benefits of metatranscriptomic in microbiome studies are: (1) gaining information on which genes are transcribed (up- or down-regulated) at a certain time-point and under defined condition, which leads to (2) enabling to gain information of potential functions of the microbiome and (3) those functional information can be used to investigate and identify metabolic pathways (Bashiardes *et al.*, 2016, Franzosa *et al.*, 2014).

Besides the mRNA-based metatranscriptomic approaches, metratranslatomic analysis focuses on sequencing only mRNA, which is used as a template for translation. Therefore, only the mRNAs that are covered by the ribosome during translation are sequenced. Only very few approaches for metatranslatome analysis have been described to this day, despite ribosome profiling being an often-used approach for single-culture analysis. Fremin *et al.* (2020) described the only published study where the translation of genes in complex microbiomes was studied on a large-scale using ribosome profiling. They could successfully show that using their adapted protocol, ribosome profiling data for uncultured fecal microbiota samples could be collected and that those results allowed to study protein synthesis across the taxa present in the stool samples. Giehren (2021, unpublished) showed that adapting the classical ribosome profiling protocol (e.g., as described by Hücker *et al.* (2017), and Neuhaus *et al.* (2017)) for mixed-bacterial cultures is possible using an in-house analysis pipeline. Moreover, they could improve the ribosome profiling protocol to make it accessible for low nucleic acid concentrations extracted from bacteria or mixed-community samples (e.g., mice gut samples).

In summary, multi-omics data, including metatranscriptome and metatranslatome analysis, allow researchers to gain deeper knowledge and understanding on functional gene expression and regulation in microbiomes at diverse environmental conditions.

1.7 Objective of the study

In the last twenty years, 16S rRNA gene sequencing substantially increased not only popularity but also in feasibility. As preparation became easily operable and sequencing more accessible and cost-efficient, 16S rRNA sequencing became a standard technique to study the microbiota of a given environment or context. Nevertheless, no general accordance on the exact protocols exists, e.g., which primer, V-regions, or DNA extraction method to use and thus, amplicon preparation methods, sequencing approaches, and the bioinformatic analysis differ from study to study.

Therefore, the overall aim of this study is to test, compare and find the most suitable 16S rRNA sequencing approach for different given demands. The following questions and tasks were answered and executed throughout the doctoral project.

1. How do samples have to be collected, stored, and treated to guarantee sample integrity? What are the most suitable DNA and RNA extraction protocols, and what is the difference in targeting RNA vs. DNA for 16S rRNA sequencing? To answer those questions, the following section 2.1 describes the best protocols for storage, extraction, and primer use for DNA and RNA-based sequencing approaches, as well as a comparison of the latter two approaches based on short amplicon 16S rRNA data.
2. How should short amplicon 16S rRNA gene sequencing be performed targeting human fecal samples? What are critical parameters that must be addressed, what are the most suitable protocol details, and which bioinformatical tools and reference databases are performing best? Section 2.2 gives recommendations on which primer, pipelines, and parameters perform best for fecal sample 16S rRNA gene sequencing.
3. How can short amplicon sequencing be adapted for samples of low biomass? To target this question, the implementation of a ddPCR step after amplicon generation was assessed. It was checked whether the implementation of this step allowed to decrease the initial gDNA input amount needed and whether the step impacted the reliability of the sequencing results.
4. How can we measure the microbial load of a sample? Considering that, a comparison of sequencing-based approaches (16S rRNA gene sequencing using spike-in controls) and cell-counting (flow cytometry) was performed. Moreover, benefits and drawbacks of each method were analyzed and are discussed.
5. How feasible and efficient are different full-length 16S rRNA sequencing approaches? What are the main benefits and drawbacks of synthetic full-length vs. long-read sequencing approaches? Different full-length sequencing approaches are analyzed, evaluated, and compared to short-amplicon sequencing results in section 2.5.
6. How can full-length 16S rRNA gene sequencing be used for detecting putative pathogenic bacteria? What is the main improvement and benefit of full-length sequencing compared to short amplicon sequencing? Using bovine milk samples, synthetic full-length 16S rRNA gene sequencing was performed and compared to short amplicon sequencing in section 2.6. Besides the comparison of those two techniques, it was investigated if full-length sequencing would improve the detection of potentially pathogenic bacteria, which are associated with mastitis in cows.

2. Results

The results of sections 2.2 and 2.6 of this dissertation were published previously in peer-reviewed journals, and the original publications are added. The results of section 2.3 are currently under review in a scientific journal. Sections 2.1, 2.4 and 2.5 are presented as drafted manuscripts which are intended to be sent for publication. The personal contributions are as follows:

Section 2.1:

Abellan-Schneyder I, Rothfischer F, Singer MT, Neuhaus K: Setting standards for 16S rRNA sequencing – storage, extraction, and handling.

Personal contribution: The study was designed by I. Abellan-Schneyder under supervision of K. Neuhaus. Experiments were performed by I. Abellan-Schneyder with the help of F. Rothfischer and M. T. Singer. Data analysis was performed by I. Abellan-Schneyder. The draft of the manuscript was written by I. Abellan-Schneyder and edited by K. Neuhaus.

Section 2.2:

Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K (2021). Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. mSphere 6:e01202-20.

Personal contribution: The study was designed by I. Abellan-Schneyder. I. Abellan-Schneyder performed wet-lab experiments and sequencing with the help of A. Sommer and Z. Sewald. M. S. Matchado and S. Reitmeier conducted the bioinformatic analysis. J. Baumbach, M. List and K. Neuhaus supervised the study and provided funding. I. Abellan-Schneyder wrote the first draft of the manuscript, which was edited by all other co-authors.

Section 2.3:

Abellan-Schneyder I, Schusser AJ, Neuhaus K (2021) How low can we go? Implementation of ddPCR allows amplicon sequencing of ultra-low amounts of gDNA from low biomass samples. Under review.

Personal contributions: I. Abellan-Schneyder conducted the main experiments and supervised A. J. Schusser, who helped to develop the protocols. Further, I. Abellan-Schneyder analyzed the data, wrote the first draft, and prepared figures. K. Neuhaus provided resources and was involved in supervision, conceptualization, and conducted final review & editing.

Section 2.4:

Abellan-Schneyder I, Ahmed M, Reitmeier S, Metwaly A, Haller D, Neuhaus K: Cell-counting in human fecal samples comparing flow cytometry versus spike-in standards in 16S rRNA gene sequencing. Personal contributions: I. Abellan-Schneyder, M. Ahmed, M. Metwaly, D. Haller and K. Neuhaus designed the study. I. Abellan-Schneyder and M. Ahmed performed the experiments. S. Reitmeier helped with the scripts for the bioinformatical spike-in analysis. A. Metwaly provided the human fecal samples. A. Metwaly, D. Haller and K. Neuhaus supervised the study. I. Abellan-Schneyder wrote the first draft of the manuscript, which was edited by K. Neuhaus.

Section 2.5:

Abellan-Schneyder I, Sewald Z, Schusser AJ, Grimm M, Neuhaus K: Synthetic versus long-read 16S rRNA gene sequencing approaches – benefits, drawbacks, and feasibility. Personal contributions: I. Abellan-Schneyder designed the study under supervision of K. Neuhaus. I. Abellan-Schneyder performed the experiments with the help of A. J. Schusser (synthetic full-length sequencing) and M. Grimm (long-read sequencing), which both were supervised by I. Abellan-Schneyder. Z. Sewald coded the synthetic full-length sequencing downstream analysis pipeline. I. Abellan-Schneyder wrote the first draft of the manuscript, which was edited by K. Neuhaus.

Section 2.6:

Abellan-Schneyder I, Siebert A, Hofmann K, Wenning M, Neuhaus K. Full-length SSU rRNA gene sequencing allows species-level detection of bacteria, archaea, and yeasts present in milk. Microorganisms 2021, 9(6):1251.

Personal contributions: I. Abellan-Schneyder and K. Neuhaus conceptualized the study. I. Abellan-Schneyder performed all experiments except for the sample collection and gDNA extraction, which was performed by A. Siebert and K. Hofmann. I. Abellan-Schneyder and K. Neuhaus analyzed, validated, and curated the data. Resources, funding, and project administration were provided M. Wenning and K. Neuhaus. The original draft was prepared and writing by I. Abellan-Schneyder. The draft was reviewed and edited by all other authors.

2.1 Setting standards for 16S rRNA gene sequencing – storage, extraction, handling

Drafted manuscript

2.1.1 Abstract

Short amplicon 16S rRNA gene sequencing is the gold standard for evaluating the microbiota of a given sample such as human gut samples, soil, or milk. After samples are taken, they must be stored and transported in an efficient and safe way to the research laboratory. Stabilizing buffers, allowing the sample to be transported at ambient temperature, facilitate this process. Nevertheless, different commercially available stabilizing agents and self-made buffers show varying degrees of stabilizing potential, potentially influencing the outcome of the analysis. Thus, the performance of three stabilizing agents on stabilizing DNA and RNA was evaluated and compared. We further investigated if different nucleic acid extraction methods added bias to our analysis. Our results demonstrate that sample collection, extraction of nucleic acids, and the later amplicon preparation should be planned carefully under consideration of the study's needs.

2.1.2 Introduction

The study of the human microbiome is challenging because it has to be performed by culture-independent analyses (Robinson *et al.*, 2010). Hence, a variety of different molecular techniques were developed, such as qPCR, fluorescence in situ hybridization, rRNA, and whole community analysis, e.g., via whole-genome sequencing, conventional metagenomics, and proteomics (Theron and Cloete, 2000).

Not only analysis methods differ among microbiome studies. Starting with the sample collection, different methods and protocols change widely in their applications and use of chemicals. To ensure the stability of microbial communities within one sample and prevent changes in the microbial composition, preservation buffers and stabilizers are widely used for microbiome research (Menke *et al.*, 2017). Commercial stabilizers such as DNA stabilizer (STRATEC Biomedical AG) or RNAlater (Thermo Fisher Scientific Inc.) are high in cost and thus not always applicable. Several studies already showed that self-made stabilizers have similar stabilizing effects compared to commercial products but are by far cheaper and relatively easy to produce. Ideally, a stabilizing solution should show simultaneous effects to stabilize DNA and RNA, and facilitate homogenization of the sample. Until now, most studies examined only stabilizing effects either DNA or RNA, but to our knowledge, no study investigated if both types of nucleic acids could be preserved within a sample at the same time

(Camacho-Sanchez *et al.*, 2013, Menke *et al.*, 2017). Thus, testing different buffers that are suitable to stabilize both DNA and RNA were considered.

The use of the SSU rRNA gene for identification (16S in bacteria/archaea and 18S in eukaryotes) was established in 1977 and revolutionized the ability to identify microbes (Pace *et al.*, 2012). Based on their ubiquitous presence, evolutionary stability, and the distinguishable changes in its variable sequences, SSU rRNA genes are an ideal phylogenetic marker to identify and distinguish between different taxa (Janda and Abbott, 2007). One of the easiest and convenient methods to identify microbial populations is sequencing of the 16S ribosomal RNA (rRNA) encoding genes, followed by a comparison to known bacterial sequence databases (Eckburg *et al.*, 2005). To produce amplicons, in general, 10-50 ng of high-quality DNA are needed. For RNA-based studies, the input amount varies but is mostly described to be between 10 ng to 10 µg per sample (Li *et al.*, 2017, Rosselli *et al.*, 2016). Thus, besides the storage of the initial sample and the extraction of nucleic acid, also amplicon generation plays an important role and is a step that can be biased due to different procedures.

Here, we want to investigate the best-practice procedure for the collection of stool samples, the extraction of both DNA and RNA, and the following 16S rRNA amplicon preparation using those nucleic acids.

2.1.3 Material and Methods

Composition of stabilizing buffers

Three different stool stabilizing buffers were tested on their efficiency to stabilize DNA and RNA in stool samples under different environmental conditions. Commercially available DNA stabilizers (STRATEC Biomedical AG) and RNAlater (Thermo Fisher Scientific Inc.) were compared to a self-made stool stabilizing solution (SStab). SStab is composed of 1,400 ml MilliQ water which was stirred while adding 60 ml 0.5 M EDTA, 37.5 ml 1 M sodium citrate (258 g), and 7.95 M Ammonium sulfate (1050 g) in 100 g amounts. The solution was cooled, pH was set to 5.2 using sulfuric acid at room temperature. Then the solution was filtered through a 0.2 µm filter cup.

Preparation of human gut sample

Stool samples were obtained from healthy volunteers after informed consent and collected in stool sample tubes (Sarstedt AG & Co.). Tubes were previously filled with 4 ml stabilizing buffer and collected samples were directly resuspended by shaking and vortexing. Samples were aliquoted and stored at -80°C.

RNA isolation

After homogenization of the thawed sample, fecal content was centrifuged (700×*g*, 4°C, 5 min) to remove the remaining solid fecal matter. One milliliter of the supernatant was transferred to a fresh microfuge tube and filled up to 2 ml with ice-cold 1× PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, and 1.8 mM KH_2PO_4). Samples were then centrifuged (9,000×*g*, 4°C, 5 min), and the supernatant was discarded. The pellet was resuspended in 0.5 ml ice-cold 1× TE buffer (composed of 10mM Tris-HCl containing 1mM EDTA x Na_2, pH 8). RNA isolation was either performed as previously described by Zoetendal *et al.* (2006) with minor changes or using the Zymo Quick-RNA Fecal/Soil Microbe Microprep Kit (Zymo Research). If the protocol by Zoetendal *et al.* (2006) was used, the following modifications were conducted. DNase digestion was not performed using the RNeasy mini kit but using TURBO DNase (Thermo Fisher Scientific Inc.) as described by the manufacturer's protocol. Afterwards, an ethanol precipitation was performed using 1/10 volume 3 M NaOAc (pH 5.2), 1/100 volume of glycogen and 3 volumes of 100% EtOH. Precipitation was performed overnight at -20°C. After centrifugation (13,400×*g*, 4°C, 20 min) and two wash steps with 80% EtOH, pellet was briefly dried and resuspended in 50 µl 1× TE buffer. If the Nanodrop measurement indicated impurity of the samples, they were cleaned up using the Zymo purification and concentrator columns (Zymo Research).

DNA isolation

DNA isolation from feces was performed using a shortened version of the method described by Godon *et al.* (1997). In brief, 2 ml aliquots were thawed on ice and vortexed several times. Then, 600 µl of the sample was transferred into a bead-beating tube and 250 µl of 4 M guanidinium thiocyanate, and 500 µl 5% N-lauroylsarcosine sodium salt were added. The samples were incubated for 60 min at 70°C while shaking at 700 rpm. Bead-beating was applied using the MP Biomedicals FastPrep24 (MP Biomedicals Inc.) thrice for 40 s at 6.5 m/s. Between the runs, samples were cooled with dry ice. Next, 15 mg poly(vinylpolypyrrolidone) were added and briefly vortexed. Samples were centrifuged for 3 min at 15,000×*g* at 4°C before transferring the supernatant into a fresh 2 ml sample tube. The supernatant was centrifuged for 3 min at 15,000×*g* and 4°C. Subsequently, 500 µl of clear supernatant were transferred into fresh 2 ml tubes and 5 µl RNase (10 mg/ml) was added. The samples were incubated for 20 min at 700 rpm and 37°C. Finally, after RNA digestion, DNA was cleaned up using NucleoSpin gDNA Clean-up filters following the manual (Macherey-Nagel). After DNA extraction, DNA concentrations were recorded using Nanodrop (NanoDrop Technologies, Inc.). By performing an agarose gel electrophoresis, it was checked whether DNA degradation was observable.

16S rRNA gene sequencing

For amplification of the variable region and addition of adapters, 1st-step PCR was performed in 50 µl volumes containing 24 ng of gDNA mixed with 1x Phusion HF buffer, 0.2 mM dNTPs, 0.125 µM of each 341F-ovh (5' TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG-CCT ACG GGN GGC WGC AG 3') and 785R-ovh Primer (5' GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G-GAC TAC HVG GGT ATC TAA TCC 3'), 0.5625% (v/v) DMSO and 0.5 U of Phusion HF II DNA polymerase. Primers were designed to have a 16S gene complementary site (marked green) as well as an adapter site to fuse the barcodes (overhang, marked blue). In a thermocycler, denaturation took place at 98°C for 40 s followed by 15 cycles of denaturation (98°C for 20 s), annealing (55°C for 40 s) and elongation (72°C for 40 s), and a final elongation step at 72°C for 2 min. The 2nd step PCR, which is needed for multiplexing, was performed in 100 µl volumes containing 10 µl of the 1st-step PCR product, 1x Phusion HF buffer, 0.2 mM dNTPs, 0.125 µM of each forward and reverse barcode used (e.g., SC501-Fw 5' AAT GAT ACG GCG ACC ACC GAG ATC TAC AC-ACG ACG TG-TCG TCG GCA GCG TC 3' and SA701-Rv: 5' CAA GCA GAA GAC GGC ATA CGA GAT-AAC TCT CG-GTC TCG TGG GCT CGG 3'), 0.25% (v/v) DMSO and 1 U of Phusion HF II DNA polymerase. 2nd step primers contained a 1st-step complementary region (part of the overhang, marked in blue), barcode sequence (marked in red), and the Illumina P5/P7 adapter sequence (marked yellow). After denaturation at 98°C for 40 s, 10 cycles of 98°C for 20 s, 55°C for 40 s and 72°C for 40 s, a final elongation step at 72°C for 2 min followed. For validation of the successful library production, 8 µl of 2nd step product were loaded onto a 1.5% (w/v) agarose gel. Electrophoresis was performed with 110 V for 40 min, and the gel was checked visually for the appearance of bands at desired size (about 570 bp) in comparison to a molecular size standard (ladder). The remaining 92 µl of 2nd step PCR product were purified with 1.8x AMPure XP beads following the manufacturers' protocol. DNA was eluted in 25 µl nuclease-free water. After measuring the concentrations of the purified samples with the help of the Qubit™ dsDNA HS Assay, samples were adjusted to 0.5 nM and pooled. Libraries for sequencing were prepared as described by the manufacturer. Pools were denatured, diluted, and mixed with denatured PhiX control solution (Illumina Inc.). Finally, 12 pM of the denatured sample libraries and 15% PhiX library were loaded onto the cartridge, and the run was started.

Small ribosomal subunit (SSU) rRNA isolation

To separate the SSU rRNA from total RNA, six different SSU rRNA extraction methods were evaluated.

1. Agarose gel extraction

First, 10 µg denatured total RNA was loaded onto a 1.5% (w/v) agarose gel (pre-stained using GelRed). For a size reference, the 1 kB DNA ladder (Thermo Fisher Scientific Inc.) was loaded. The gel was run for 40 min at 110 V before illumination on a blue LED light source was performed. Then, the SSU rRNA was cut out from the full gel. The agarose gel slab was dissolved and cleaned up with a Zymo Gel Recovery Kit (Zymo Research).

2. Agarose gel squeeze extraction

Following the same setup of the first separation protocol for the extraction of SSU rRNA from total RNA, 10 µg sample were loaded on an agarose gel. Then the gel part, including the SSU rRNA, was cut out and instead of recovering the RNA through a kit and column-based clean-up, the agarose gel piece, including the SSU rRNA, was simply squeezed in between two parafilm covered glass slides, and the exited liquid was recovered.

3. MOPS gel extraction

3-(N-morpholino)propanesulfonic acid (MOPS) gel electrophoresis is a common method used for RNA separation. 1% MOPS agarose gels were prepared using 10 ml 5x MOPS buffer (10 mM $MgSO_4$, 0.5 M MOPS, 2.5 M NaCl), 36 ml nuclease-free water, and 0.5 g agarose. The mixture was boiled in a microwave, cooled to about 55°C, and then 8.8 ml 37% formaldehyde and 1 µl GelRed were added in a fume hood. After the gel was cooled down, 1 µg denatured total RNA was mixed up with 2x RNA loading dye (Thermo Fisher Scientific Inc.) and loaded on the gel. The gel was run at 110 V for 40 min. Bands containing the SSU rRNA were cut out and gel extraction was performed using the Zymo Gel Recovery Kit (Zymo Research).

4. PAGE gel breaker extraction

Denaturing PAGE gels were produced using 6 M urea in a 3.75% PAGE set-up. The sample pockets of the PAGE gel were washed with buffer, and the gel was pre-run for 25 min at 140 V. Sample wells were washed again twice before the samples were loaded to prevent urea accumulation on the bottom of the sample wells. The gel was run for 70 min at 140 V and post-stained in SYBR Gold nucleic acid gel stain (Thermo Fisher Scientific Inc.). Under illumination, the bands were cut out of the gel. For size reference, several markers with different ranges were used.

5. E-Gel electrophoresis

The E-Gel precast agarose electrophoresis system (Thermo Fisher Scientific Inc.) was used in the double-comb format. The denatured total RNA sample was loaded in the top row comb

and gel was run until the SSU rRNA reached the bottom row. Then, the SSU rRNA was extracted by pipetting the liquid from the lower pocket.

6. Well-gel electrophoresis

Based on the E-Gel method, a similar method revolving around a normal agarose gel electrophoresis was performed. For this, a normal 50 ml 1.5% (w/v) agarose gel was set up and pre-stained using the GelRed pre-stain. Instead of only one comb for loading, two combs (an upper and lower comb) were inserted into the gel. Sample loading of up to 10 µg of extracted total RNA was performed. The gel tray was not fully covered with running buffer but only filled up to half of the gel thickness to enable a later recovery from the lower pocket without losing material to the surrounding buffer. Electrophoresis was performed until the SSU rRNA reached the lower pocket. The sample was pipetted out and used for further analysis.

RNA/DNA quality control

To ensure that the RNA products were of sufficient quality and quantity, RNA samples were controlled on a 2100 Bioanalyzer system (Agilent Technologies). Depending on the RNA concentration, either the RNA Nano 6000 or Pico 6000 reagent kit (Agilent Technologies) was used.

Direct 16S rRNA sequencing

For reverse transcription, 120 ng RNA was used. Therefore, the RNA was mixed with 1 µl 10 mM dNTPs (Sigma), 0.13 µM 785R-primer, and filled up to 15 µl with nuclease-free water. Samples were incubated for 5 min at 65°C and then cooled down for at least 1 min on ice. Reverse transcription was performed by adding 4 µl SSIV buffer, 1 µl 100 mM DTT, 1 µl SUPERase•IN RNase Inhibitor, and 1 µl SuperScript IV reverse transcriptase (Thermo Fisher Scientific Inc.). The reaction was performed at 55°C for 10 min followed by incubation at 80°C for 10 min. cDNA was amplified by PCR. The 1st-step PCR was performed in 50 µl volumes containing 5 µl cDNA product, 1x Phusion HF Buffer, 0.2 mM dNTPs (Sigma), 0.125 µM of each 341F-primer and 785R-primer, 0.5625% (v/v) DMSO, and 0.25 µl of Phusion HF II DNA polymerase (Thermo Fisher Scientific Inc.). PCR was performed as followed: 98°C for 40 s, followed by 15 cycles of 98°C for 20 s, annealing at 55°C for 40 s and 72°C for 40 s followed by a final extension step at 72°C for 2 min. Samples were cooled down to 8°C in a final storage step.

To enable multiplexing, barcodes were added in a 2nd-step PCR. For this, a 50 µl PCR was prepared using 2 µl of the 1st-step PCR product, 1x Phusion HF Buffer, 0.2 mM dNTPs, 0.125 µM of each forward and reverse barcode primer, which were designed as previously described by Klindworth *et al.* (2012), 0.25% (v/v) DMSO and 0.5 µl of Phusion HF II DNA polymerase. PCR conditions were 98°C for 40 s, 10 cycles of 98°C for 20 s, 55°C for 40 s and

72°C for 40 s as well as a final extension step at 72°C for 2 min. Final amplicons were validated, cleaned-up, normalized, pooled, and sequenced as described for the DNA standard method under 16S rRNA gene sequencing.

Sequencing data analysis

Data were analyzed as previously described (Lagkouvardos *et al.*, 2015). In brief, raw sequence reads were processed using IMNGS (Lagkouvardos *et al.*, 2016), an in-house-developed pipeline based on UPARSE (Edgar, 2013). Parameters were used as default settings were set. Further analyses were performed in Rhea (Lagkouvardos *et al.*, 2017).

2.1.4 Results

Evaluation of preservation buffers and storage solutions for nucleic acid extraction from stool

DNA and RNA yields and qualities differed between the different stabilizing buffers at different time points (Fig 2.1.1A). Samples stored in DNA stool stabilizer (DS) showed weaker RNA extraction potential and more degradation than samples stored in RNAlater (RL) or self-made stool stabilizer (SStab). RNA extracted from DS samples, stored for longer than two days, was prone to be degraded (Fig 2.1.1B). DNA yields were generally the highest for DNA extracted from samples stored in DS (except stored one week at RT) but were comparable in quality to those extracted from RL or SStab (Fig 2.1.1C). RNA yields, however, did not show a general trend in terms of higher concentrations based on a certain stabilizing agent (Fig 2.1.1D).

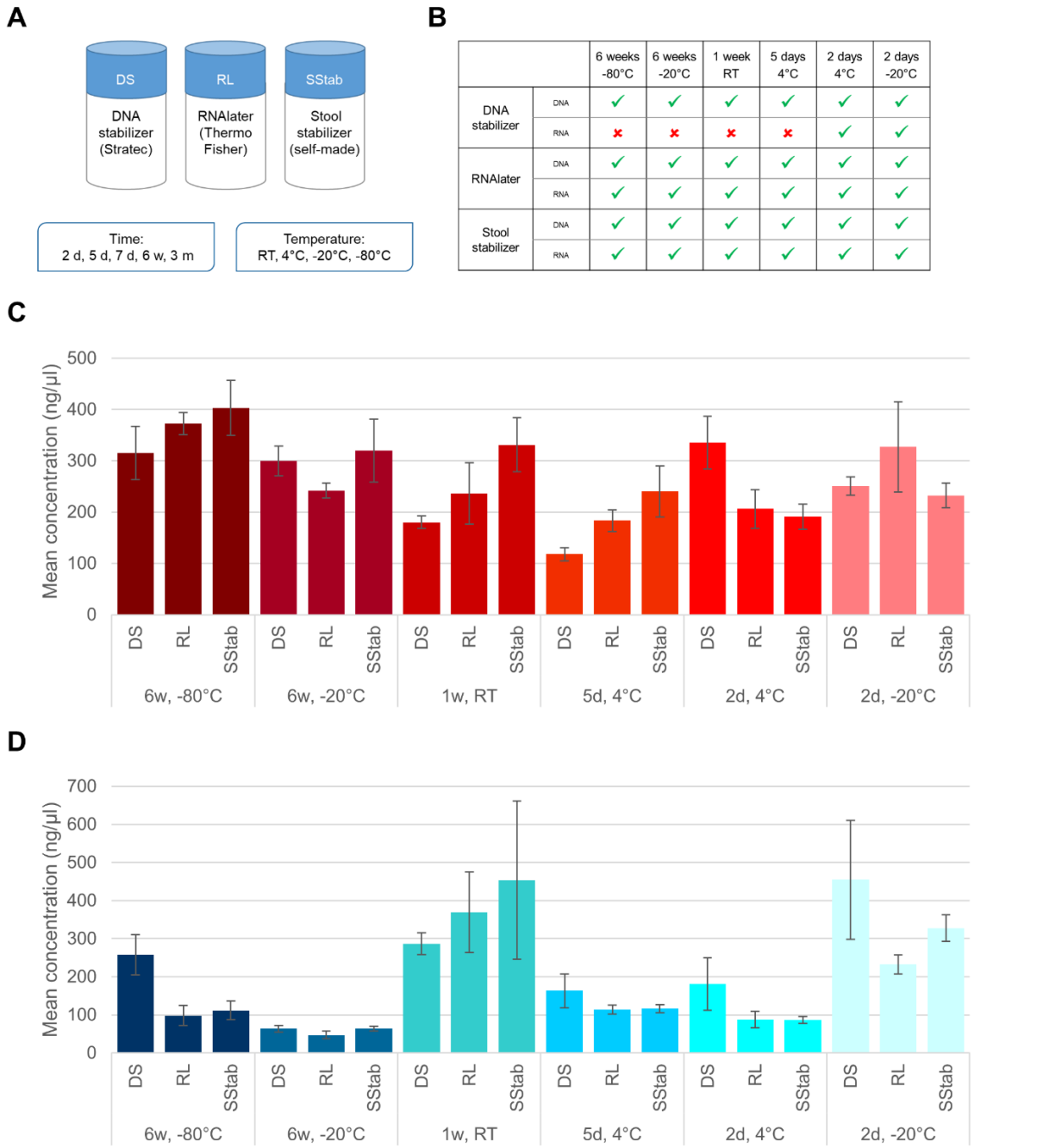

Figure 2.1.1: Three different stabilizing agents were compared, the commercially available DNA stabilizer (DS), RNAlater (RS), and the self-made stool stabilizer (SStab). DNA and RNA extraction were performed after storing fecal samples for a different time period at different temperatures (**A**). RNA and DNA extraction was successful for most samples (marked with a green hook, symbolizing effective) besides RNA extraction of stool samples stored in DS (marked with a red cross for fail) for longer than two days (**B**). Mean concentration of DNA (**C**) and RNA (**D**) in the respective tested conditions. For every condition, three biological replicates were tested, error bars represent the variation of the replicates.

In a further experiment, it was assessed whether 16S rRNA gene sequencing profiles generated from the same human donor samples, either stored in DS or SStab, produced comparable results. As distances in the non-metric multi-dimensional scaling (NMDS) plot were shown to be small (Fig 2.1.2A) and clustering was due to sample origin and not stabilizing solution, when analyzing the phylogenetic tree (Fig 2.1.2B), comparability of the stabilizing

agents was given. Here it should be noticed that samples which were prviously stored (2 days, -20°C) were used.

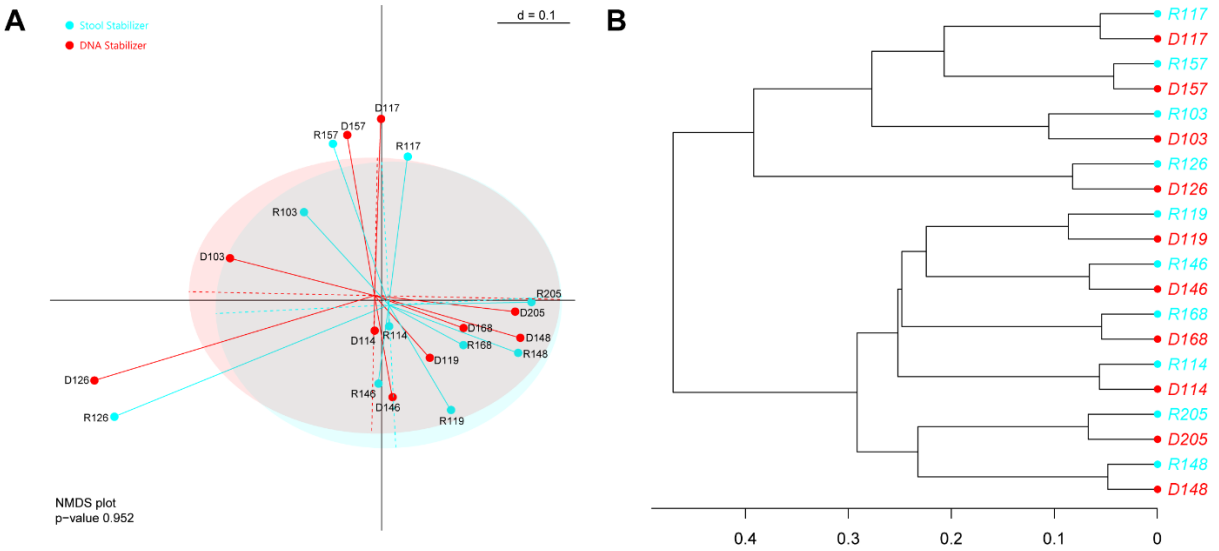

Figure 2.1.2: Beta-diversity analysis of human samples either stored in DS or SStab. NMDS plot calculated from the generalized UniFrac dissimilarity matrix (**A**) and hierarchical clustering of the samples in a phylogenetic tree (**B**) show that clustering is due to sampling origin (human donor) and not due to stabilizing agent used.

Improving total RNA and SSU rRNA extraction

For the extraction of total RNA, the performance of a previously described protocol using a phenol-chloroform-based (P/C) method (Zoetendal *et al.*, 2006) was compared to using a commercially available kit system (Zymo Quick-RNA Fecal/Soil Microbe Microprep Kit). Overall, the kit-based approach performed superior as a lower degree of degradation was observed in the kit-based electropherogram (Fig 2.1.3). Moreover, the rRNA ratio of 16S/23S, which should ideally be ≥ 2, was drastically decreased in the P/C based method compared to the kit-based extraction, which could also be observed in small 23S rRNA bands at the agarose gel loaded with RNA extracted using the P/C based method (Fig 2.1.3A). Thus, even though the P/C-based method is cheap, kit-based methods are preferred as those do not lead to partly degraded total RNA.

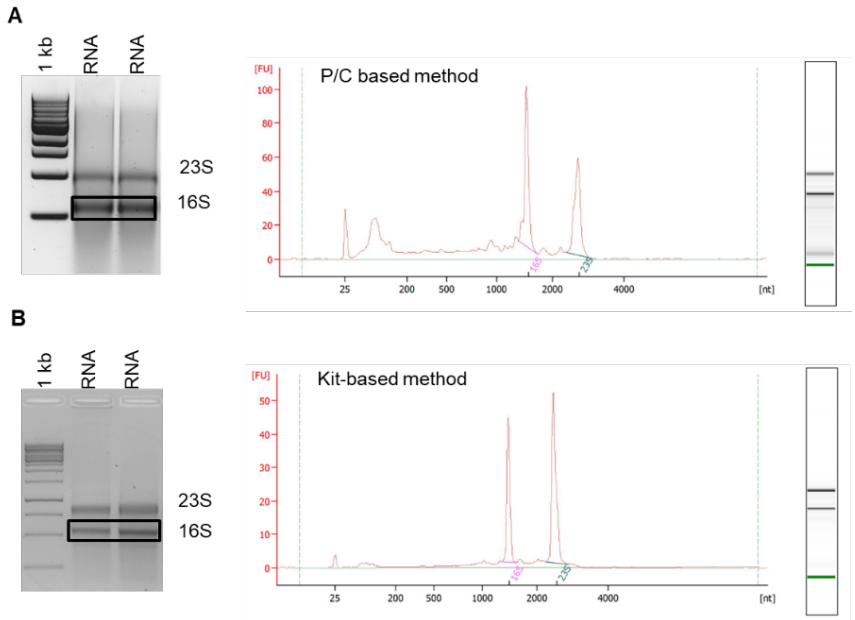

Figure 2.1.3: Agarose gel electrophoresis and electropherograms of total RNAs isolated either using a phenol-chloroform-based method (**A**) or the Zymo Quick-RNA Fecal/Soil Microbe Microprep Kit (**B**).

In the next step, it was investigated how SSU rRNA could be extracted from total RNA efficiently. Six different extraction protocols were tested and compared (see section *SSU rRNA isolation* in the method section of this part for further details). Except for the MOPS extraction method, for all other five methods, SSU rRNA could be successfully extracted and verified using a Bioanalyzer (Fig 2.1.4). Using either the squeeze or PAGE extraction, for most samples SSU rRNA could be gained and detected. Nevertheless, also the E-gel or Well-gel extraction led to good results. Using the agarose extraction protocol, only smaller amounts of SSU rRNA could be satisfactorily be shown.

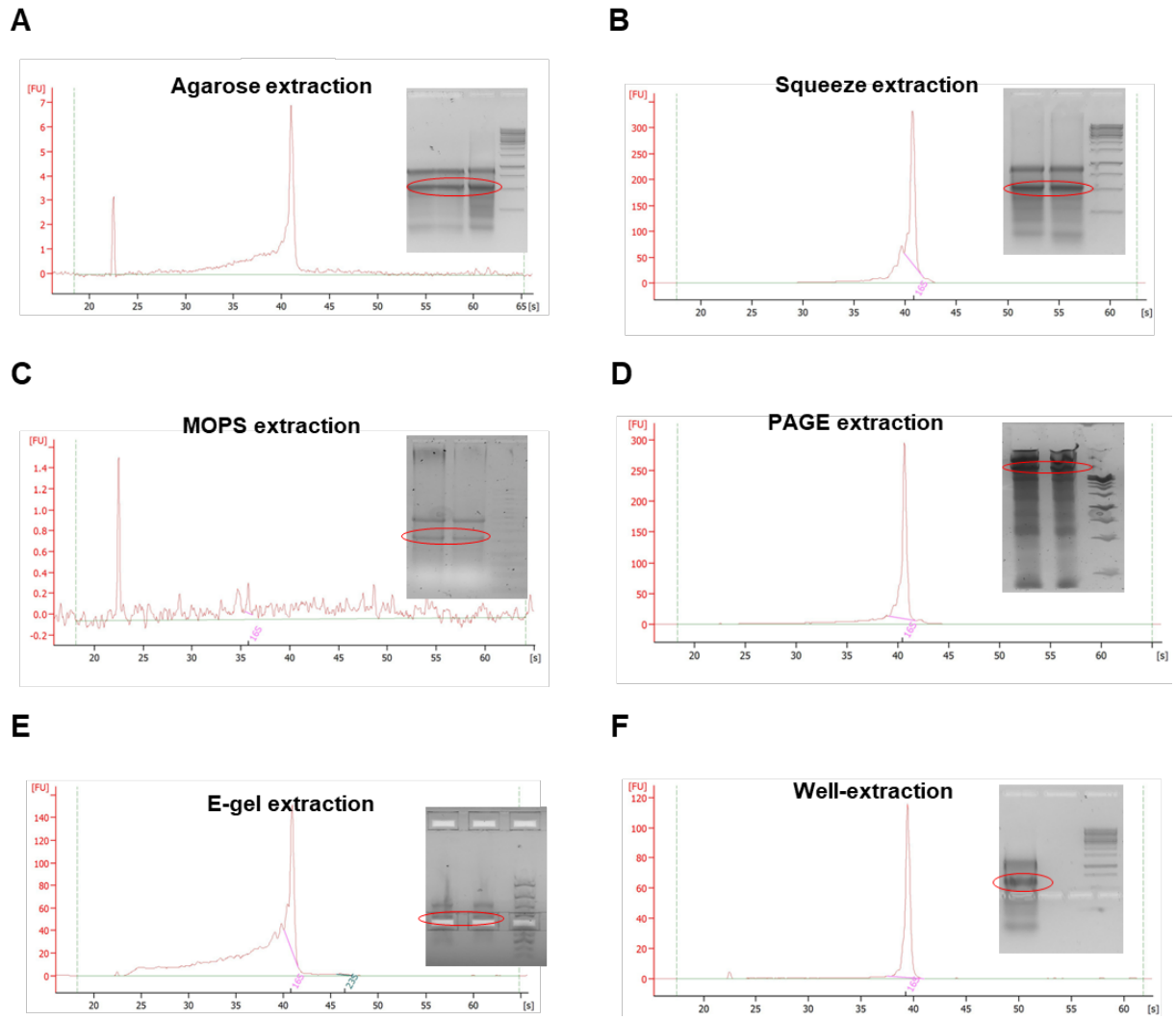

Figure 2.1.4: Agarose gel electrophoresis and electropherograms of SSU rRNAs isolated by six different extraction approaches: agarose extraction (**A**), squeeze extraction (**B**), MOPS extraction (**C**), PAGE extraction (**D**), E-gel extraction (**E**) and Well-extraction (**F**). Besides the MOPS extraction, all extractions methods allowed to successfully isolate clean SSU rRNA. The red oval in the gels marks the SSU rRNA band.

16S rRNA gene versus direct-16S rRNA sequencing

The use of DNA or RNA as a starting molecule for amplicon sequencing was investigated next. Amplicons using both nucleic acid types (DNA/RNA) as starting molecules were prepared. In general, it must be noticed that the use of RNA instead of DNA for amplicon sequencing of the 16S rRNA changed the analysis outcome due to resulting differences in the taxonomical profiles and absolute abundances of the detected taxa. At the DNA level, the presence or absence of different taxa can be determined, whereas, on RNA level, conclusions about the metabolic active taxa are possible. DNA-based rRNA gene sequencing resulted on average in 33,051 reads per sample, whereas the read count was on average lower for the direct rRNA sequencing with 21,350 reads (Tab 2.1.1). For the DNA-based approach, a mean of 145 OTUs per sample was found, and for the direct rRNA approach, on average, 132 OTUs. The β-diversity revealed that using RNA or DNA as a starting molecule leads to significant differences

in the sample composition (Fig 2.1.5A). The hierarchical clustering tree shows that only one of the ten tested samples clustered by sample origin and not by processing methodology (Fig 2.1.5B). This indicates that the results of those two different sequencing strategies cannot not directly be compared.

Table 2.1.1: Number of sequences per sample after every processing step in the IMNGS pipeline.

Method	Sample_ID	Demultiplexing	Merging	EE-filtering	Chimeras-Artifacts	OTU Abundance filter
direct rRNA	R109	18,153	14,027	14,026	9,553	9,210
	R136	19,430	15,173	15,171	10,252	10,053
	R139	20,181	15,520	15,519	9,762	9,463
	R141	27,662	17,043	17,038	15,790	15,002
	R144	19,503	15,771	15,770	11,419	11,211
	R146	18,068	14,432	14,428	9,591	9,297
	R147	19,866	15,656	15,654	10,552	10,301
	R148	21,906	17,475	17,472	12,424	11,800
	R149	23,229	18,319	18,316	10,939	10,271
	R151	25,500	20,044	20,041	12,222	11,528
rRNA gene	D109	37,179	27,971	27,969	25,687	24,694
	D136	35,456	25,748	25,748	23,062	22,214
	D139	36,246	26,646	26,644	24,783	24,365
	D141	29,063	20,912	20,912	18,824	17,860
	D144	34,875	25,431	25,427	24,193	23,918
	D146	30,341	21,357	21,357	19,623	18,753
	D147	29,503	21,376	21,373	19,231	18,711
	D148	34,518	25,099	25,098	22,975	21,949
	D149	38,050	27,894	27,894	25,616	24,473
	D151	25,282	18,279	18,276	16,931	15,809

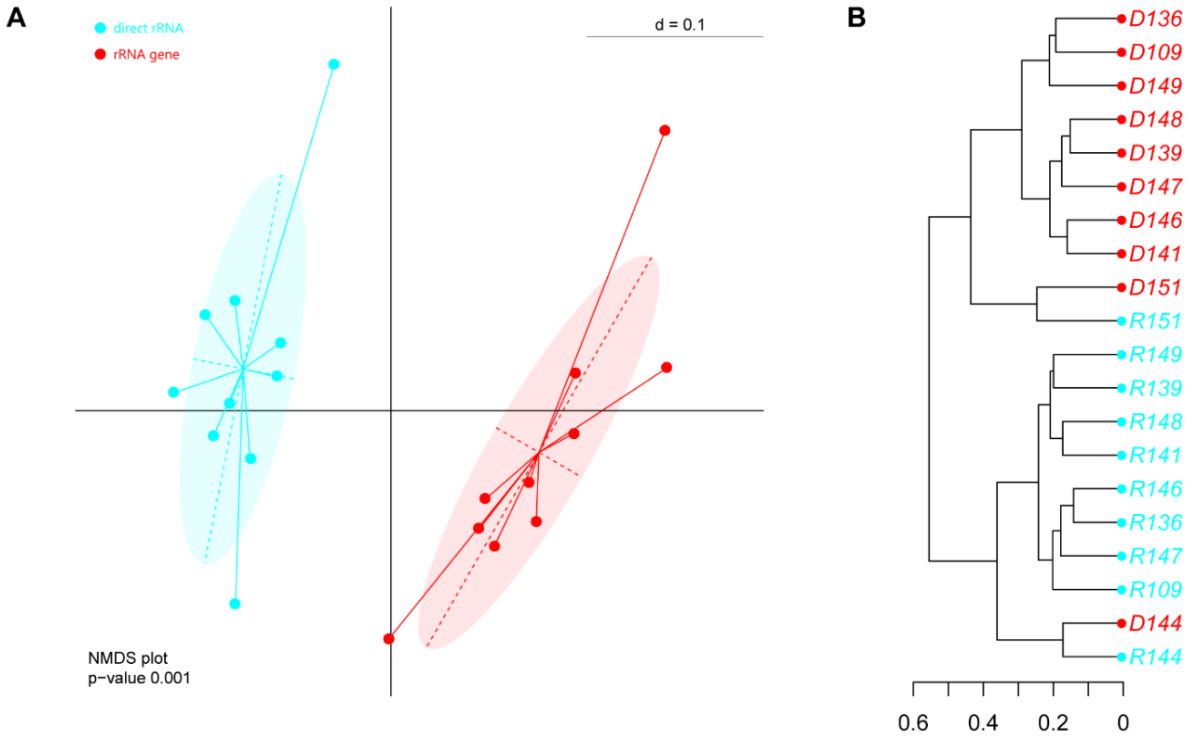

Figure 2.1.5: β-diversity reveals that using RNA or DNA as a starting molecule leads to significant differences in the analysis result (**A**). The hierarchical clustering tree shows that only one of the ten tested samples clustered by sample origin and not by processing methodology (**B**).

Phylogenetic analysis at the family level was performed for two of the tested samples. It was checked whether normalization by 16S-gene copy number affects the trend that direct rRNA sequenced samples (i.e., starting from RNA) differ in their composition compared to the rRNA gene sequenced samples (i.e., starting from DNA). Interestingly, it was shown that normalization for copy number of genes present in each genome did not affect the general trend (Fig 2.1.6A).

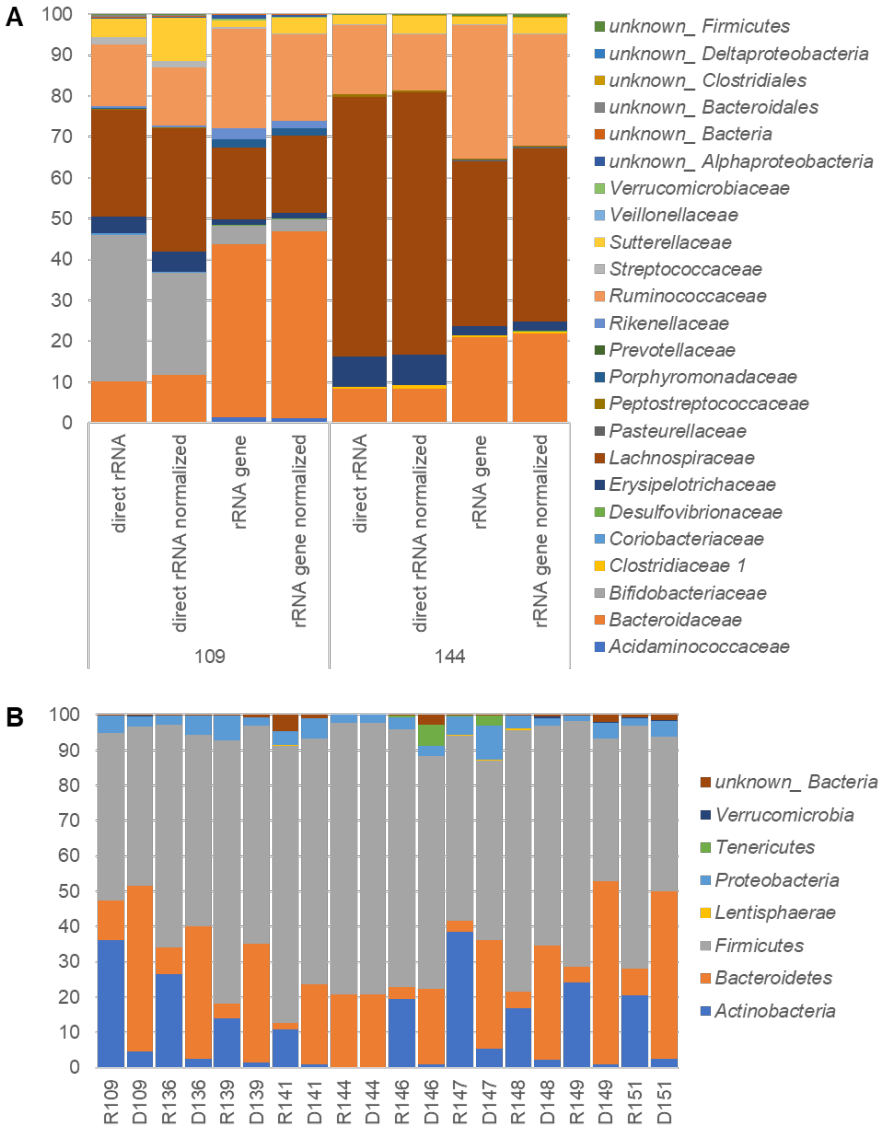

Figure 2.1.6: Normalization by 16S copy number shows no distinct shift in the microbial composition analysis. Direct rRNA samples differ in their composition compared to the rRNA gene sequenced samples. Normalization for copy number does not affect this trend (**A**). Phylogenetic distribution at phyla level of bacterial composition analyzed by rRNA gene copy (D-samples) or direct rRNA sequencing (R-samples) of selected samples (**B**).

Already, at the phylum level, differences in the phylogenetic distribution are observable (Fig 2.1.6B). Moreover, it was detected that especially the *Firmicutes*/*Bacteroides* (F/B) ratio is influenced by targeting either RNA or DNA. In general, F/B ratios are much higher for direct rRNA samples due to an increased proportion detected as *Bacteroidetes* (Tab 2.1.2).

2.1.5 Discussion

Evaluation of preservation buffers and storage solutions for RNA and DNA extraction from stool samples

In the last couple of years, studies investigating the human gut microbiome increased not only in total numbers but also in the size of participants (Chen *et al.*, 2020). Usually, samples are either taken at a hospital or study center or at home. If samples are taken at home, transportation to the study facility must be performed, and control over the storage conditions during transport is at best limited. Here, stabilizing agents, allowing samples to be transported at ambient temperature, increased in their demand and interest. Nonetheless, to our knowledge, it was not systematically tested before if the stabilizing buffers used were able to stabilize both RNA and DNA within human stool samples. Thus, the effect of different preservation buffers was tested on stabilizing DNAs and RNAs originating from human fecal samples under different environmental conditions. It was found that RL and the self-made SStab showed a better stabilizing potential than the commercial DNA stool stabilizer for stabilizing RNA. DNA extraction from samples stored in the different tested buffers resulted in similar quality and quantity of the extracted gDNA product. However, stability was only tested and valued based on overall yields and evident stability of the RNA and DNA (based on agarose gel visualization). Thus, no conclusion about the stabilizing potential of the initial sample composition could be made. Nevertheless, comparing our results to published data (Camacho-Sanchez *et al.*, 2013, Menke *et al.*, 2017), self-made stabilizing solutions seem to be good alternatives to expensive commercially available stabilizing reagents. This was further supported by 16S rRNA gene sequencing comparing human samples either stored in SStab or DS. Here, fecal samples from the same individual highly correlated regardless of the stabilizing agent used. But, one has to notice that if no preservation is needed, e.g., when samples can be directly processed or immediate freezing is possible, samples should be stored without buffer to enable also further experimental procedures such as cell-counting using flow cytometry, metabolome or proteome-analysis (Gorokhova, 2005, Hickl *et al.*, 2019). Also, multiple freezing and thawing cycles should be avoided, especially to guarantee RNA stability (Fouhy *et al.*, 2015b, Yu *et al.*, 2017). So, storing the samples in smaller sample volumes, e.g., in 600 µl portions, would be best.

Before storing, accurate mixing of the sample must be performed to enable full homogenization, which is important to allow the most efficient stabilizing effects. Only under these requirements, accuracy and efficiency of the stabilizing reagent can be assured.

Table 2.1.2: Percental distribution of samples at phylum-level either prepared as DNA (D-samples) or direct rRNA (R-samples) amplicons.

	R109	D109	R136	D136	R139	D139	R141	D141	R144	D144	R146	D146	R147	D147	R148	D148	R149	D149	R151	D151
Actinobacteria	36.12	4.60	26.45	2.61	14.12	1.48	10.87	1.02	0.11	0.00	19.44	1.01	38.50	5.36	16.82	2.36	24.26	1.06	20.44	2.52
Bacteroidetes	11.20	46.87	7.54	37.39	3.99	33.77	1.88	22.67	8.31	20.78	3.30	21.27	3.23	30.79	4.75	32.27	4.35	51.86	7.70	47.42
Firmicutes	47.51	45.18	63.24	54.33	74.69	61.75	78.52	69.53	89.45	76.87	73.22	66.00	52.39	50.98	74.03	62.26	69.57	40.51	68.94	43.87
Lentisphaerae	0.00	0.00	0.08	0.01	0.02	0.01	0.14	0.08	0.00	0.00	0.02	0.01	0.21	0.06	0.61	0.08	0.00	0.00	0.00	0.01
Proteobacteria	4.94	2.91	2.64	5.61	6.94	2.27	3.91	5.67	2.10	2.33	3.45	3.07	5.37	9.87	3.64	2.15	1.58	4.31	2.03	4.44
Tenericutes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	5.89	0.19	2.90	0.00	0.00	0.00	0.00	0.00	0.00
Verrucomicrobia	0.00	0.41	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.05	0.00	0.00	0.04	0.44	0.00	0.36	0.27	0.32
unknown_*Bacteria*	0.23	0.04	0.05	0.05	0.24	0.71	4.69	1.02	0.03	0.00	0.25	2.69	0.11	0.03	0.10	0.43	0.24	1.90	0.62	1.43
F/B ratio	4.24	0.96	8.39	1.45	18.72	1.83	41.77	3.07	10.76	3.70	22.19	3.10	16.22	1.66	15.59	1.93	15.99	0.78	8.95	0.93

Improving total RNA and SSU rRNA extraction

Even though several publications comparing DNA extraction methods from stool exist (e.g., Claassen *et al.*, 2013, Fiedorová *et al.*, 2019, Gryp *et al.*, 2020, Videnska *et al.*, 2019, Yang *et al.*, 2020), we could not find a recent publication, comparing RNA extraction kits for human stool samples. Therefore, we tested whether the established protocol by Zoetendal *et al.* (2006) or the Zymo Quick RNA isolation kit for stool samples performed superior when compared to each other. We found that the kit-based approach outperformed the phenol-chloroform-based approach of Zoetendal *et al*. in terms of quantity and purity. Nevertheless, it should be noticed that the kit-based system did not recover small RNAs very well and might, therefore, not be suitable for metatranslatomic or metatranscriptomic approaches targeting those.

Further, several different methods for the isolation of the SSU rRNA were intensively tested on their possibility to segregate the SSU easily, quickly, and reliably from total RNAs. Out of the tested methods, a few were found functional possibilities, with PAGE extraction performing overall best and E- and Well-gel extraction performing well. Quality control using capillary gel electrophoresis measures, e.g., the Bioanalyzer RNA Kits, are essential, since only when using those, the quantity and quality of the extracted SSU rRNA could be reliably determined without losing too much of the samples.

RNA versus DNA 16S rRNA sequencing approaches

Generally, direct RNA and rRNA gene sequencing approaches as such are not comparable. While rRNA gene sequencing is representing the microbial abundance due to rRNA gene numbers, with some deviation due to the ribosomal operon numbers within each genome present in the sample, direct rRNA sequencing is more an indicator of current microbial activity (Blazewicz *et al.*, 2013, Rosselli *et al.*, 2016). Even though it has been shown that some limitations for the use of rRNA as an indicator of microbial activity exist (Blazewicz *et al.*, 2013), other studies showed that using direct rRNA approaches could drastically expanse our knowledge about the active microorganisms within a given community (Kamke *et al.*, 2010, Rosselli *et al.*, 2016). Here, we showed that using rRNA gene or direct rRNA sequencing influences the microbial composition analysis, already on the phyla level. Interestingly, it could be shown that the F/B ratio is dramatically different when comparing rRNA gene sequencing and direct rRNA sequencing. This might be interesting for a lot of nutritional and health studies, as the F/B ratio is known to be correlated with obesity and other diseases (Koliada *et al.*, 2017, Ley *et al.*, 2006). Further investigation, e.g., supported by a nutritional intervention study might be conducted to confirm that hypothesis.

2.1.6 Conclusion

Here we demonstrated that it is of great importance to test and document the conditions in which samples were stored and/or processed. We found that self-made stabilizer (SStab) is ideal to store samples. Further, samples should be stored as cold as possible (≤ -80°C) for the long term. If further analyses, such as proteomic or metabolomic approaches are under consideration, native samples (samples stored without preservation buffer) should be collected as well. Here, flash freezing in liquid nitrogen (and storage therein) is possibly the best, however, this was not tested in the current study. RNA extraction for SSU rRNA gene sequencing using the Zymo extraction kit worked very well, as RNA extracted by this approach was stable and clean and not even partially degraded in contrast to phenol-chloroform-based methods. Next, SSU rRNA should be extracted from total RNA using PAGE extraction. In general, 16S rRNA gene sequencing should be performed with DNA as a starting material when the microbiota is the subject of analysis, while direct 16S rRNA sequencing can be applied to gain information about active organisms.

2.2 Primer, pipeline, parameters: Issues in 16S rRNA gene sequencing

Summary

The procedures for performing and analyzing 16S rRNA gene sequencing are still not standardized. Therefore, we used three different mock communities of increasing complexity and 33 human fecal samples to study the effects of different primer pairs, reference databases, clustering approaches, and specific settings of parameters used in the bioinformatical analysis on the resulting taxonomic profiles.

Concerning the primers, we evaluated which primer pair is most suitable for the analysis of human stool samples out of six commonly used primer pairs, spanning different V-regions. We found that either primer pair 341F/785R, spanning the V3-V4 region, or 27F/338R, spanning the V1-V2 regions, should be used, as those produced when using mock communities, the most comparable results to the expected proportions of the included bacteria.

Next, we evaluated which reference database would be most suitable to use. Here we found that either Silva or RDP should be used for further analysis. GreenGenes was found to be outdated, and therefore, we strongly recommend avoiding this database.

We further evaluated whether the use of different clustering approaches such as the standard OTU approach or clustering approaches which include denoising steps (ASVs or zOTUs) affected the resulting taxonomical profiles severely. Here we found out that clustering was improved by using denoising approaches (ASVs or zOTUs), but generally, differences were minor compared to the biological differences produced due to targeting different V-regions or the use of different primer pairs.

In the last step, we checked whether the use of different pipeline settings drastically influenced resulting taxonomic profiles. Concerning that, we could show that the setting for truncation is an important parameter. When targeting V4, the truncation should be set to 250 bp and 180 bp for forward and reverse reads, respectively. However, we found that different settings do not affect the resulting taxonomical profiles severely. Nevertheless, every pipeline setting should be listed and described in publications to evaluate and assess possible biasing factors when comparing different datasets.

Generally, we recommend creating specific and sufficiently complex mock communities to test the desired study design and to check for possible biasing or influencing factors that could occur during amplicon generation and downstream sequence analysis steps.

Complex mock communities are of importance as we could show that less complex mock communities do not reflect complex samples and thus do not show possible problematic features or biases that would occur when targeting complex samples.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE

Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing

Isabel Abellan-Schneyder,[a] Monica S. Matchado,[b] Sandra Reitmeier,[a] Alina Sommer,[a] Zeno Sewald,[a] Jan Baumbach,[b,c,d] Markus List,[b] Klaus Neuhaus[a]

[a]Core Facility Microbiome, ZIEL—Institute for Food & Health, Technische Universität München, Freising, Germany
[b]Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technische Universität München, Freising, Germany
[c]Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[d]Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany

ABSTRACT Short-amplicon 16S rRNA gene sequencing is currently the method of choice for studies investigating microbiomes. However, comparative studies on differences in procedures are scarce. We sequenced human stool samples and mock communities with increasing complexity using a variety of commonly used protocols. Short amplicons targeting different variable regions (V-regions) or ranges thereof (V1-V2, V1-V3, V3-V4, V4, V4-V5, V6-V8, and V7-V9) were investigated for differences in the composition outcome due to primer choices. Next, the influence of clustering (operational taxonomic units [OTUs], zero-radius OTUs [zOTUs], and amplicon sequence variants [ASVs]), different databases (GreenGenes, the Ribosomal Database Project, Silva, the genomic-based 16S rRNA Database, and The All-Species Living Tree), and bioinformatic settings on taxonomic assignment were also investigated. We present a systematic comparison across all typically used V-regions using well-established primers. While it is known that the primer choice has a significant influence on the resulting microbial composition, we show that microbial profiles generated using different primer pairs need independent validation of performance. Further, comparing data sets across V-regions using different databases might be misleading due to differences in nomenclature (e.g., *Enterorhabdus* versus *Adlercreutzia*) and varying precisions in classification down to genus level. Overall, specific but important taxa are not picked up by certain primer pairs (e.g., *Bacteroidetes* is missed using primers 515F-944R) or due to the database used (e.g., *Acetatifactor* in GreenGenes and the genomic-based 16S rRNA Database). We found that appropriate truncation of amplicons is essential and different truncated-length combinations should be tested for each study. Finally, specific mock communities of sufficient and adequate complexity are highly recommended.

IMPORTANCE In 16S rRNA gene sequencing, certain bacterial genera were found to be underrepresented or even missing in taxonomic profiles when using unsuitable primer combinations, outdated reference databases, or inadequate pipeline settings. Concerning the last, quality thresholds as well as bioinformatic settings (i.e., clustering approach, analysis pipeline, and specific adjustments such as truncation) are responsible for a number of observed differences between studies. Conclusions drawn by comparing one data set to another (e.g., between publications) appear to be problematic and require independent cross-validation using matching V-regions and uniform data processing. Therefore, we highlight the importance of a thought-out study design including sufficiently complex mock standards and appropriate V-region choice for the sample of interest. The use of processing pipelines and parameters must be tested beforehand.

KEYWORDS 16S rRNA gene sequencing, amplicon sequencing, variable regions, clustering, bioinformatic settings, microbiome, databases, mock communities

Citation Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K. 2021. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. mSphere 6:e01202-20. https://doi.org/10.1128/mSphere.01202-20.

Editor Susannah Green Tringe, U.S. Department of Energy Joint Genome Institute

Address correspondence to Klaus Neuhaus, neuhaus@tum.de.

Received 26 November 2020
Accepted 4 February 2021
Published 24 February 2021

The human gut microbiome is a complex environment hosting a large number of different bacteria. A cost-effective method to determine the bacterial composition of, e.g., human fecal samples is to sequence amplicons targeting the 16S rRNA gene. Microbial compositions of diverse environments, which are influenced by different factors or conditions (e.g., sampling time point, targeted rRNA region, response to health or disease, sequencing strategy, machinery, depth, and read lengths), were also studied with this method (1–7).

The 16S rRNA gene spans about 1,500 bp and is structured in highly conserved regions interspersed with nine variable regions (V-regions), V1 to V9 (8, 9). The conserved regions can be used for primer binding and thus allow for capturing a greater number of different bacterial taxa, sometimes including or not including archaea, while the variable regions permit the discrimination of these taxa within different microbial environments (10). However, differences between the conserved regions and, therefore, differences in primer annealing result in an unequal amplification of bacteria present in a sample (11). Depending on the particular V-region that was targeted, differences in the sequencing results and taxonomic outcome occurred, which led to misinterpretation (12, 13). Further, not every variable region has the same sensitivity, i.e., allowing separation of closely related taxa (14). Concerning archaea, the applicability of certain primer pairs has been covered well in previous studies (12, 15, 16).

Second-generation sequencers, e.g., Illumina's MiSeq, enable sequencing of amplicons up to 600 bp with high accuracy. This length allows targeting about one to three adjacent variable regions of the 16S rRNA gene using "universal" primers for the conserved regions. In a subsequent PCR, sequencing adapters are added to the amplicons (17). After a cleanup step, the amplicon libraries are sequenced. The resulting reads are used to analyze similarities and differences between samples with different microbial compositions (e.g., alpha- and beta-diversity) (18). In contrast, full-length 16S rRNA gene sequencing is possible by using third-generation sequencers, for instance, Oxford Nanopore MinION (19) and the PacBIOs Sequel (20), which were introduced in 2009 and 2008, respectively. The greatest advantage is the long read length (up to 10,000 bp) and sequencing on a single-molecule level in a short time. These long reads enable an improved identification of bacterial taxa, as shown in several recent studies (21–27). Nevertheless, significant drawbacks include the relatively high error rate (up to 15% per sequence) (28, 29), limited applicability in high-throughput studies, higher general costs, and even less standardization of protocols and analysis pipelines. However, despite the widespread use of 16S rRNA gene sequencing, there is a need to better understand the differences between the targeted region and the data analysis pipeline chosen in amplicon sequencing of the 16S rRNA genes.

For short-amplicon sequencing, a literature survey showed that the regions V1-V2/V3 (30, 31), V3-V4/V5 (32–34), and V4 (35, 36) are most commonly used. However, the taxonomic classification differs considerably when targeting different variable regions (37), affecting attempts to perform cross-study comparison and leading to further biases in compositional analysis, where short-amplicon primers are not as universal as desired (11, 38). Since the taxonomic resolution seems to differ for some phyla for different variable regions (39), closely related bacterial species and genera might be indistinguishable (40). Moreover, the choice of bioinformatic processing pipelines and analysis tools is known to influence the results (41–44). Different 16S rRNA gene-specific taxonomic classification methods, such as Mothur (45), Qiime (46), Qiime2 (47), DADA2 (48), and others, were developed. During data processing, sequences are clustered into operational taxonomic units (OTUs) at a threshold of 97% sequence similarity. Sequence representatives, i.e., sequences with the least mismatches to other sequences in a cluster, are used for taxonomic assignment. Amplicon sequence variants (ASVs) or zero-radius OTUs (zOTUs) have been suggested as alternatives to OTUs (48, 49), as they correct for sequencing errors by different denoising approaches. In contrast to OTUs, these clusters are supposed to contain reads originating only from the same bacterial species, enabling a cross-study comparison (49, 50). In any case, after clustering, sequences are classified for

taxonomic assignment using databases of known 16S rRNA gene sequences, e.g., GreenGenes (GG) (51), the Ribosomal Database Project (RDP) (52), Silva (53), the genomic-based 16S rRNA Database (GRD) (54), or The All-Species Living Tree (LTP) (55). Not only different pipelines and reference databases but also settings of a given pipeline influence the results and are an often-overlooked bias in microbiome studies (42, 56–58). Nevertheless, some biases occurring in 16S rRNA gene amplicon sequencing have already been addressed in the past. Well-studied biasing factors, for instance, include sampling and storage procedures (59–63), DNA extraction methods (64–68), choice of variable region and primers (12, 36, 69–72), library preparation and sequencing strategies (73–76), and sequence data processing, including denoising, taxonomic classification, and the use of distinct bioinformatic tools (42, 56–58). Further, the use of negative controls and mock communities as internal standards to detect contamination or aberrancies in the sequencing results was proposed (77–79).

In this study, we joined several of these separate issues to raise awareness that the combination of primer sequence choice, clustering methods, reference database, and analysis parameters must be considered thoroughly to avoid increased bias. Thus, we created a large benchmark data set of 16S rRNA gene amplicon sequences, targeting different V-regions of the 16S rRNA gene, and systematically tested different software tools with different sets of parameters for the analysis. We sequenced three mock communities of increasing complexity with known composition, along with complex human fecal samples for comparison.

RESULTS

We systematically assessed the global influence of multiple parameters in mock communities of known composition and in human samples (Fig. 1). First, the choice of primers targeting different variable regions of the 16S rRNA gene was evaluated. We show that primer choice influences the taxonomic composition, visible in a multidimensional scaling (MDS) plot of samples originating from the same donor (Fig. 2). Second, we investigated how, and in what magnitude, the use of different clustering approaches and taxonomy assignment methods influences the results for the classification of bacterial taxonomies.

Primer choice influences the estimated microbial composition. A set of different 16S rRNA gene sequencing primer pairs covering one, two, or three of the variable regions V1 to V9 is commonly used for the analysis of microbial compositions. Depending on the input material (e.g., human gut samples, water analysis, sludge, food research, etc.), different primer pairs are used. In this study, we investigated seven different primer pairs, 27F-338R (V1-V2), 27F-534R (V1-V3), 341F-785R (V3-V4), 515F-806R (V4), 515F-944R (V4-V5), 939F-1378R (V6-V8), and 1115F-1492R (V7-V9), for the analysis of human gut samples and mock communities (Fig. 1 and Table 1). The use of different primer pairs led to primer-specific and not mainly donor-specific clustering of human stool samples (Fig. 2). These differences varied according to the analyzed taxonomic level. Differences were found to be less pronounced at higher taxonomic levels, e.g., phylum level compared to genus level (Fig. 2A and C). When analyzing samples from the same human donor but sequenced using different primer pairs, some taxa are unique for certain primer pairs. For instance, when analyzing human sample 1 (Fig. 3), *Verrucomicrobia* was detected only when using 341F-785R (V3-V4), 515F-806R (V4), 939F-1378R (V6-V8), and 1115F-1492R (V7-V9) primers and not 27F-338R (V1-V2), 27F-534R (V1-V3), or 515F-944R (V4-V5). Comparisons of samples derived from the same human donor but sequenced using different primer pairs become even more difficult at the genus level (see Fig. S2 in the supplemental material). This was mainly due to differences in the prevalence of genera when using different V-regions. A large number of reads were not classified down to genus level in either one or several V-regions and were thus considered "unknown." Importantly, the 515F-944R (V4-V5) primer pair seemed to produce results with only a few overlaps with other primer pairs (Fig. 2) and displayed a low abundance of *Bacteroidetes* (Fig. 3; Fig. S2). We analyzed whether this was due to a much lower theoretical coverage of known bacterial species. Therefore, all primers were

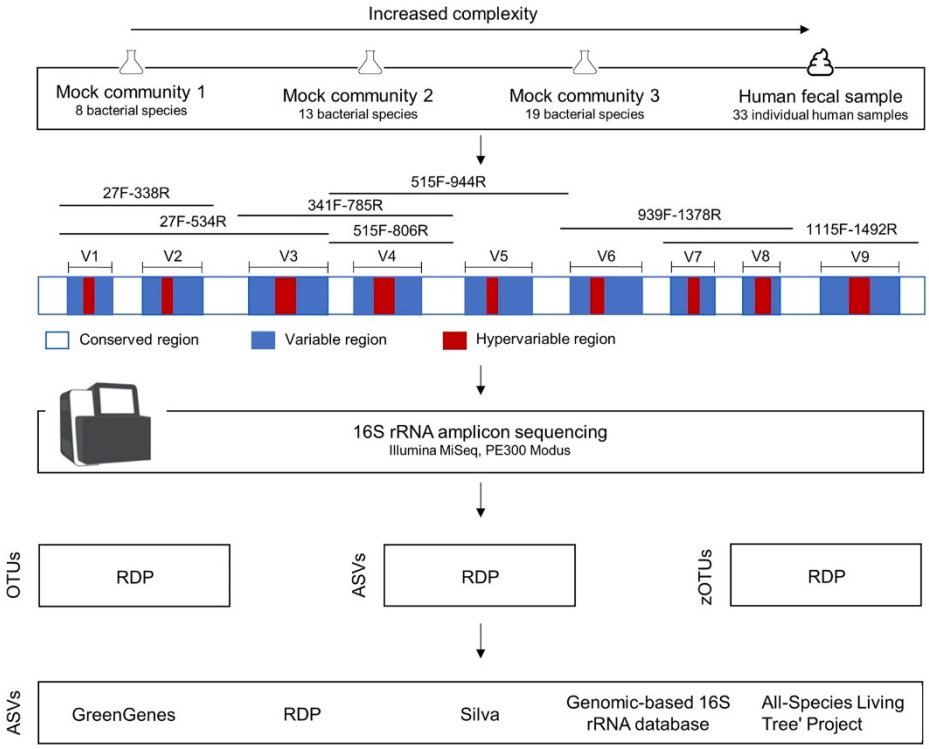

FIG 1 Overview of the analysis strategies used in this study. DNAs from different sample types with increasing complexity (i.e., 3 mock communities and 33 human stool samples) were extracted. Amplicons were generated using different primer pairs targeting different V-regions and sequenced on an Illumina MiSeq. Afterwards, the impacts of different clustering approaches and reference databases on the microbial profiles were investigated.

evaluated *in silico* for their theoretical coverage on all bacterial genera using the Silva database. While the theoretical coverage for 515F-944R (V4-V5) primers was lower than for the primer pairs 27F-338R (V1-V2), 27F-534R (V1-V3), 341F-785R (V3-V4), and 515F-806R (V4), we found the theoretical coverage for primer pairs 939F-1378R (V6-V8) and 1115F-1492R (V7-V9) to be even lower (Table S2). Thus, we believe that the low coverage of *Bacteroidetes* is the main reason for primer pair 515F-944R (V4-V5) to form an outlier.

Clustering approaches have minor influence on taxonomic profiles. In addition to the 97%-identity OTU approach, ASV clustering gained a lot of attention in the latest studies (43). Due to its improved resolution and thus better comparability of results between different studies, it is nowadays a popular and often favored method. In this study, we tested whether different clustering approaches have an influence on the assigned taxonomic profiles for the ZIEL-I mock community. Thus, we compared ASVs, zOTUs, and OTUs. Overall, the clustering methodology seemed to have only a minor effect on the assigned taxonomic composition compared to the effect of primer choice (Fig. 4A). Again, the 515F-944R (V4-V5) primer pair showed profiles distinct from those found for all other primer pairs used, no matter which clustering was used. Differences observed for each clustering approach were mainly due to identification problems at the genus level. When using the ASV approach for clustering the data, *Bacillus* could not be classified down to genus level. In contrast, this was possible when using zOTU and OTU approaches. Similarly, *Enterococcus* was not assigned correctly by the 27F-534R (V1-V3) primer pair using the ASV approach. Overall, we found that ASVs

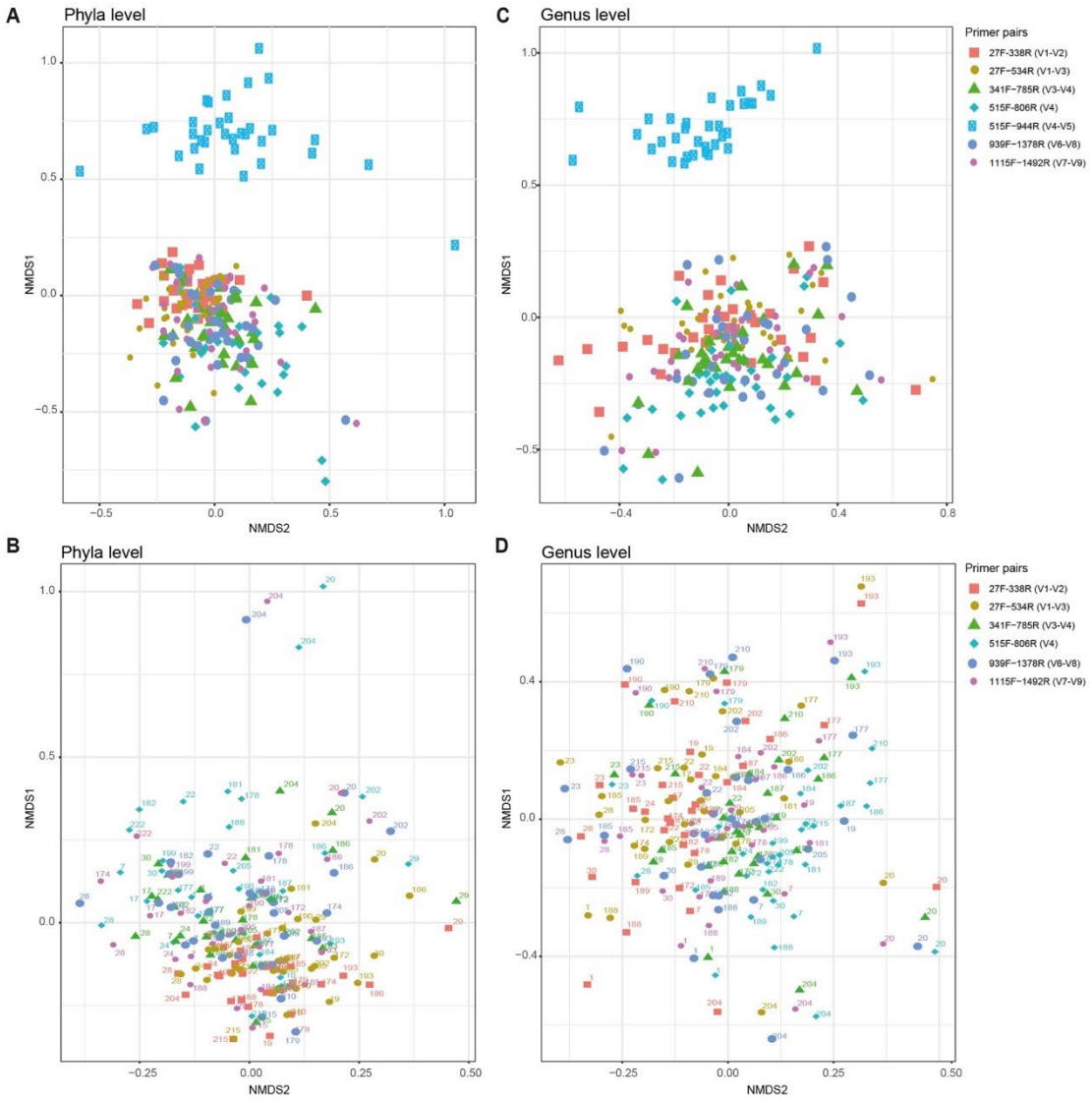

FIG 2 NMDS plots for the microbiome composition of human samples. Sample similarity is shown at phylum level (A and B) and at genus level (C and D). Different primer pairs are indicated to the right for all panels. Top panels (A and C) include processing the V4-V5 region, while for the bottom panels (B and D) this region has been omitted since results using 515F-944R primers (blue squares in panels A and C) fall separately from all other clusters. Labeling of the samples in the bottom panels (B and D) is based on donor number.

performed best for most of the other genera, as differences between theoretical values and expected amounts of the distinct taxa were the smallest here (Table S3). The additional analysis of a human sample subset resulted in results comparable to those for the ZIEL-I mock community (example of one representative sample is shown in Fig. 4B). Differences in taxonomic profiles are more dependent on primer pairs used than on clustering approach. Smaller variations occurred mostly due to problems assigning genera; e.g., identification of members of the *Lachnospiraceae* family on the genus level is not possible for zOTUs when using primer pairs 515F-944R (V4-V5) and

TABLE 1 V-region-specific forward and reverse primers and annealing temperature for 1st step PCR

V-region	Forward primer	Reverse primer	Forward sequence (5′–3′)	Reverse sequence (5′–3′)	Specificity	Annealing temp (°C)	Reference
V1-V2	27F	338R	AGA GTT TGA TYM TGG CTC AG	GCT GCC TCC CGT AGG AGT	Universal[a]	57	Salter et al. (115)
V1-V3	27F	534R	AGA GTT TGA TYM TGG CTC AG	ATT ACC GCG GCT GCT GG	Universal	57	Walker et al. (84)
V3-V4	341F	785R	CCT ACG GGN GGC WGC AG	GAC TAC HVG GGT ATC TAA TCC	Universal	55	Klindworth et al. (70)
V4	515F	806R	GTG CCA GCM GCC GCG GTA A	GGA CTA CHV GGG TWT CTA AT	Universal	53	Caporaso et al. (116)
V4-V5	515F	944R	GTG CCA GCM GCC GCG GTA A	GAA TTA AAC CAC ATG CTC	Bacterial	53	Fuks et al. (117)
V6-V8	939F	1378R	GAA TTG ACG GGG GCC CGC ACA AG	CGG TGT GTA CAA GGC CCG GGA ACG	Bacterial	58	Lebuhn et al. (118)
V7-V9	1115F	1492R	CAA CGA GCG CAA CCC T	TAC GGY TAC CTT GTT ACG ACT T	Bacterial	51	Turner et al. (119)

[a]Universal, binds to archaea and bacteria.

1115F-1492R (V7-V9). Still, neither OTU nor zOTU clustering caused a larger bias, and thus, the influence of clustering is limited.

Sample taxonomies are influenced by reference databases. Ideally, the 16S rRNA gene sequences should reflect the organism the sequence came from. However, this depends not only on the primer pairs used or how sequence data were extracted from the raw data but also on the quality of the reference database and thus the taxonomic classification. We systematically tested five different databases commonly used: GG, RDP, Silva, GRD, and LTP.

When analyzing the Zymo mock community, which includes only eight different bacteria, we observed just a few minor differences in the assigned taxonomy for different V-regions used. Further, differences were relatively minor using different reference databases in the analysis (Fig. 5A). Using RDP for primer pair 515F-806R (V4), *Bacillus* could not be classified at the genus level but was at least assigned to *Bacillales* at the family level. The classification of *Escherichia/Shigella* was most accurate when using Silva or RDP as a reference database; thus, it displayed the lowest deviation from the ideal composition of the mock community. GG could not identify *Escherichia/Shigella* and *Listeria* at the genus level and showed poor results. When using the Zymo mock community, GG might be dismissed as an inferior database, but all other parameters seemed to have no major impact. However, as a mock community of only eight bacterial species provides only limited insights, we used two further, more complex mock communities.

The ZIEL-I mock community consists of 13 species in 13 genera (Fig. 5B) and uses bacteria, which would be expected in the gut. Analyzing this, GG performed worst again. No genus-level classification for *Acetatifactor*, *Bacillus*, *Clostridium*, and *Pseudomonas* was possible using GG as a reference. GRD classified neither *Bacillus* nor *Pseudomonas* down to genus level. The other databases worked reasonably well but with some differences between V-regions. As before, 515F-944R (V4-V5) data performed worst. Only 4 to 8 taxa were classified at genus level, whereas between 9 and 13 taxa (Table 2) were identified when analyzing the data generated by using the primer pair 341F-785R (V3-V4). *Actinomyces*, *Alistipes*, *Bacteroides*, *Cellulosimicrobium*, *Parabacteroides*, and *Flavonifractor* were not detected with the primer pair 515F-944R (V4-V5) at the genus level irrespective of the reference database used.

The ZIEL-II mock community increased the complexity of the comparison by including 19 bacteria in 18 genera. Furthermore, we purposely included species which showed difficulties in past experiments (data not shown). Again, the 515F-944R (V4-V5) primer pair showed inadequate performance irrespective of the database. Using the Silva database, 14 to 18 taxa were classified at genus level for primer pair 341F-785R (V3-V4), whereas only 7 to 9 taxa were found for data corresponding to primer pair 515F-944R (V4-V5) at genus level (Table 2). *Akkermansia* could not be identified using the 27F-338R (V1-V2) primers (Fig. 5C). *Microbacterium* was underrepresented when using the 341F-785R (V3-V4) primers. *Enterobacter* and *Ruminococcus* were best classified by Silva. Generally, most accurate taxonomic classifications were possible when using Silva or RDP as the reference database. Silva even had the smallest amount of unknown genus-level identifications, followed by RDP, LTP, GRD, and GG.

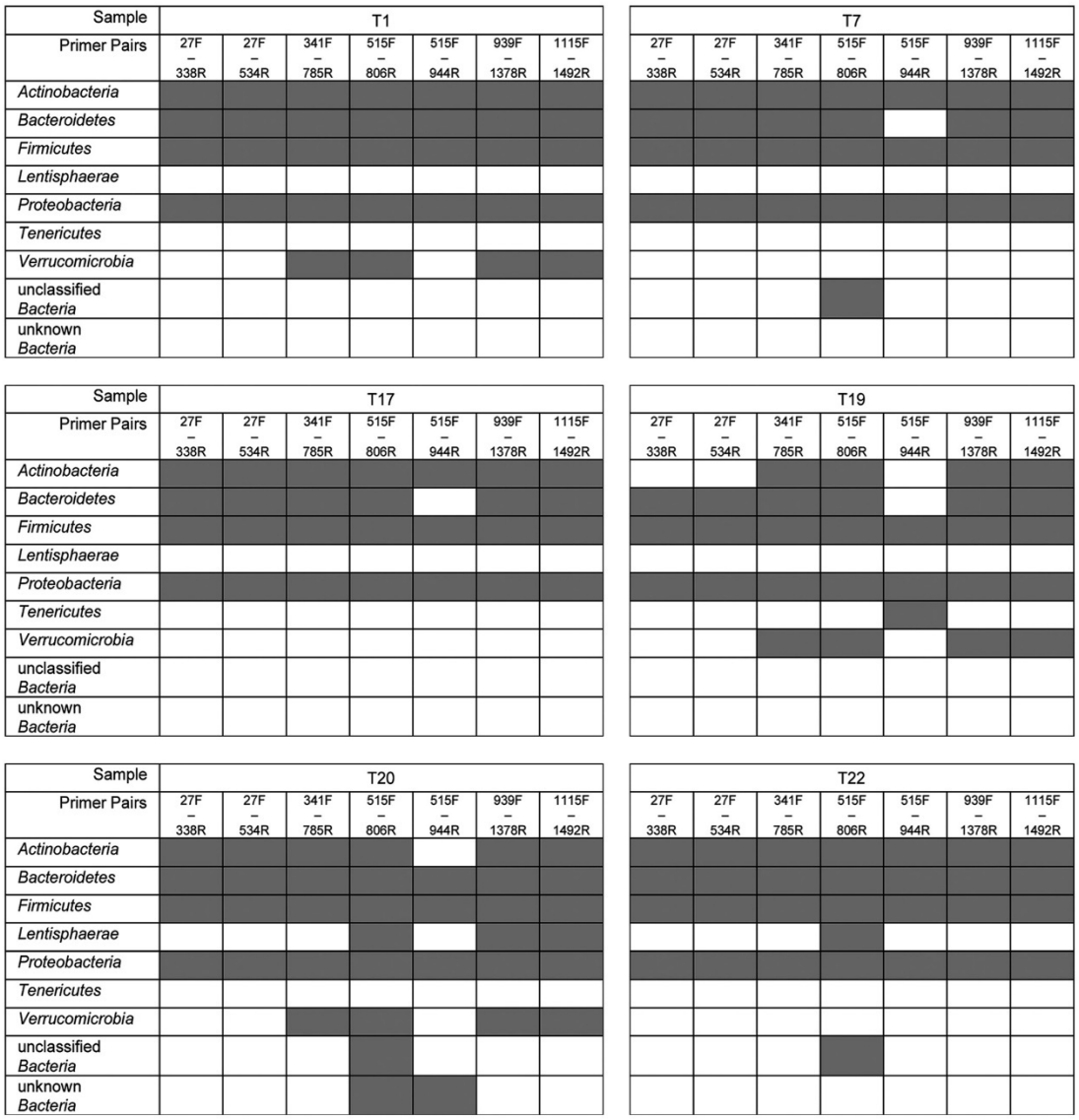

FIG 3 Presence-and-absence map of human samples on phylum level for different V-regions. Gray represents present taxa, and white represents absent taxa. Primers and their V-region spanning are given in Table 1.

Specific pipeline settings have minor influences on taxonomic classification. As clustering methodologies showed a minor influence and the use of different reference databases a more severe impact on taxonomical profiles, we also assessed the potential influence of specific pipeline parameters. As ASVs performed slightly better than zOTUs and OTUs, we focused our comparison on ASVs. Processing steps include removal of primers and adapters, trimming of low-quality reads, chimera removal, and merging of paired-end reads. The removal of all primer and adapter sequences is required for ASV production. Incorrect removal or insufficient trimming leads to loss of sequences in the merging and

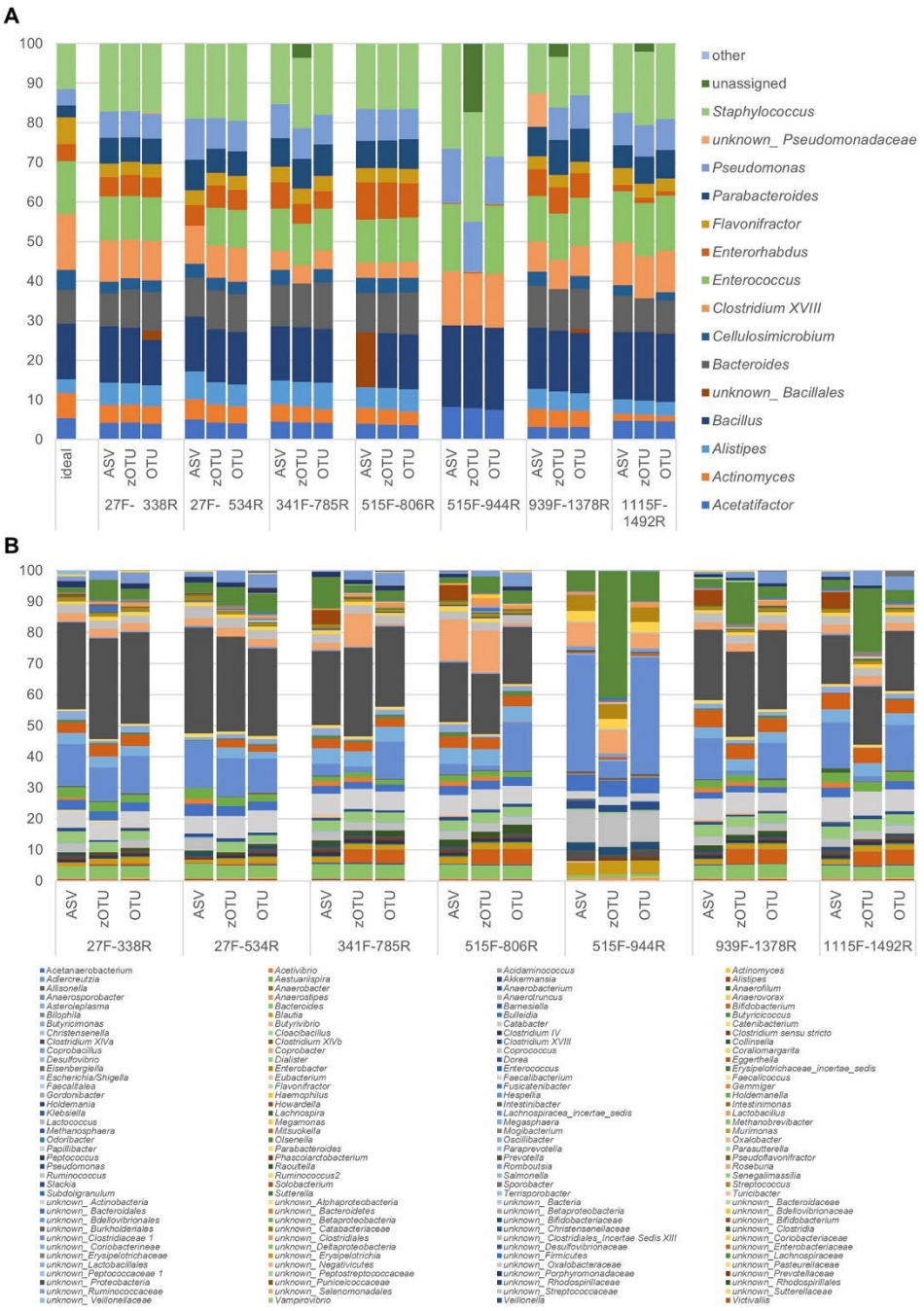

FIG 4 Comparison of the influence of the clustering method on taxonomic designation for the ZIEL-I mock community (A) and an example of a representative human sample T1 (B). The genus-level composition is shown according to ASVs, zOTUs, and OTUs as indicated. "Other" represents taxa not matching the composition of the mock community, while "unassigned" represents reads that could not be assigned to any taxonomic classification (RDP was used as a reference database). Primers and their V-region spanning are given in Table 1.

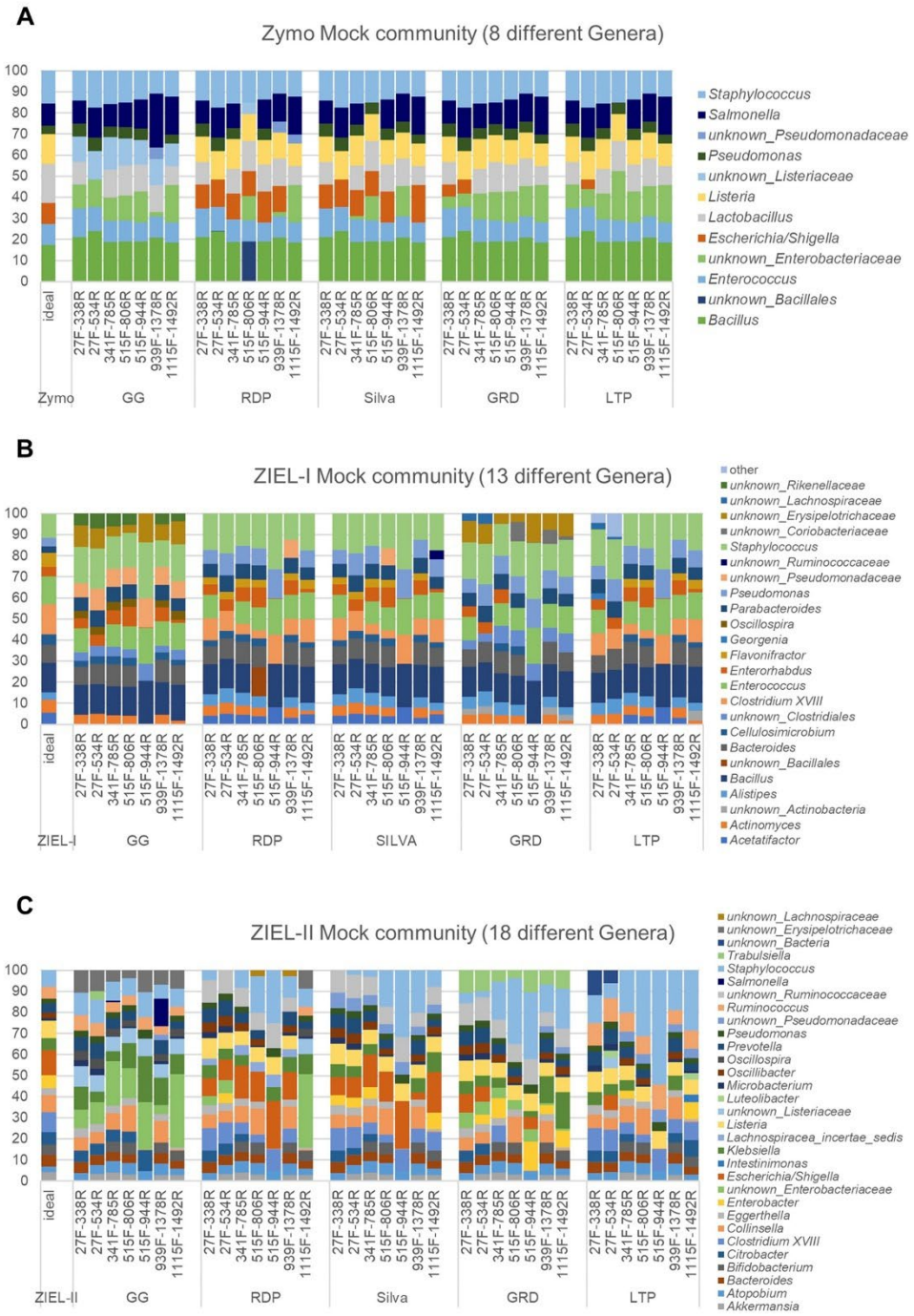

FIG 5 Comparison of mock communities sequenced over different V-regions, processed using different databases as references (GG, GreenGenes; RDP, Ribosomal Database Project; GRD, the genomic-based 16S rRNA database; LTP, The All-Species Living Tree Project) at genus level. Primers and their V-region spanning are given in Table 1.

TABLE 2 Number of ASVs and number of assigned taxa[a]

		Zymo (8 species)		ZIEL-I (13 species)		ZIEL-II (19 species)	
Primer pairs	Database	ASVs	Assigned taxa	ASVs	Assigned taxa	ASVs	Assigned taxa
27F-338R (V1-V2)	GG	19	6 [2]	19	9 [4]	34	13 [3]
	RDP	19	8	19	13	34	16 [2]
	Silva	19	8	19	13	34	17
	GRD	19	8 [1]	33	9 [4]	34	14 [3]
	LTP	19	7 [1]	33	11 [2]	34	13 [2]
27F-534R (V1-V3)	GG	23	5 [2]	36	8 [4]	45	15 [3]
	RDP	23	7 [1]	36	12	45	17 [1]
	Silva	23	7	36	12	45	18 [1]
	GRD	24	7 [1]	36	8 [4]	45	16 [3]
	LTP	23	7 [1]	36	9 [2]	45	15 [3]
341F-785R (V3-V4)	GG	12	6 [2]	22	9 [4]	33	15 [3]
	RDP	12	8	22	13	33	17 [1]
	Silva	12	8 [1]	22	13	33	18 [1]
	GRD	12	7 [1]	22	10 [2]	33	14 [3]
	LTP	12	7 [1]	22	13	33	17 [1]
515F-806R (V4)	GG	9	6 [2]	18	9 [4]	25	14 [3]
	RDP	9	5 [3]	18	12 [1]	25	15 [2]
	Silva	9	7 [1]	18	12 [1]	25	15 [1]
	GRD	9	7 [1]	18	9 [3]	25	12 [3]
	LTP	9	6 [1]	18	13	25	14 [1]
515F-944R (V4-V5)	GG	9	6 [2]	9	4 [3]	18	7 [3]
	RDP	9	8	9	7	18	9 [1]
	Silva	9	8	9	7	18	8 [1]
	GRD	9	7 [1]	9	4 [3]	18	8 [3]
	LTP	9	7 [1]	9	7	18	7 [1]
939F-1378R (V6-V8)	GG	17	5 [3]	21	9 [4]	33	13 [4]
	RDP	17	7 [2]	21	12 [1]	33	15 [2]
	Silva	17	7 [1]	21	13	33	16 [1]
	GRD	17	7 [1]	21	8 [4]	33	15 [3]
	LTP	17	7 [1]	21	13	33	16 [1]
1115F-1492R (V7-V9)	GG	15	6 [2]	18	9 [4]	36	13 [3]
	RDP	15	6 [2]	18	13	36	16 [3]
	Silva	15	8	18	12 [1]	36	17 [1]
	GRD	15	7 [1]	18	8 [4]	36	15 [3]
	LTP	15	7 [1]	18	12 [1]	36	16 [1]

[a]Assigned taxa are at the genus level; brackets indicate that taxa are unknown at the genus level. The Zymo, ZIEL-I, and ZIEL-II mock communities contain 8, 13, and 19 bacterial species, respectively (for ZIEL-II, 18 at genus level, when *Escherichia/Shigella* fall into one cluster). Shading in green indicates good identification (the darker the better), while yellow and darker shading indicates inferior outcomes.

chimera removal steps. Ambiguous nucleotides would, for example, cause a problem, as default merging settings require a minimum overlap length of 20 bp and identical sequences in forward and reverse reads. Still, we expected the truncation step to have the largest impact on the results. In general, truncation is important to reduce the influence of low-quality bases at the end of the sequence reads. The truncated length for forward and reverse reads can be decided based on two factors: quality scores and amplicon length. However, there is a trade-off between read quality and read length for efficient merging. In this study, we performed the truncation step with different combinations of truncated length for forward and reverse reads for the ZIEL-I mock community for the V4 region (primer pair 515F-806R). Different ranges of forward (250 to 280 bp) and reverse (180 to 250 bp) read lengths were selected based on the quality (*q*) score (≥20) and amplicon length. We found that changes in the forward and reverse truncated lengths directly influence the percentages of sequence counts retained after that step (Fig. 6A). For instance, when the forward read length is set to 250 bp and the reverse read length to be 180 bp, 90% of the input reads were retained. The percentage of retained reads gradually decreased from 90% to 68% when increasing the reverse read length. The same trend was observed for forward 260-bp and reverse truncated length combinations (180 to 250 bp). However, using a forward read length of 270 bp or 280 bp combined with a reverse read length between 180 and 250 bp resulted in a lower percentage of retained reads, ranging from 85% to 65%. The lower number and, thus, reduced percentage of retained reads are mostly due to a decreased number of reads passing the filter. Subsequently, only this decreased number of reads was processed during denoising and merging steps (see Table S3).

The association between the percentage of reads retained and the number of ASVs obtained after those processing steps was also evaluated. The slight differences in the retained percentage of reads for different truncated length combinations did not drastically affect the number of features obtained. The total number of ASVs varied from 10 to 20 for different combinations of truncated lengths for the ZIEL-I mock community. Using truncated lengths of 250 bp and 180 bp for forward and reverse reads, respectively, resulted in 20 ASVs, while other length combinations obtained only 10 to 15 ASVs (Fig. 6B).

To check whether the observed differences in detected ASVs (e.g., 10 versus 20) arose from contaminated reads not corresponding to bacteria included in the ZIEL-I mock community, we performed a local BLAST search. We checked the reads produced by different forward and reverse read combinations against the reference sequence and used a cutoff of ≥97% identity, ≥90% coverage, and E value of ≤0.00001. BLAST results of each forward and reverse read combination showed that 91 to 100% of the ASVs were mapped against the reference sequence of the mock community. The highest number of mismatches was found to be 1 (Fig. 6C). Only a very few nonhits, which did not reach the above-mentioned BLAST cutoffs, were obtained. Nevertheless, truncation for each amplicon length should be tested since low-quality bases impair read clustering.

Mock communities will, irrespective of the number of species added, never fully reflect complex microbial communities. Thus, we analyzed whether truncation showed an impact on a complex microbial community similar to that for the mock community used before. To this end, we used the previously analyzed 33 human stool samples as the test set. We found that the percentage of reads retained after truncation showed lower variations than for the mock community. The largest number of reads retained was identified for setting 250 bp and 180 bp for forward and reverse reads, respectively (Fig. 6D). Interestingly, when using 250 bp for the forward read, the percentage of retained reads decreased from 89% to 67% when increasing the reverse read length from 180 to 250 bp. Thus, insufficient removal of low-quality read sections (i.e., wrong bases) inhibits merging. The number of ASVs varied from 1,219 (250 bp forward/250 bp reverse) to 2,363 (for 270 bp/180 bp) for different combinations of truncated lengths (Fig. 6E), which led us to investigate whether different numbers of ASVs affect taxonomic assignments at genus level. Toward this end, we analyzed the number of generated ASVs for 280-bp forward reads in combination with different reverse read lengths. The number of ASVs varied from 2,057 (for 280 bp/250 bp) to 2,231 (for 280 bp/180 bp).

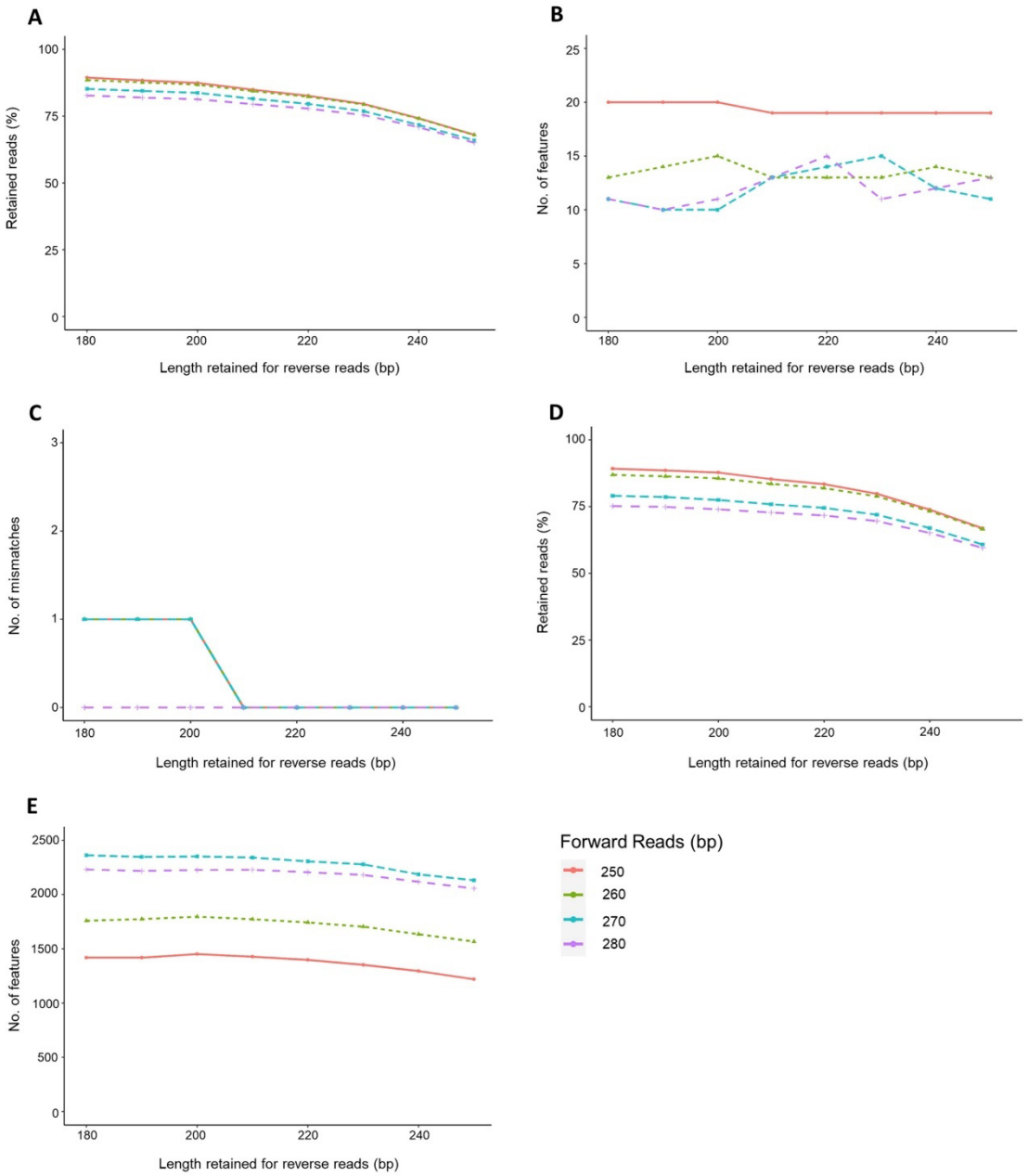

FIG 6 (A and B) The effects of different lengths of forward and reverse reads after truncation on the percentage of sequences retained after denoising (A) and number of features obtained (B) for the ZIEL-I mock community. The numbers of mismatches obtained after local BLAST search against reference sets are shown; these were used in order to test the accuracy of the ASV predictions (C). (D and E) Analysis of human data set on retained reads after denoising and truncation (D) and number of features obtained (E) for each read-length combination.

TABLE 3 Bacterial taxa at genus level influenced by primer choice and selected reference database[a]

Genera	27F-338R (V1-V2)		27F-534R (V1-V3)		341F-785R (V3-V4)		515F-806R (V4)		515F-944R (V4-V5)		939F-1378R (V6-V8)		1115F-1492R (V7-V9)	
Acetatifactor	○	○	○	○	○	○	-	-	-	-	-	-	○	○
Actinomyces	-	-	○	○	-	-	-	-	×	×	-	-	-	-
Akkermansia	×	×	○	○	○	○	-	-	×	×	-	-	-	-
Alistipes	-	-	-	-	-	-	-	-	×	×	-	-	○	○
Atopobium	-	-	-	-	-	-	-	-	-	-	-	-	○	○
Bacillus	○	○	○	○	○	+	×	+	-	-	○	○	○	○
Bacteroides	○	○	+	+	+	+	+	+	×	×	○	○	○	○
Bifidobacterium	-	-	-	-	+	+	○	○	×	×	+	+	+	+
Cellulosimicrobium	-	-	-	-	-	-	○	○	×	×	-	-	-	-
Citrobacter	○	○	○	-	-	-	×	×	×	×	×	×	×	×
***Clostridium* XVIII**	○	○	○	○	-	-	-	-	+	+	-	-	○	○
Collinsella	○	○	-	-	+	+	-	-	-	-	-	-	-	-
Eggerthella	-	-	○	○	○	+	+	+	×	×	+	+	-	-
Enterobacter	×	○	-	-	×	○	×	×	×	×	×	×	×	-
Enterococcus	○	○	○	○	○	○	○	○	○	○	○	○	+	+
Enterorhabdus	○	○	○	○	-	-	-	-	-	-	-	-	-	-
Escherichia/Shigella	○	○	○	○	-	-	○	○	-	-	○	-	×	-
Flavonifractor	-	-	-	-	-	-	-	-	×	×	-	-	-	×
Klebsiella	○	○	○	○	○	○	×	×	○	×	○	○	-	-
Lactobacillus	-	-	×	×	-	-	○	○	-	○	-	-	-	-
Listeria	+	+	○	○	○	○	○	○	○	○	○	○	○	○
Microbacterium	-	-	-	-	-	-	○	○	×	×	-	-	+	+
Oscillibacter	-	-	-	-	-	-	○	○	×	×	○	○	×	-
Parabacteroides	-	-	-	-	-	-	-	-	×	×	-	-	-	-
Prevotella	-	-	-	-	○	○	+	+	×	×	-	-	○	○
Pseudomonas	-	-	-	-	-	-	-	○	-	-	×	-	-	-
Ruminococcus	×	×	×	×	×	×	×	×	×	×	×	×	○	×
Salmonella	+	+	-	-	○	+	×	×	-	-	-	-	-	-
Staphylococcus	+	○	-	+	○	○	-	-	-	-	+	-	○	○
Database	RDP	Silva	RDP	Silva	RDP	Silva	RDP	Silva	RDP	Silva	RDP	Silva	RDP	Silva

[a]RDP (left column for each V-region) and Silva (right column for each V-region) were used as reference databases. +, <5% difference from reference (shaded green); ○, 5 to 25% difference from the reference (shaded white), −, >25% difference from the reference (shaded light brown); ×, not detected at genus level (shaded dark brown). In bold are bacterial genera present in more than one mock community; therefore, mean values were calculated for these species to estimate their performance.

The number of different genera (including unknown and unclassified entries) varied from 131 (for 280 bp/250 bp) to 143 (for 280 bp/190 and 200 bp).

Selection of primer, pipeline, parameters, and complexity of the ecosystem influences taxonomic classification. Using three different mock communities, we were able to show differences in taxonomic compositions that were due to differences in used primer pairs, reference databases, clustering methods, or specific settings. We determined a set of bacterial taxa which are biased due to primer choice as well as the reference database (Table 3). Of note, we observed that there is a strong association between the correct assignment of taxa and the complexity of the mock community. For example, *Staphylococcus* was included in all three mock communities. This species was well characterized when using the Zymo mock community but poorly represented when using the more complex mock communities ZIEL-I and ZIEL-II (Table S3). Moreover, we evaluated the influence of specific primers and their comparability in a large population-based cohort ($n = 1{,}976$ subjects). Amplicon sequencing was performed targeting the V1-V2 and the V3-V4 regions of the 16S rRNA gene (1). For the V3-V4 region, the same primer set as in this work was used. However, for V1-V2, the same primer region was used but the forward primer (27F) did not include the degenerated bases Y and M (80). This led, for example, to a complete loss of identification of *Bifidobacterium* but to an identification of *Akkermansia*. These findings strengthen our hypothesis that methodological settings influence the outcome and, thus, the results that are generated out of 16S rRNA gene sequencing data. We would like to highlight the need for transparency to increase reproducibility and comparability.

DISCUSSION

For short-amplicon 16S rRNA gene sequencing, primers spanning more than one V-region are commonly used, which enhances precision in identifying bacteria compared to a single region. Some of the most frequently used primer pairs enclose V1-V3, V3-V4, and V3-V5, which were used in large population-based cohorts, e.g., the Human Microbiome Project and others (1, 33, 34). Nevertheless, each different primer pair or V-region used will cause bias in the data. In addition, sampling and sample storage, sample processing (including DNA extraction and amplicon generation), sequencing analysis, and data processing introduce further bias. In the last 10 years, many of these factors were studied for a variety of ecosystems, e.g., the human gut (31, 40, 59, 68, 74, 81–84), oral and skin microbiomes (64, 85, 86), food-related ecosystems (87, 88), and environmental microbiomes such as water, marine environments, and sludge (16, 69, 72, 89–91). Nevertheless, the combination of different bias-causing factors was rarely studied. In this study, we analyzed the effects of choice of primer, reference databases, clustering method, and specific pipeline settings in combination on human stool samples and mock communities with increasing complexity using recent approaches. We wanted to highlight the contribution of each these factors to the precision of taxonomic assignment, providing the scientific community with up-to-date guidelines for experimental design and data analysis. Anticipating conclusions, each experimental setting (e.g., cohort and environment) needs to be tested up front for best performance using different experimental settings and strategies.

First, the effect of different primer pairs on the corresponding microbial profile was evaluated. Irrespective of the reference database, the primer pair 341F-785R (V3-V4) slightly outperformed the other combinations and is, therefore, a justified choice for human gut samples. This is also in accordance with Thijs et al. (71), who suggested the primer pair 341F-785R to be a good match for soil and plant-associated bacterial microbiome studies, and Rausch et al. (92), who recommended the use of the V3-V4 region over V1-V2. The sequences produced by using the primer pair 515F-944R (V4-V5) performed well when analyzing the microbiota profile of the Zymo mock community but showed poor performance on the more complex ZIEL-I and ZIEL-II mock communities, suggesting that the primer combination may not be suitable for complex microbial ecosystems at all. This highlights also the importance of including mock communities in routinely performed 16S rRNA gene analysis, as a theoretical sequence analysis by Yang et al. (14) suggested the V4-V5 region to be a good match based on its robustness in representing the full-length 16S rRNA sequences and, therefore, theoretically seemed to be a good primer pair. However, it did not perform well when real samples were used.

Obviously, mock communities do not fully reflect the complexity of a microbial community as it is seen in, e.g., human stool samples. Therefore, we included 33 human fecal samples in our analysis as well. Here, phylum-level classification is robust across the use of different primer pairs targeting different V-regions for *Bacteroidetes* (except 515F-944R), *Proteobacteria*, and *Firmicutes*. In contrast, the detection of *Actinobacteria*, *Tenericutes*, *Lentisphaerae*, and *Verrucomicrobia* varied across the use of different primer pairs, highlighting that the choice of primer should be considered carefully. Intraindividual comparison at genus level showed a high degree of variability across the different targeted regions. This was due to many unknown or unclassified taxa at genus level as well as a generally large number of different taxa. This highlights the need for ecosystem-specific reference databases (93, 94) and new bioinformatic tools that can integrate data across V-regions by taking into account region-specific bias. Here, we notice a need for large-scale studies covering multiple V-regions, which would allow for training taxonomic classifiers that can dynamically account for any region-specific bias. This would possibly be obsolete by sequencing the full-length 16S rRNA gene, although sequencing would still be influenced by the primer choice, i.e., 27F and 1492R, for nearly full-length sequencing. Full-length 16S rRNA gene sequencing is possible by using third-generation sequencing strategies (24, 26, 27) or by the generation of short reads that are later *de novo* assembled to a synthetic full-length sequence (95). Those methods seemingly offer taxonomic identification down to species or even strain level (27). Both approaches are not yet well

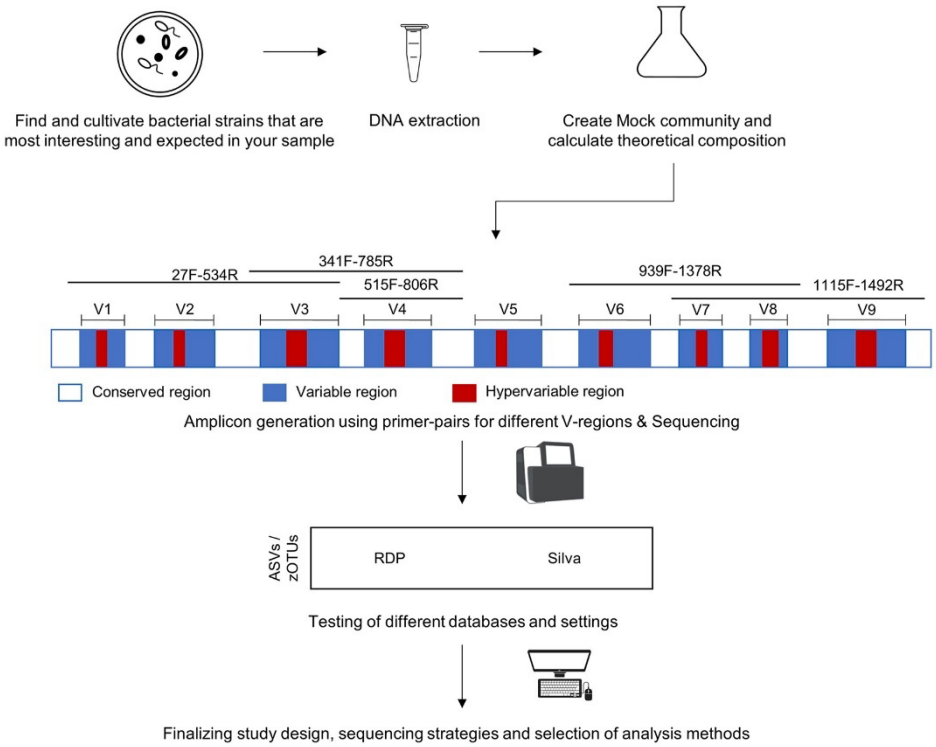

FIG 7 Recommended validation strategy before starting new microbiome studies, especially for uncommon environments. Even existing commonly used parameter combinations might be reevaluated. Thus, complex mock communities should be used and sequenced, testing a variety of different primer pairs for best performance within the environment of interest. Despite their being of minor influence, we still recommend using clustering approaches that include denoising steps (e.g., DADA2 generating ASVs) and recommend the seemingly well-curated and up-to-date databases RDP and Silva as references.

established for high-throughput sequencing and are not cost-efficient, reproducible, or easy in handling and thus need further investigation to be competitive. Further, long-read sequencing still suffers from comparably high error rates (29, 96).

It is known that the use of different bioinformatic pipelines can have an impact on the determined microbiota composition (40, 43, 65, 97). However, the influence of reference databases for taxonomic prediction was, to our knowledge, not intensively studied. In this study, we evaluated the performance of five different databases using three different mock communities. We tested the ability of each database to identify the correct taxonomy and assessed how well the known diversity of the mock samples could be captured by each database. Our finding illustrated that the Silva and RDP databases were the most accurate 16S rRNA gene databases, showing similar performances consistently superior to those of GRD, LTP, and GG in terms of true positives at the genus level. GG failed to classify *Escherichia/Shigella*, *Listeria*, *Acetatifactor*, *Bacillus*, *Clostridium*, and *Pseudomonas*, in line with the results of Park and Won (98), who found GG to be subpar compared to Silva. GG was last updated in 2013, and any usage is highly questionable.

In addition to the above, we found that quality assessment for each particular database could be conducted only when using a variety of V-regions and a sufficient complex mock community. Low-complexity mock communities using common bacteria did not reveal database issues. Thus, low-complexity mock communities might be used as positive controls in existing pipelines for general quality monitoring, but they are not recommended for detecting fundamental issues when setting up a new study,

TABLE 4 Composition of the ZIEL-I mock community[a]

Species	Amt of gDNA used (ng)	Genome size (bp)	16S rRNA gene copy no.	Theoretical abundance (%)
Actinomyces bowdenii	12	3,103,770	3	6.3
Enterorhabdus mucosicola	12	3,009,822	2	4.3
Cellulosimicrobium cellulans	12	3,850,000	3	5.1
Bacteroides sartorii	12	5,377,291	7	8.5
Alistipes sp.	12	3,734,239	2	3.5
Bacillus subtilis	12	4,215,606	9	14.0
Parabacteroides goldsteinii	12	6,751,539	7	6.8
Flavonifractor plautii	12	4,306,691	2	3.0
Clostridium ramosum	12	3,235,195	7	14.2
Enterococcus hirae	12	2,962,227	6	13.3
Acetatifactor muris	12	6,013,646	5	5.4
Staphylococcus warneri	12	2,860,455	5	11.4
Pseudomonas sp.	12	6,342,352	4	4.1

[a]Genome sizes were determined according to entries in EzBioCloud (101), and 16S rRNA gene copy number was determined according to entries in rrnDB (103).

pipeline, or laboratory. Further, concerning other body sites (or environments), specific mock communities of sufficient complexity should be used. Certainly, the addition of ubiquitous bacteria, like the skin commensal *Cutibacterium acnes* in humans and other such bacteria, should be considered.

A third factor influencing taxonomic assignment is constituted by the denoising and OTU clustering steps of data analysis. To investigate this aspect, we compared classical OTUs generated by ≥97% clustering Qiime1, ASVs generated by DADA2 denoising (48), and zOTUs generated by the USEARCH denoising algorithm (49, 99). The numbers of features identified by these clustering approaches were nearly identical across all three approaches for the tested mock community. ASV clustering performed well in the human data sets despite the increased complexity, supporting the results of previous studies (42, 100), which suggests that ASVs are the current best choice, as they showed the highest accordance with the theoretical composition of the tested mock community. However, zOTUs performed very similarly and are more robust and user-friendly concerning the input.

Specific settings, e.g., the truncation length, influence the number of reads retained for further analysis steps, as we have demonstrated. Selecting a suitable truncation

TABLE 5 Composition of the ZIEL-II mock community[a]

Species	Amt of gDNA used (ng)	Genome size (bp)	16S rRNA gene copy no.	Theoretical abundance (%)
Prevotella copri	12	3,784,859	4	4.2
Collinsella aerofaciens	12	2,463,631	5	8.0
Atopobium parvulum	12	1,543,805	1	2.6
Eggerthella lenta	12	3,500,501	3	3.4
Bifidobacterium longum	12	2,402,802	3	4.9
Clostridium ramosum	12	3,703,302	9	9.6
Staphylococcus epidermidis	12	2,520,741	5	7.8
Klebsiella pneumoniae	12	5,589,189	8	5.6
Escherichia coli LF82	12	4,881,487	7	5.6
Shigella flexneri	12	4,551,801	7	6.1
Oscillibacter valericigenes	12	4,470,622	3	2.6
Akkermansia muciniphila	12	2,760,363	3	4.3
Ruminococcus gnavus	12	3,415,781	5	5.8
Bacteroides vulgatus	12	5,063,322	7	5.4
Pseudomonas aeruginosa	12	6,612,169	4	2.4
Citrobacter freundii	12	5,300,882	8	5.9
Enterobacter cloacae	12	5,030,416	8	6.3
Listeria welshimeri	12	2,819,373	6	8.4
Microbacterium flavum	12	6,818,507	2	1.2

[a]Genome sizes were determined according to entries in EzBioCloud (101), and 16S rRNA gene copy number was determined according to entries in rrnDB (103).

length is of importance, as too-short reads have short or missing overlaps that lead to problems during merging. Conversely, too-long reads can be difficult to merge, as they show lower sequence quality. The varying number of detected ASVs for different truncation lengths is linked to the trade-off between incorporating reads of lower quality and the sensitivity for detecting low-abundance genera. By systematically reducing the reverse read length, the number of rarely observed sequences increased, as sequencing errors decrease. This highlights an important role for this parameter in the reproducibility of analysis results. To assess this potential bias, we suggest using sufficiently complex mock communities of known composition to determine suitable truncation lengths. Further, it is important to report this parameter (as well as all others) with respect to reproducibility of analysis results.

In summary, our results across 3 mock communities and 33 human samples suggest using primers for the V3-V4 region, which show good overall performance for human gut samples. As a reference database, we recommend using either Silva or RDP. Even though only minor differences were observed between clustering methods, we currently recommend using ASVs or zOTUs, with negligible difference between the two. Regarding pipeline settings, we suggest that truncated length combinations should be tested for the primer pairs used in each study. For example, we would suggest for V4 reads truncated to 250 bp and 180 bp for forward and reverse, respectively. However, the last settings depend on the amplicon lengths of the V-regions. To guarantee comparable and reliable results, we recommend creating specific (i.e., reflecting the targeted microbial environment) and sufficiently complex mock community to test whether the study design and the analysis pipelines will be suitable for the bacterial community of interest or type of sample desired (Fig. 7).

MATERIALS AND METHODS

Preparation of human gut samples. Stool samples were obtained from healthy volunteers (33 subjects) and collected in stool sample tubes (Sarstedt AG & Co.). Tubes had been prefilled under a clean bench with 8 ml of stabilizing buffer (1,400 ml of Milli-Q water supplemented with 60 ml of 0.5 M EDTA, 37.5 ml of 1 M sodium citrate, and 1.05 kg of ammonium sulfate [pH 5.2] and sterile filtered using a 0.2-μm filter). A stainless steel mixing bead of 5.5 mm (MP Biomedicals) was added to facilitate homogenization of the crude stool in the stabilizing fluid. The stool was directly resuspended by shaking and vortexing. All samples were aliquoted (in 600-μl portions) and stored at −80°C until DNA extraction.

Preparation of mock communities. A mock community is a defined *in vitro*-created mixture of microbial cells. For validation, three different mock communities were used, (i) the ZymoBIOMICS microbial community DNA standard (Zymo Research; catalog no. D6306) with 8 bacterial species, (ii) a more complex in-house mock community (ZIEL-I) including 13 different bacterial species (Table 4), and (iii) another in-house mock community (ZIEL-II) with even more increased complexity including 19 different bacterial species (Table 5). For the in-house mock communities, common gut-related bacterial species were used. The mock community ZIEL-II included such species, which seemed to be influenced by targeted V-region in preliminary results (data not shown). Bacteria were cultured as described in Table S1 and harvested after 2 to 3 days by centrifugation. Pellets were resuspended in stabilizing buffer and stored at −80°C until further processing. After genomic DNA (gDNA) extraction was performed for each strain separately (see below), strain identities were verified by Sanger sequencing. Afterwards, mock communities were constructed by pooling 12 ng of bacterial gDNA per strain. The theoretical composition was calculated according to the formula described for the Zymo mock community by Zymo Research: 16S rRNA gene copy number = total genomic DNA (g) × unit conversion constant (bp/g)/genome size (bp) × 16S rRNA gene copy number per genome. Genome sizes were determined by the 16S reference database EzBioCloud (101). If the genome size for the species included was not available in the database, the closest relative (based on 16S rRNA gene identity) was used for genome size estimation instead. In cases in which only the genus of the bacterium used in the mock community is known, mean genome sizes including all species listed in the database of the genus were used. The 16S rRNA gene copy number was determined from rrnDB (102, 103) as a reference database, also using the closest relative as a surrogate or using mean values of 16S rRNA gene numbers if specific values were not available. Overall, the three different mock communities were sequenced (see below) in duplicates (ZIEL-I) or triplicates (Zymo and ZIEL-II). For further analyses (see below), we used the mean values of the taxonomic compositions of the replicates (all replicates are shown in Fig. S3 to S5).

Extraction of gDNA. gDNA was isolated with a modified protocol of Godon et al. (104) as described previously (105). Briefly, either 600 μl of pure bacterial culture or 600 μl of frozen stool samples (i.e., bacteria in stabilizer fluid) was thawed on ice and vortexed. Samples were transferred into a 2-ml bead-beating tube (MP Biomedicals), and 250 μl of 4 M guanidinium thiocyanate and 500 μl of 5% sodium *N*-lauroylsarcosine were added. The mixture was incubated at 70°C for 60 min with shaking (700 rpm). Next, cells were disrupted by bead-beating using a FastPrep24 instrument (MP Biomedicals). Bead-beating was conducted three times for 40 s at 6.5 m/s, with cooling with dry ice. Processed samples were stored

on ice. Subsequently, 15 mg of polyvinylpolypyrrolidone was added to each sample, with brief mixing. Samples were centrifuged for 3 min at 15,000 × *g* and 4°C, and the supernatant was transferred into a fresh 2-ml sample tube. To every sample, 5 μl of RNase A (10 mg/ml) was added and samples were incubated for 20 min at 37°C with moderate shaking (700 rpm). DNA was purified using gDNA columns (Macherey-Nagel) following the manufacturer's instructions. Finally, gDNA was eluted in 100 μl of elution buffer provided in the kit. Concentrations and purity were checked using the NanoDrop system (Thermo Scientific), and samples were stored at 4°C (up to 5 days) or at −20°C thereafter.

Primer selection and *in silico* testing. Primers for commonly used V-regions were chosen after a literature survey. *In silico* tests of primer specificity were conducted using Silva TestPrime 1.0 (http://www.arb-silva.de/search/testprime/) using standard settings with zero mismatches.

Library preparation of different variable regions of the 16S rRNA gene. For amplification of the variable regions (Fig. S1) and addition of adapter binding sites for sequencing, a 1st-step PCR was performed in a 50-μl total volume. Each reaction mixture contained 24 ng of gDNA, 1× Phusion HF buffer, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.125 μM each forward and reverse primer, 7.5% dimethyl sulfoxide (DMSO), and 0.25 μl of Phusion HF II DNA polymerase (Thermo Fisher). PCR was performed as follows: 98°C for 40 s, 15 cycles of 98°C for 20 s, the V-region specific annealing temperature (Table 1) for 40 s, and 72°C for 40 s, and a final extension step at 72°C for 2 min.

Barcodes enabling multiplexing were added in the 2nd-step PCR. For this, a 100-μl PCR mixture was prepared using 10 μl of the 1st-step PCR product, 1× Phusion HF buffer, 0.2 mM dNTPs, 0.125 μM each forward and reverse barcode primer, 0.25% DMSO, and 0.5 μl of Phusion HF II DNA polymerase. PCR conditions were 98°C for 40 s, 10 cycles of 98°C for 20 s, 55°C for 40 s, and 72°C for 40 s, and a final extension step at 72°C for 2 min. Further details and work time estimations are found in the work of Reitmeier et al. (105).

Library quality check and sequencing. For validation and quality assurance, 8 μl of the 2nd-step PCR product was loaded onto a 1.5% agarose gel. The remaining 92 μl of the 2nd-step PCR product was purified with AMPure XP beads using a ratio of 1.8 times (i.e., addition of 180 μl of beads to 100 μl of sample). Concentrations of the final PCR products were measured in triplicates using a Qubit (Thermo Fisher). Each sample was adjusted to 0.5 nM, and all samples were pooled and sequenced in paired-end modus for 2 × 300 bp (PE300) using a MiSeq system (Illumina, Inc.) following the manufacturer's instructions. The final DNA concentration of the library was 12 pM, and 15% (vol/vol) PhiX was added.

Primer-specific feature classifiers. User-generated feature classifiers accounting for unique characteristics introduced by sample preparation, sequencing primer, and read length perform generally better than the naive classifiers trained on full-length sequences (106). In order to improve the taxonomic classification, five different databases were used to generate primer-specific feature classifiers, namely, GreenGenes (GG) (51), the Ribosomal Database Project (RDP) (52), Silva (53), the genomic-based 16S rRNA Database (GRD) (54), and The All-Species Living Tree (LTP) database (55). Feature classifiers were built for each V-region or primer pair using the *q2-feature-classifier* (107), which is a naive Bayes taxonomic classifier implemented in Qiime2-2019.10 (47).

OTU clustering using Qiime1. We consider Qiime-UCLUST (108) a popular example of an OTU-generating method as well as the recently proposed USEARCH-UNOISE3 (49, 99) (described below). Qiime-UCLUST clusters sequence reads at ≥97% sequence identity. UCLUST clustering was performed in Qiime1 as follows. Forward and reverse primer sequences and the low-quality reads ($q \leq 20$) of demultiplexed paired-end reads were removed by cutadapt 2.10 (109). The trimmed reads were joined by *multiple_join_paired_ends.py* to create a single fasta file of all samples using *multiple_split_libraries_fastq.py*. OTU abundance tables were generated using the UCLUST clustering method through the script *pick_de_novo_otus.py* script in Qiime1. OTU mapping files along with representative sequences, alignment of sequences, and taxonomic alignment files were generated during the *de novo* clustering steps. The RDP database was used as a reference database for defining OTUs at ≥97% sequence similarity.

zOTU generation using UNOISE. USEARCH-UNOISE3 aims to reconstruct exact biological sequences from the samples into zOTUs. Paired-end raw reads were merged using the *fastq_mergepairs* script of USEARCH version 11 (108), and the primer sequences were removed using the *fastx_truncate* script. Merging and primer removal steps were conducted before quality filtering, as primer removal reduces the expected errors and merging before quality filtering improves the base call error estimates captured in the overlapping regions as suggested by the author of USEARCH/UPARSE (110). Processed reads were deduplicated and *de novo* clustered into zOTUs. RDP database (project release 11) was used for taxonomic assignment of the representative zOTU sequences.

ASV generation using nf-core/ampliseq pipeline. The three mock communities and human data sets were analyzed using the nfcore/ampliseq nextflow pipeline (111, 112). nfcore/ampliseq is a Qiime2-based end-to-end solution for processing 16S rRNA gene amplicon sequencing data. The quality of raw sequencing reads was assessed by FastQC (113). Primer sequences and bases with low-quality scores were trimmed using cutadapt (109). The DADA2 (48) package wrapped inside the nf-core/ampliseq pipeline was used for denoising and constructing ASVs. Based on the quality profile and amplicon length, truncated lengths for forward (250 to 280 bp) and reverse reads (180 to 260 bp) were used in the DADA2 denoising steps to study the relationships between the truncated lengths and number of ASVs generated.

Data visualization using Rhea. Data visualization was performed with the R-based pipeline Rhea (114), a collection of R-scripts for 16S rRNA gene sequencing data analysis. After normalization, alpha-diversity and beta-diversity were determined and visualized. Taxonomic classification was conducted down to genus level.

Data visualization for human samples. To determine differences of the microbiota composition by targeting different V-regions, a multivariate analysis was performed using the *vegan* R-package. Therefore, a Bray-Curtis distance between samples was calculated based for relative abundance values

on phylum and genus levels and grouped according to targeted V-region. First, two dimensions of the nonmetric MDS (NMDS) plot were visualized by using *ggplot2*, and data points were labeled according to targeted V-region.

Data availability. Raw sequencing data are available at the Sequence Read Archive under the accession number PRJNA674596.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.1 MB.

FIG S2, PDF file, 0.7 MB.

FIG S3, PDF file, 0.1 MB.

TABLE S1, PDF file, 0.02 MB.

TABLE S2, PDF file, 0.02 MB.

TABLE S3, PDF file, 0.1 MB.

TABLE S4, PDF file, 0.1 MB.

TABLE S5, PDF file, 0.2 MB.

ACKNOWLEDGMENTS

We thank Thomas Clavel, Theresa Streidl, and David Wylensek (Research Group Functional Microbiome, RWTH Aachen), Annemarie Siebert and Michaela Kreitmeier (Chair of Microbial Ecology, TUM), and Nico Gebhardt (Chair of Nutrition & Immunology, TUM) for providing bacterial strains. Further, we thank Annika Naumann, Andrea Isabel Proaño Vasco, and Caroline Ziegler for excellent technical assistance.

I.A.-S. was funded by the ZIEL—Institute for Food & Health with a grant for a doctorate position and partially funded by a grant of the Research Foundation of Dairy Science at the Technical University of Munich (VFMF), both given to K.N. J.B. was partially funded by VILLUM Young Investigator Grant no. 13154. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation)—Projektnummer 395357507—SFB 1371.

REFERENCES

1. Reitmeier S, Kiessling S, Clavel T, List M, Almeida EL, Ghosh TS, Neuhaus K, Grallert H, Linseisen J, Skurk T, Brandl B, Breuninger TA, Troll M, Rathmann W, Linkohr B, Hauner H, Laudes M, Franke A, Le Roy CI, Bell JT, Spector T, Baumbach J, O'Toole PW, Peters A, Haller D. 2020. Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. Cell Host Microbe 28:258–272.e6. https://doi.org/10.1016/j.chom.2020.06.004.
2. Hamady M, Knight R. 2009. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19:1141–1152. https://doi.org/10.1101/gr.085464.108.
3. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. 2012. Next-generation sequencing technologies for environmental DNA research. Mol Ecol 21:1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x.
4. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun 8:1784. https://doi.org/10.1038/s41467-017-01973-8.
5. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. 2016. Genetic determinants of the gut microbiome in UK twins. Cell Host Microbe 19:731–743. https://doi.org/10.1016/j.chom.2016.04.017.
6. Hiergeist A, Reischl U, Gessner A, Priority Program 1656 Intestinal Microbiota Consortium/ quality assessment participants. 2016. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. Int J Med Microbiol 306:334–342. https://doi.org/10.1016/j.ijmm.2016.03.005.
7. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One 9:e94249. https://doi.org/10.1371/journal.pone.0094249.
8. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87:4576–4579. https://doi.org/10.1073/pnas.87.12.4576.
9. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761–2764. https://doi.org/10.1128/JCM.01228-07.
10. Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555. https://doi.org/10.1016/j.mimet.2003.08.009.
11. Martinez-Porchas M, Villalpando-Canchola E, Ortiz Suarez LE, Vargas-Albores F. 2017. How conserved are the conserved 16S-rRNA regions? PeerJ 5:e3036. https://doi.org/10.7717/peerj.3036.
12. Fischer MA, Güllert S, Neulinger SC, Streit WR, Schmitz RA. 2016. Evaluation of 16S rRNA gene primer pairs for monitoring microbial community structures showed high reproducibility within and low comparability between datasets generated with multiple archaeal and bacterial primer pairs. Front Microbiol 7:1297. https://doi.org/10.3389/fmicb.2016.01297.
13. Martínez-Porchas M, Vargas-Albores F. 2017. An efficient strategy using k-mers to analyse 16S rRNA sequences. Heliyon 3:e00370. https://doi.org/10.1016/j.heliyon.2017.e00370.
14. Yang B, Wang Y, Qian P-Y. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17:135. https://doi.org/10.1186/s12859-016-0992-y.
15. Pausan MR, Csorba C, Singer G, Till H, Schöpf V, Santigli E, Klug B, Högenauer C, Blohs M, Moissl-Eichinger C. 2019. Exploring the archaeome: detection of archaeal signatures in the human body. Front Microbiol 10:2796. https://doi.org/10.3389/fmicb.2019.02796.
16. Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. 2018. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. Environ Microbiol Rep 11:487–494. https://doi.org/10.1111/1758-2229.12684.
17. Berry D, Ben Mahfoudh K, Wagner M, Loy A. 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl Environ Microbiol 77:7846–7849. https://doi.org/10.1128/AEM.05220-11.
18. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res 35:e120. https://doi.org/10.1093/nar/gkm541.
19. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S,

Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA. 2008. The potential and challenges of nanopore sequencing. Nat Biotechnol 26:1146–1153. https://doi.org/10.1038/nbt.1495.

20. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133–138. https://doi.org/10.1126/science.1162986.
21. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, Iwai T, Imanishi T, Hirota K. 2019. Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION nanopore sequencer. FEBS Open Bio 9:548–557. https://doi.org/10.1002/2211-5463.12590.
22. Curren E, Yoshida T, Kuwahara VS, Leong SCY. 2019. Rapid profiling of tropical marine cyanobacterial communities. Reg Stud Mar Sci 25:100485. https://doi.org/10.1016/j.rsma.2018.100485.
23. Benítez-Páez A, Portune KJ, Sanz Y. 2016. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. Gigascience 5:4. https://doi.org/10.1186/s13742-016-0111-z.
24. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. 2019. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the *rrn* operon [version 2; peer review: 2 approved, 3 approved with reservations]. F1000Res 7:1755. https://doi.org/10.12688/f1000research.16817.2.
25. Martijn J, Lind AE, Schön ME, Spiertz I, Juzokaite L, Bunikis I, Pettersson OV, Ettema TJG. 2019. Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. Environ Microbiol 21:2485–2498. https://doi.org/10.1111/1462-2920.14636.
26. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 47:e103. https://doi.org/10.1093/nar/gkz569.
27. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029. https://doi.org/10.1038/s41467-019-13036-1.
28. Rhoads A, Au KF. 2015. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289. https://doi.org/10.1016/j.gpb.2015.08.002.
29. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY, Popitsch N, Ip CLC, Roberts HE, Salatino S, Lockstone H, Lunter G, Taylor JC, Buck D, Simpson MA, Donnelly P. 2019. Sequencing of human genomes with nanopore technology. Nat Commun 10:1869. https://doi.org/10.1038/s41467-019-09637-5.
30. Hoffmann C, Hill DA, Minkah N, Kirn T, Troy A, Artis D, Bushman F. 2009. Community-wide response of the gut microbiota to enteropathogenic *Citrobacter rodentium* infection revealed by deep sequencing. Infect Immun 77:4668–4678. https://doi.org/10.1128/IAI.00493-09.
31. Alcon-Giner C, Caim S, Mitra S, Ketskemety J, Wegmann U, Wain J, Belteki G, Clarke P, Hall LJ. 2017. Optimisation of 16S rRNA gut microbiota profiling of extremely low birth weight infants. BMC Genomics 18:841. https://doi.org/10.1186/s12864-017-4229-x.
32. Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, Brodie EL, Malamud D, Poles MA, Pei Z. 2010. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol 16:4135–4144. https://doi.org/10.3748/wjg.v16.i33.4135.
33. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. 2012. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. PLoS One 7:e39315. https://doi.org/10.1371/journal.pone.0039315.
34. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi VR, Brooks P, Buck GA, Buhay CJ, The Human Microbiome Project Consortium, et al. 2012. A framework for human microbiome research. Nature 486:215–221.
35. Claesson MJ, O'Sullivan O, Wang Q, Nikkila J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW. 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One 4:e6669. https://doi.org/10.1371/journal.pone.0006669.
36. Ghyselinck J, Pfeiffer S, Heylen K, Sessitsch A, De Vos P. 2013. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. PLoS One 8:e71360. https://doi.org/10.1371/journal.pone.0071360.
37. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. 2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data 6:190007. https://doi.org/10.1038/sdata.2019.7.
38. Barb JJ, Oler AJ, Kim H-S, Chalmers N, Wallen GR, Cashion A, Munson PJ, Ames NJ. 2016. Development of an analysis pipeline characterizing multiple hypervariable regions of 16S rRNA using mock samples. PLoS One 11:e0148047. https://doi.org/10.1371/journal.pone.0148047.
39. Pinna NK, Dutta A, Monzoorul Haque M, Mande SS. 2019. Can targeting non-contiguous V-regions with paired-end sequencing improve 16S rRNA-based taxonomic resolution of microbiomes?: an *in silico* evaluation. Front Genet 10:653. https://doi.org/10.3389/fgene.2019.00653.
40. Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. 2015. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. J Proteomics Bioinform 8:12. https://doi.org/10.4172/jpb.1000381.
41. Marizzoni M, Gurry T, Provasi S, Greub G, Lopizzo N, Ribaldi F, Festari C, Mazzelli M, Mombelli E, Salvatore M, Mirabelli P, Franzese M, Soricelli A, Frisoni GB, Cattaneo A. 2020. Comparison of bioinformatics pipelines and operating systems for the analyses of 16S rRNA gene amplicon sequences in human fecal samples. Front Microbiol 11:1262. https://doi.org/10.3389/fmicb.2020.01262.
42. Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One 15:e0227434. https://doi.org/10.1371/journal.pone.0227434.
43. Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, Corby P, Guo Y, Ruff RR, Alekseyenko AV, Li X, Saxena D. 2020. The influences of bioinformatics tools and reference databases in analyzing the human oral microbial community. Genes 11:878. https://doi.org/10.3390/genes11080878.
44. Balvočiūtė M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? BMC Genomics 18:114. https://doi.org/10.1186/s12864-017-3501-4.
45. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09.
46. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth.f.303.
47. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9.
48. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869.

49. Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375. https://doi.org/10.1093/bioinformatics/bty113.
50. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639–2643. https://doi.org/10.1038/ismej.2017.119.
51. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072. https://doi.org/10.1128/AEM.03006-05.
52. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642. https://doi.org/10.1093/nar/gkt1244.
53. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219.
54. Laboratory for Integrated Bioinformatics, Center for Integrative Medical Sciences. 2015. GRD—Genomic-based 16S ribosomal RNA database, Riken (Japan). https://metasystems.riken.jp/grd/.
55. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, Ludwig W, Glöckner FO, Rosselló-Móra R. 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 31:241–250. https://doi.org/10.1016/j.syapm.2008.07.001.
56. Bailén M, Bressa C, Larrosa M, González-Soltero R. 2020. Bioinformatic strategies to address limitations of 16rRNA short-read amplicons from different sequencing platforms. J Microbiol Methods 169:105811. https://doi.org/10.1016/j.mimet.2019.105811.
57. Escobar-Zepeda A, Godoy-Lozano EE, Raggi L, Segovia L, Merino E, Gutiérrez-Rios RM, Juarez K, Licea-Navarro AF, Pardo-Lopez L, Sanchez-Flores A. 2018. Analysis of sequencing strategies and tools for taxonomic annotation: defining standards for progressive metagenomics. Sci Rep 8:12034. https://doi.org/10.1038/s41598-018-30515-5.
58. Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF, Dutton RJ, Turnbaugh PJ, Knight R, Caporaso JG. 2016. mockrobiota: a public resource for microbiome bioinformatics benchmarking. mSystems 1:e00062-16. https://doi.org/10.1128/mSystems.00062-16.
59. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. 2015. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. PLoS One 10:e0134802. https://doi.org/10.1371/journal.pone.0134802.
60. Sinha R, Chen J, Amir A, Vogtmann E, Shi J, Inman KS, Flores R, Sampson J, Knight R, Chia N. 2016. Collecting fecal samples for microbiome analyses in epidemiology studies. Cancer Epidemiol Biomarkers Prev 25:407–416. https://doi.org/10.1158/1055-9965.EPI-15-0951.
61. Burz SD, Abraham AL, Fonseca F, David O, Chapron A, Béguet-Crespel F, Cénard S, Le Roux K, Patrascu O, Levenez F, Schwintner C, Blottière HM, Béra-Maillet C, Lepage P, Doré J, Juste C. 2019. A guide for *ex vivo* handling and storage of stool samples intended for fecal microbiota transplantation. Sci Rep 9:8897. https://doi.org/10.1038/s41598-019-45173-4.
62. Choo JM, Leong LE, Rogers GB. 2015. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5:16350. https://doi.org/10.1038/srep16350.
63. Fouhy F, Deane J, Rea MC, O'Sullivan O, Ross RP, O'Callaghan G, Plant BJ, Stanton C. 2015. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. PLoS One 10:e0119355. https://doi.org/10.1371/journal.pone.0119355.
64. Teng F, Darveekaran Nair SS, Zhu P, Li S, Huang S, Li X, Xu J, Yang F. 2018. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. Sci Rep 8:16321. https://doi.org/10.1038/s41598-018-34294-x.
65. Ducarmon QR, Hornung BVH, Geelen AR, Kuijper EJ, Zwittink RD. 2020. Toward standards in clinical microbiota studies: comparison of three DNA extraction methods and two bioinformatic pipelines. mSystems 5:e00547-19. https://doi.org/10.1128/mSystems.00547-19.
66. Gryp T, Glorieux G, Joossens M, Vaneechoutte M. 2020. Comparison of five assays for DNA extraction from bacterial cells in human faecal samples. J Appl Microbiol 129:378–388. https://doi.org/10.1111/jam.14608.
67. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Abnet CC, The Microbiome Quality Control Project Consortium, Knight R, White O, Huttenhower C. 2017. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol 35:1077–1086. https://doi.org/10.1038/nbt.3981.
68. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung F-E, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M, Brown JRM, Carton T, Cools-Portier S, Daigneault M, Derrien M, Druesne A, de Vos WM, Finlay BB, Flint HJ, Guarner F, Hattori M, Heilig H, Luna RA, van Hylckama Vlieg J, Junick J, Klymiuk I, Langella P, Le Chatelier E, Mai V, Manichanh C, Martin JC, Mery C, Morita H, O'Toole PW, Orvain C, Patil KR, Penders J, Persson S, Pons N, Popova M, Salonen A, Saulnier D, Scott KP, Singh B, Slezak K, Veiga P, Versalovic J, Zhao L, Zoetendal EG, Ehrlich SD, Dore J, Bork P. 2017. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 35:1069–1076. https://doi.org/10.1038/nbt.3960.
69. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414. https://doi.org/10.1111/1462-2920.13023.
70. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41:e1. https://doi.org/10.1093/nar/gks808.
71. Thijs S, Op De Beeck M, Beckers B, Truyens S, Stevens V, Van Hamme JD, Weyens N, Vangronsveld J. 2017. Comparative evaluation of four bacteria-specific primer pairs for 16S rRNA gene surveys. Front Microbiol 8:494. https://doi.org/10.3389/fmicb.2017.00494.
72. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. 2015. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771. https://doi.org/10.3389/fmicb.2015.00771.
73. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, Knights D, Beckman KB. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol 34:942–949. https://doi.org/10.1038/nbt.3601.
74. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT, Yooseph S, Biggs W, Nelson KE, Venter JC. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci U S A 112:14024–14029. https://doi.org/10.1073/pnas.1519288112.
75. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17:55. https://doi.org/10.1186/s12864-015-2194-9.
76. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112–5120. https://doi.org/10.1128/AEM.01043-13.
77. Yeh Y-C, Needham DM, Sieradzki ET, Fuhrman JA. 2018. Taxon disappearance from microbiome analysis reinforces the value of mock communities as a standard in every sequencing run. mSystems 3:e00023-18. https://doi.org/10.1128/mSystems.00023-18.
78. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, Braun J, McWeeney S. 2019. Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. mSystems 4:e00290-19. https://doi.org/10.1128/mSystems.00290-19.
79. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. 2015. The microbiome quality control project: baseline study design and future directions. Genome Biol 16:276. https://doi.org/10.1186/s13059-015-0841-8.
80. Thaiss Christoph A, Zeevi D, Levy M, Zilberman-Schapira G, Suez J, Tengeler Anouk C, Abramson L, Katz Meirav N, Korem T, Zmora N, Kuperman Y, Biton I, Gilad S, Harmelin A, Shapiro H, Halpern Z, Segal E, Elinav E. 2014. Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. Cell 159:514–529. https://doi.org/10.1016/j.cell.2014.09.048.
81. Bellali S, Lagier JC, Raoult D, Bou Khalil J. 2019. Among live and dead bacteria, the optimization of sample collection and processing remains essential in recovering gut microbiota components. Front Microbiol 10:1606. https://doi.org/10.3389/fmicb.2019.01606.
82. Ma J, Sheng L, Hong Y, Xi C, Gu Y, Zheng N, Li M, Chen L, Wu G, Li Y, Yan J, Han R, Li B, Qiu H, Zhong J, Jia W, Li H. 2020. Variations of gut microbiome profile under different storage conditions and preservation periods: a multi-dimensional evaluation. Front Microbiol 11:972. https://doi.org/10.3389/fmicb.2020.00972.

83. Penington JS, Penno MAS, Ngui KM, Ajami NJ, Roth-Schulze AJ, Wilcox SA, Bandala-Sanchez E, Wentworth JM, Barry SC, Brown CY, Couper JJ, Petrosino JF, Papenfuss AT, Harrison LC, ENDIA Study Group. 2018. Influence of fecal collection conditions and 16S rRNA gene sequencing at two centers on human gut microbiota analysis. Sci Rep 8:4386. https://doi.org/10.1038/s41598-018-22491-7.
84. Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome 3:26. https://doi.org/10.1186/s40168-015-0087-4.
85. Bjerre RD, Hugerth LW, Boulund F, Seifert M, Johansen JD, Engstrand L. 2019. Effects of sampling strategy and DNA extraction on human skin microbiome investigations. Sci Rep 9:17287. https://doi.org/10.1038/s41598-019-53599-z.
86. Meisel JS, Hannigan GD, Tyldsley AS, SanMiguel AJ, Hodkinson BP, Zheng Q, Grice EA. 2016. Skin microbiome surveys are strongly influenced by experimental design. J Invest Dermatol 136:947–956. https://doi.org/10.1016/j.jid.2016.01.016.
87. De Filippis F, Parente E, Zotta T, Ercolini D. 2018. A comparison of bioinformatic approaches for 16S rRNA gene profiling of food bacterial microbiota. Int J Food Microbiol 265:9–17. https://doi.org/10.1016/j.ijfoodmicro.2017.10.028.
88. Xue Z, Kable ME, Marco ML. 2018. Impact of DNA sequencing and analysis methods on 16S rRNA gene bacterial community analysis of dairy products. mSphere 3:e00410-18. https://doi.org/10.1128/mSphere.00410-18.
89. Fredriksson NJ, Hermansson M, Wilen BM. 2013. The choice of PCR primers has great impact on assessments of bacterial community diversity and dynamics in a wastewater treatment plant. PLoS One 8:e76431. https://doi.org/10.1371/journal.pone.0076431.
90. Shah M. 2014. An application of sequencing batch reactors in the identification of microbial community structure from an activated sludge. J Applied Environ Microbiol 2:176–184.
91. Brandt J, Albertsen M. 2018. Investigation of detection limits and the influence of DNA extraction and primer choice on the observed microbial communities in drinking water samples using 16S rRNA gene amplicon sequencing. Front Microbiol 9:2140. https://doi.org/10.3389/fmicb.2018.02140.
92. Rausch P, Rühlemann M, Hermes BM, Doms S, Dagan T, Dierking K, Domin H, Fraune S, von Frieling J, Hentschel U, Heinsen F-A, Höppner M, Jahn MT, Jaspers C, Kissoyan KAB, Langfeldt D, Rehman A, Reusch TBH, Roeder T, Schmitz RA, Schulenburg H, Soluch R, Sommer F, Stukenbrock E, Weiland-Bräuer N, Rosenstiel P, Franke A, Bosch T, Baines JF. 2019. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 7:133. https://doi.org/10.1186/s40168-019-0743-1.
93. Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, Albertsen M, Nielsen PH. 2020. Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax). mBio 11:e01557-20. https://doi.org/10.1128/mBio.01557-20.
94. F Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, Lemon KP. 2020. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 8:65. https://doi.org/10.1186/s40168-020-00841-w.
95. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. 2018. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol 36:190–195. https://doi.org/10.1038/nbt.4045.
96. Loit K, Adamson K, Bahram M, Puusepp R, Anslan S, Kiiker R, Drenkhan R, Tedersoo L. 2019. Relative performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens. Appl Environ Microbiol 85:e01368-19. https://doi.org/10.1128/AEM.01368-19.
97. Almeida A, Mitchell AL, Tarkowska A, Finn RD. 2018. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience 7:giy054. https://doi.org/10.1093/gigascience/giy054.
98. Park S-C, Won S. 2018. Evaluation of 16S rRNA databases for taxonomic assignments using mock community. Genomics Inform 16:e24. https://doi.org/10.5808/GI.2018.16.4.e24.
99. Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv https://doi.org/10.1101/081257:081257.
100. Caruso V, Song X, Asquith M, Karstens L. 2019. Performance of microbiome sequence inference methods in environments with varying biomass. mSystems 4:e00163-18. https://doi.org/10.1128/mSystems.00163-18.
101. Yoon S-H, Ha S-M, Kwon S, Lim J, Kim Y, Seo H, Chun J. 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 67:1613–1617. https://doi.org/10.1099/ijsem.0.001755.
102. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. 2001. rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic Acids Res 29:181–184. https://doi.org/10.1093/nar/29.1.181.
103. Roller BRK, Stoddard SF, Schmidt TM. 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat Microbiol 1:16160. https://doi.org/10.1038/nmicrobiol.2016.160.
104. Godon JJ, Zumstein E, Dabert P, Habouzit F, Moletta R. 1997. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. Appl Environ Microbiol 63:2802–2813. https://doi.org/10.1128/AEM.63.7.2802-2813.1997.
105. Reitmeier S, Kiessling S, Neuhaus K, Haller D. 2020. Comparing circadian rhythmicity in the human gut microbiome. STAR Protoc 1:100148. https://doi.org/10.1016/j.xpro.2020.100148.
106. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90. https://doi.org/10.1186/s40168-018-0470-z.
107. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267. https://doi.org/10.1128/AEM.00062-07.
108. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
109. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12.
110. Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998. https://doi.org/10.1038/nmeth.2604.
111. Peltzer A, Straub D, Patel H. 2019. nf-core/ampliseq: Ampliseq version 1.1.2. https://doi.org/10.5281/zenodo.3585924.
112. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 38:276–278. https://doi.org/10.1038/s41587-020-0439-x.
113. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
114. Lagkouvardos I, Fischer S, Kumar N, Clavel T. 2017. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ 5:e2836. https://doi.org/10.7717/peerj.2836.
115. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87. https://doi.org/10.1186/s12915-014-0087-z.
116. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108:4516–4522. https://doi.org/10.1073/pnas.1000080107.
117. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. 2018. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome 6:17. https://doi.org/10.1186/s40168-017-0396-x.
118. Lebuhn M, Hanreich A, Klocke M, Schlüter A, Bauer C, Pérez CM. 2014. Towards molecular biomarkers for biogas production from lignocellulose-rich substrates. Anaerobe 29:10–21. https://doi.org/10.1016/j.anaerobe.2014.04.006.
119. Turner S, Pryer KM, Miao VP, Palmer JD. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol 46:327–338. https://doi.org/10.1111/j.1550-7408.1999.tb04612.x.

Supplement

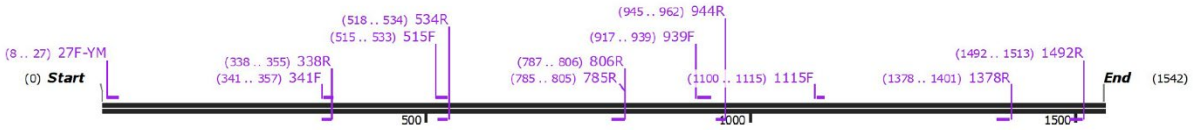

Figure S2.2.1: Positions of primers mapped onto the Escherichia coli 16S rRNA gene. Graphic designed using SnapGene software (from GSL Biotech).

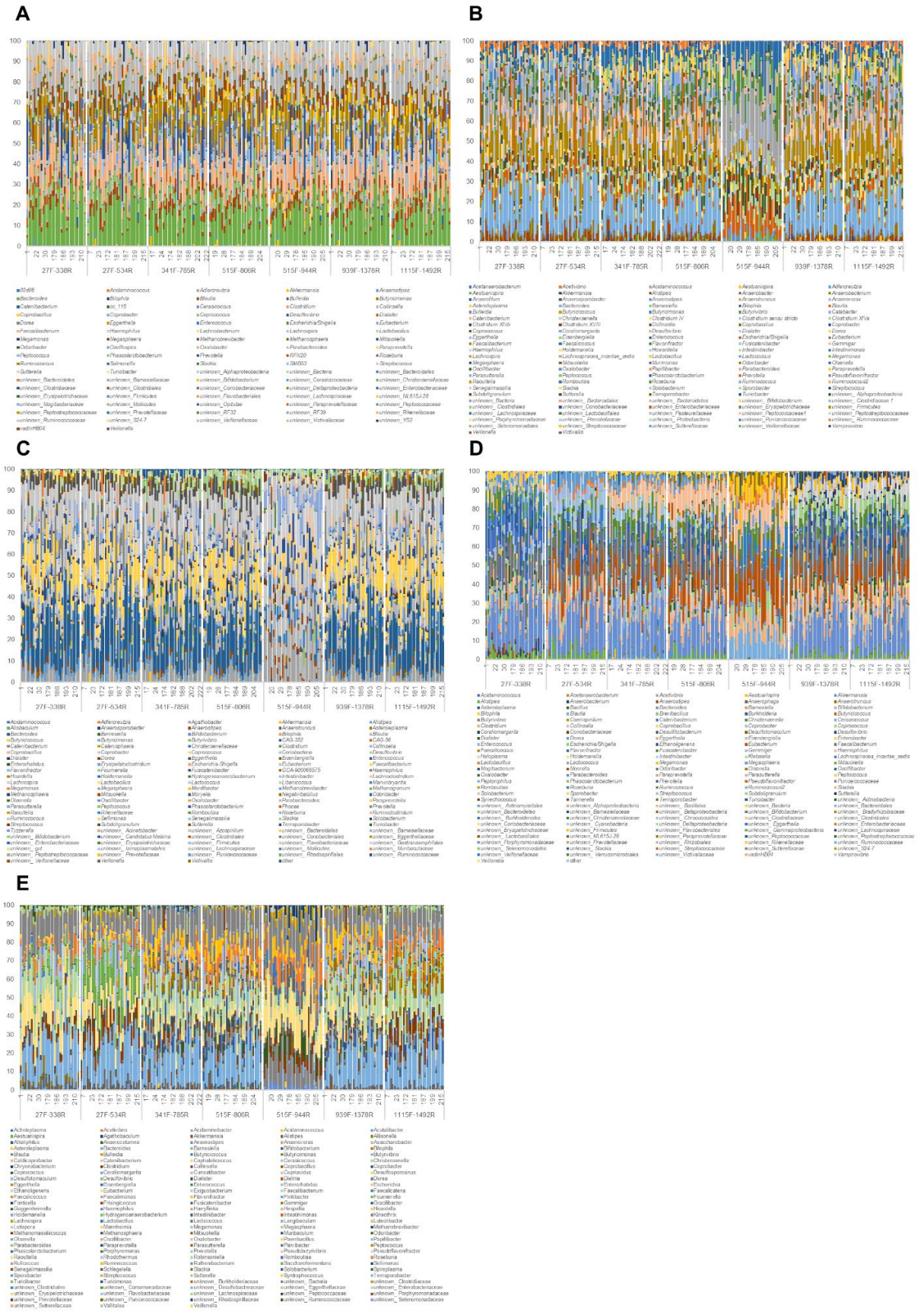

Figure S2.2.2: Human samples at genus level, using GG (**A**), RDP (**B**), Silva (**C**), GRD (**D**), and LTP (**E**) as reference databases. The primer pairs span the following V-regions: 27F-338R, V1-V2; 27F-534R, V1-V3; 341F-785R, V3-V4; 515F-806R, V4; 515F-944R, V4-V5; 939F-1378R, V6-V8; and 1115F-1492R, V7-V9.

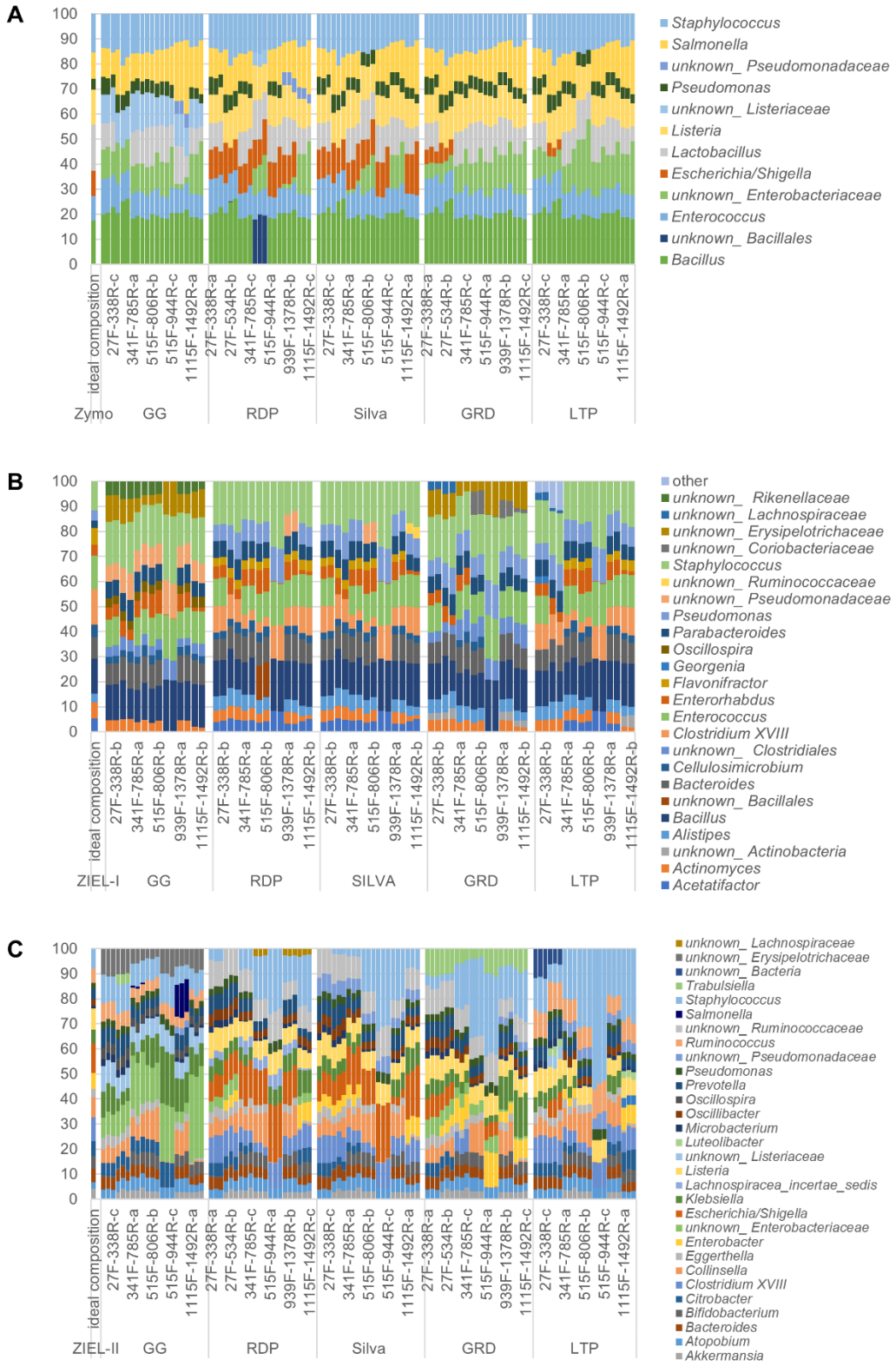

Figure S2.2.3: Comparison of Zymo (**A**), ZIEL-I (**B**), and ZIEL-II (**C**) mock sequenced over different V-regions, processed using different databases as references (GG, GreenGenes; RDP, Ribosomal Database Project; GRD, genomic-based 16S ribosomal RNA database; LTP, The All-Species Living Tree Project). The primer pairs span the following V-regions: 27F-338R, V1-V2; 27F-534R, V1-V3; 341F-785R, V3-V4; 515F-806R, V4; 515F-944R, V4-V5; 939F-1378R, V6-V8; and 1115F-1492R, V7-V9.

Table S2.2.1: Culture conditions of bacterial strains used for the ZIEL mock communities.

Mock-community	Name	Aerobe/ anaerobe	Temperature (°C)	Cultivation media
ZIEL-I-mock	*Actinomyces bowdenii*	aerobe	37	TSA/TSB
	Enterorhabdus mucosicola	anaerobe	37	WCA
	Cellulosimicrobium cellulans	anaerobe	37	WCA
	Bacteroides sartorii	anaerobe	37	WCA
	Alistipes sp.	anaerobe	37	WCA
	Bacillus subtilis	aerobe	37	TSA/TSB
	Parabacteroides goldsteinii	anaerobe	37	WCA
	Flavonifractor plautii	anaerobe	37	WCA
	Clostridium ramosum	anaerobe	37	WCA
	Enterococcus hirae	aerobe	37	TSA/TSB
	Acetatifactor muris	anaerobe	37	WCA
	Staphylococcus warneri	aerobe	37	TSA/TSB
	Pseudomonas sp.	aerobe	37	TSA/TSB
ZIEL-II- mock	*Prevotella copri*	anaerobe	37	WCA
	Collinsella aerofaciens	anaerobe	37	WCA
	Atopobium parvulum	anaerobe	37	PYG
	Eggerthella lenta	anaerobe	37	NB / WCA
	Bifidobacterium longum	anaerobe	37	WCA
	Clostridium ramosum	anaerobe	37	WCA
	Staphylococcus aureus	aerobe	37	TSA/TSB
	Klebsiella pneumoniae	aerobe	28	TSA/TSB
	Escherichia coli LF82	anaerobe	37	WCA
	Shigella flexneri	aerobe	37	NB / TSA
	Oscillibacter valericigenes	anaerobe	30	PYG
	Akkermansia muciniphila	anaerobe	37	Schaedler
	Ruminococcus gnavus	anaerobe	37	WCA
	Bacteroides vulgatus	anaerobe	37	WCA
	Pseudomonas aeruginosa	aerobe	37	TSA/TSB
	Citrobacter freundii	aerobe	37	TSA/TSB
	Enterobacter cloacae	aerobe	28	TSA/TSB
	Listeria welshimeri	aerobe	37	TSA/TSB
	Microbacterium flavum	aerobe	28	TSA/TSB

Table S2.2.2**:** In silico evaluation for used primers.

		Coverage of Kingdom	Coverage of Phyla				
V-Region	Primer	*Bacteria* [%]	*Actinobacteria* [%]	*Bacteroidetes* [%]	*Firmicutes* [%]	*Proteobacteria* [%]	*Verrucomicrobia* [%]
V1-V2	27F-338R	75.9	76	83.6	80.4	81.8	1.2
V1-V3	27F-534R	73.8	69	82.8	78.6	80.7	76.1
V3-V4	341F-785R	82.8	78.1	88.2	83.7	85.5	83.3
V4	515F-806R	82.9	78.2	86.8	84.4	86.1	76.4
V4-V5	515F-944R	48.8	3.7	41.4	60.5	66.6	3.1
V6-V8	939F-1378R	44.6	56.6	50.5	32.2	53	21.2
V7-V9	1115F-1492R	23.1	52.6	0.9	25.7	29.2	6.2

Table S2.2.3: Overview of numbers and percentages of retained reads after each processing step while generating ASVs.

ZIEL-I mock community						
Forward read length (bp)	Reverse read length (bp)	% of input passing filter	% of input de-noised	% of input merged (retained reads)	number of features (ASVs)	number of mismatches (BLAST mismatch)
250	180	90.3	90.0	89.5	20	1
250	190	89.3	89.1	88.5	20	1
250	200	88.3	88.1	87.5	20	1
250	210	85.7	85.4	85.0	19	0
250	220	83.4	83.2	82.7	19	0
250	230	80.3	80.1	79.6	19	0
250	240	74.9	74.6	74.3	19	0
250	250	68.4	68.2	68.1	19	0
260	180	89.2	89.0	88.5	13	1
260	190	88.4	88.2	87.7	14	1
260	200	87.5	87.3	86.8	15	1
260	210	85.1	84.9	84.4	13	0
260	220	82.9	82.7	82.3	13	0
260	230	80.0	79.8	79.4	13	0
260	240	74.7	74.5	74.1	14	0
260	250	68.3	68.1	68.0	13	0
270	180	87.7	87.3	85.2	11	1
270	190	87.1	86.6	84.5	10	1
270	200	86.3	85.9	83.7	10	1
270	210	84.2	83.8	81.5	13	0
270	220	82.2	81.8	79.6	14	0
270	230	79.5	79.0	76.9	15	0
270	240	74.3	73.8	71.7	12	0
270	250	68.1	67.8	66.0	11	0
280	180	85.2	84.8	82.7	11	0
280	190	84.7	84.2	81.9	10	0
280	200	84.1	83.6	81.4	11	0
280	210	82.3	81.9	79.5	13	0
280	220	80.7	80.2	77.8	15	0
280	230	78.3	77.8	75.4	11	0
280	240	73.5	73.1	70.8	12	0
280	250	67.6	67.0	65.1	13	0

Human dataset					
Forward read length (bp)	Reverse read length (bp)	% of input passing filter	% of input de-noised	% of input merged (retained reads)	number of features (ASVs)
250	180	90.6	90.0	89.2	1418
250	190	89.9	89.3	88.6	1418
250	200	89.0	88.4	87.8	1451
250	210	86.6	86.0	85.3	1427
250	220	84.8	84.2	83.4	1398
250	230	81.4	80.8	79.8	1352
250	240	75.6	75.0	73.8	1295
250	250	68.6	68.1	66.9	1219
260	180	89.2	88.2	86.9	1759
260	190	88.6	87.7	86.3	1774
260	200	87.9	86.9	85.6	1796
260	210	85.7	84.8	83.5	1773
260	220	84.1	83.1	81.9	1743
260	230	80.9	80.0	78.8	1704
260	240	75.2	74.4	73.3	1634
260	250	68.4	67.6	66.6	1567
270	180	87.3	84.5	79.0	2363
270	190	86.8	84.1	78.6	2347
270	200	86.2	83.5	77.5	2352
270	210	84.4	81.7	75.9	2341
270	220	82.9	80.2	74.5	2306
270	230	80.0	77.4	72.0	2279
270	240	74.7	72.1	67.0	2186
270	250	68.1	65.6	60.8	2132
280	180	83.9	81.1	75.2	2231
280	190	83.5	80.8	74.9	2218
280	200	83.1	80.4	74.0	2226
280	210	81.7	79.0	72.8	2228
280	220	80.6	77.9	71.7	2206
280	230	78.1	75.5	69.6	2181
280	240	73.4	70.8	65.1	2117
280	250	67.3	64.8	59.5	2057

Table S2.2.4: Comparison of difference to expected theoretical values of the ZIEL-I mock community. In green, lowest difference from ideal amount in mock.

	27F-338R (V1-V2)			27F-534R (V1-V3)			341F-785R (V3-V4)			515F-806R (V4)			515F-944R (V4-V5)			939F-1378R (V6-V8)			1115F-1492R (V7-V9)		
	ASV	zOTU	OTU	ASV	zOTU	OTU	ASV	zOTU	OTU	ASV	zOTU	OTU	ASV	zOTU	OTU	ASV	zOTU	OTU	ASV	zOTU	OTU
Acetatifactor	1.25	1.12	1.45	0.32	1.15	1.39	0.93	1.14	1.22	1.47	1.63	1.76	-2.82	-2.57	-2.05	2.18	2.32	2.19	0.73	0.71	0.83
Actinomyces	1.71	1.78	1.90	1.22	1.72	1.91	1.93	2.33	2.93	2.28	2.48	2.92	6.30	6.30	6.30	1.77	2.05	2.27	4.51	4.77	4.91
Alistipes	-2.07	-1.93	-1.80	-3.52	-2.15	-1.97	-2.56	-2.84	-3.30	-1.78	-1.90	-2.13	3.50	3.50	3.49	-1.55	-1.33	-0.98	-0.19	-0.11	-0.06
Bacillus	-0.22	-0.05	2.46	0.14	0.65	0.75	0.25	0.25	0.38	14.00	0.15	0.12	-6.61	-6.80	-6.80	-1.43	-1.35	-1.18	-3.04	-3.35	-3.25
unknown_ Bacillales	0.00	0.00	-2.33	0.00	0.00	0.00	0.00	0.00	0.00	-13.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.06	0.00	0.00	-0.02
Bacteroides	0.05	-1.05	-1.33	-1.31	-1.30	-1.18	-1.99	-2.55	-3.16	-1.43	-1.65	-2.13	8.50	8.50	8.47	-2.16	-1.90	-1.58	-0.59	0.01	0.14
Cellulosimicrobium	2.23	2.21	2.25	1.63	1.91	2.04	1.43	5.10	1.64	1.22	1.32	1.44	5.10	5.10	5.10	1.62	5.10	1.89	2.46	5.10	3.01
Clostridium XVIII	3.84	4.47	4.30	4.58	5.99	5.58	9.35	9.51	9.52	10.35	10.26	10.12	0.54	1.00	0.63	6.53	6.77	6.57	3.28	3.65	3.72
Enterococcus	2.16	2.25	2.22	13.30	3.79	3.80	2.68	2.85	2.73	2.45	2.30	2.10	-3.68	13.30	-3.95	1.84	1.60	1.18	0.43	-0.25	-0.64
Enterorhabdus	-0.63	-0.98	-0.67	-0.93	-1.34	-0.71	-2.37	-0.67	-0.04	-5.07	-4.96	-4.34	4.04	4.05	4.08	-2.39	-2.36	-1.95	2.73	3.08	3.28
Flavonifractor	3.42	3.44	3.37	3.05	3.66	3.20	2.78	2.88	2.94	3.15	3.21	3.15	6.80	6.76	6.76	3.46	3.70	3.90	2.52	3.20	3.55
Parabacteroides	-3.55	-3.20	-3.38	-4.74	-3.13	-3.13	-4.17	-4.47	-5.04	-3.86	-4.02	-4.44	3.00	3.00	2.99	-4.41	-5.81	-5.41	-2.80	-3.89	-4.12
Pseudomonas	-2.50	-2.48	-2.23	-6.23	-3.59	-3.70	-4.51	-3.68	-3.41	-3.97	-3.65	-3.48	-9.62	-8.60	-8.01	4.10	-4.09	-4.19	-4.00	-3.85	-3.74
unknown Pseudomonadaceae	0.00	0.00	-0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-8.47	0.00	0.00	0.00	0.00	0.00
Staphylococcus	-5.77	-5.60	-6.00	-7.59	-7.40	-8.07	-3.87	-6.30	-6.51	-5.09	-5.25	-5.18	-15.15	-16.40	-17.11	-1.18	-1.45	-1.76	-6.15	-7.05	-7.71
unassigned	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-3.61	0.00	0.00	0.00	0.00	0.00	-17.15	0.00	0.00	-3.31	0.00	0.00	-2.12	0.00
other	0.00	-0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.07	0.00	0.00	0.00	0.00	0.00	-0.02	0.00

Table S2.2.5: Bacterial taxa at genus level influenced by primer-choice and reference database. +, <5% difference from reference (shaded green); ○, 5 to 25% difference from the reference (shaded white); −, >25% difference from the reference (shaded light brown); ×, not detected at genus level (shaded dark brown). In yellow are bacterial genera present in more than one mock community. The primer pairs span the following V-regions: 27F-338R, V1-V2; 27F-534R, V1-V3; 341F-785R, V3-V4; 515F-806R, V4; 515F-944R, V4-V5; 939F-1378R, V6-V8; and 1115F-1492R, V7-V9.

		GG							RDP							Silva							GRD							LTP						
Origin	Genus	27F-338R	27F-534R	341F-785R	515F-806R	515F-944R	939F-1378R	1115F-1492R	27F-338R	27F-534R	341F-785R	515F-806R	515F-944R	939F-1378R	1115F-1492R	27F-338R	27F-534R	341F-785R	515F-806R	515F-944R	939F-1378R	1115F-1492R	27F-338R	27F-534R	341F-785R	515F-806R	515F-944R	939F-1378R	1115F-1492R	27F-338R	27F-534R	341F-785R	515F-806R	515F-944R	939F-1378R	1115F-1492R
ZIEL-I	*Acetatifactor*	×	×	×	×	×	×	×	○	○	○	-	-	-	○	○	○	○	-	-	-	○	×	×	×	×	×	×	×	×	×	○	-	-	-	×
ZIEL-I	*Actinomyces*	-	○	-	-	×	-	-	-	○	-	-	×	-	-	-	○	-	-	×	-	-	-	○	-	-	×	-	-	-	○	-	-	×	-	-
ZIEL-II	*Akkermansia*	×	○	○	-	×	-	-	×	○	○	-	×	-	-	×	○	○	-	×	-	-	×	○	○	-	×	-	-	×	○	○	-	×	-	×
ZIEL-I	*Alistipes*	×	×	×	×	×	×	×	-	-	-	-	×	-	○	-	-	-	-	×	-	○	-	-	-	-	×	-	○	-	-	-	-	×	-	○
ZIEL-II	*Atopobium*	-	-	-	-	-	-	○	-	-	-	-	-	-	○	-	-	-	-	-	-	○	-	-	-	-	-	-	○	-	-	-	-	-	-	○
Zymo	*Bacillus*	-	-	○	○	○	○	○	○	-	○	×	○	○	○	○	-	○	○	○	○	○	○	-	○	○	○	○	○	○	-	○	○	○	○	○
ZIEL-I	*Bacillus*	+	+	+	+	-	○	○	+	+	+	×	-	○	○	+	+	+	+	-	○	○	+	+	+	+	-	○	○	+	+	+	+	-	○	○
ZIEL-II	*Bacteroides*	○	○	○	○	×	○	-	○	○	○	○	×	○	-	○	○	○	○	×	○	-	○	○	○	○	×	○	-	○	○	○	○	×	○	-
ZIEL-I	*Bacteroides*	+	○	○	○	×	-	○	+	○	○	○	×	-	○	+	○	○	○	×	-	○	+	○	○	○	×	-	○	+	○	○	○	×	-	○
ZIEL-II	*Bifidobacterium*	-	-	+	○	×	+	+	-	-	+	○	×	+	+	-	-	+	○	×	+	+	-	-	+	○	×	+	+	-	-	+	○	×	+	+
ZIEL-I	*Cellulosimicrobium*	-	-	-	○	×	-	-	-	-	-	○	×	-	-	-	-	-	○	×	-	-	×	×	-	○	×	×	×	×	×	-	○	×	-	-
ZIEL-II	*Citrobacter*	○	○	+	○	-	×	×	○	○	-	×	×	×	×	○	-	-	×	×	×	×	○	○	×	×	×	×	×	○	○	-	×	×	○	-
ZIEL-II	*Clostridium XVIII*	×	×	×	×	×	×	×	○	+	-	-	○	-	×	○	+	-	-	○	-	○	×	×	○	×	×	×	×	○	+	-	-	○	-	○
ZIEL-I	*Clostridium XVIII*	×	×	×	×	×	×	×	-	-	-	-	+	-	○	-	-	-	-	+	-	○	×	×	×	×	×	×	×	-	-	-	-	+	-	○
ZIEL-II	*Collinsella*	○	-	+	-	-	-	-	○	-	+	-	-	-	-	○	-	+	-	-	-	-	○	-	×	-	-	-	-	○	-	+	-	-	-	-
ZIEL-II	*Eggerthella*	-	○	○	+	×	+	-	-	○	○	+	×	+	-	-	○	+	+	×	+	-	-	○	○	+	×	+	-	-	○	○	+	×	+	-
ZIEL-II	*Enterobacter*	×	×	×	×	×	×	×	×	-	×	×	×	×	×	○	-	○	×	×	×	-	×	-	-	×	-	-	-	×	×	-	×	×	×	-
Zymo	*Enterococcus*	-	○	+	+	○	+	+	-	○	○	+	○	+	+	-	○	○	+	○	+	+	-	○	○	+	○	+	+	-	○	○	+	○	+	+
ZIEL-I	*Enterococcus*	○	×	○	○	-	○	+	○	×	○	○	-	○	+	○	×	○	○	-	○	+	○	×	○	○	-	○	+	○	×	○	○	-	○	+
ZIEL-I	*Enterorhabdus*	○	○	-	-	-	-	-	○	○	-	-	-	-	-	○	○	-	-	-	-	-	○	○	-	×	×	×	×	○	○	-	-	-	-	-
Zymo	*Escherichia/ Shigella*	×	×	×	×	×	×	×	○	-	○	○	-	○	×	○	-	○	○	-	×	-	-	-	×	×	×	×	×	×	-	×	×	×	×	×
ZIEL-II	*Escherichia/ Shigella*	×	×	×	×	×	×	×	-	○	-	○	-	○	×	-	○	-	○	-	-	-	-	-	×	×	-	×	×	×	-	×	×	×	×	×
ZIEL-I	*Flavonifractor*	×	×	×	×	×	×	×	-	-	-	-	×	-	-	-	-	-	-	×	-	×	×	×	×	×	×	×	×	×	×	-	-	×	-	-
ZIEL-II	*Klebsiella*	-	○	○	-	-	-	-	○	○	○	×	○	○	-	○	○	○	×	×	○	-	○	○	○	×	-	-	-	○	○	○	×	×	○	×
Zymo	*Lactobacillus*	-	×	-	○	○	-	-	-	×	-	○	-	-	-	-	×	-	○	○	-	-	-	×	-	○	-	-	-	-	×	-	○	-	-	-
Zymo	*Listeria*	×	×	×	×	×	×	×	○	+	+	○	○	○	○	○	+	+	○	○	○	○	○	+	+	○	○	○	○	○	+	+	○	○	○	○
ZIEL-II	*Listeria*	×	×	×	×	×	×	×	○	○	○	○	+	○	×	○	○	○	○	+	○	○	○	○	○	○	+	○	○	○	○	○	○	+	○	○
ZIEL-II	*Microbacterium*	-	-	-	○	×	-	+	-	-	-	○	×	-	+	-	-	-	○	×	-	+	-	-	-	○	×	-	+	-	-	-	○	×	-	+
ZIEL-II	*Oscillibacter*	×	×	×	×	×	×	×	-	-	-	○	×	○	×	-	-	-	○	×	○	-	-	-	-	○	×	○	-	×	×	-	○	×	○	×
ZIEL-I	*Parabacteroides*	-	-	-	-	×	-	-	-	-	-	-	×	-	-	-	-	-	-	×	-	-	-	-	-	-	×	-	-	-	-	-	-	×	-	-
ZIEL-II	*Prevotella*	-	-	○	+	×	-	○	-	-	○	+	×	-	○	-	-	○	+	×	-	○	-	-	○	+	×	-	○	-	-	○	+	×	-	○
Zymo	*Pseudomonas*	-	-	○	-	-	×	+	-	-	○	×	-	×	×	-	-	○	-	-	-	+	-	-	○	○	-	-	+	-	-	○	○	-	-	+
ZIEL-II	*Pseudomonas*	○	-	○	○	-	×	○	○	-	○	×	-	×	○	○	-	○	○	-	○	○	○	-	○	○	-	○	○	○	-	○	○	-	○	○
ZIEL-I	*Pseudomonas*	×	×	×	×	×	×	×	-	-	-	-	-	×	-	-	-	-	×	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ZIEL-II	*Ruminococcus*	-	○	○	○	+	-	○	×	×	×	×	×	×	○	×	×	×	×	×	×	×	×	×	×	×	×	×	×	-	-	-	-	-	○	-
Zymo	*Salmonella*	+	-	+	○	-	-	-	+	-	○	×	-	-	-	+	-	+	×	-	-	-	+	-	○	+	-	-	-	+	-	○	×	-	-	-
Zymo	*Staphylococcus*	○	○	+	+	○	-	○	○	○	+	+	○	-	○	○	○	+	+	○	-	○	○	○	+	+	○	-	○	○	○	+	+	○	-	○
ZIEL-II	*Staphylococcus*	-	-	○	○	-	○	○	-	×	○	-	-	-	○	×	-	-	-	-	-	+	-	-	-	-	-	-	-	-	○	-	-	-	-	-
ZIEL-I	*Staphylococcus*	-	-	-	-	-	○	-	-	-	-	-	-	○	-	-	-	-	-	-	○	-	-	-	-	-	-	○	-	-	-	-	-	○	-

2.3 How low can we go? Implementation of ddPCR allows amplicon sequencing of ultra-low amounts of gDNA from low biomass samples

Drafted manuscript, in submission

2.3.1 Abstract

Background: One limiting factor of short amplicon 16S rRNA gene sequencing approaches is the use of low gDNA amounts in the amplicon generation step. Especially for low-biomass samples, insufficient or even commonly undetectable DNA amounts can limit or even prohibit further analysis in standard protocols.

Results: Using a newly established protocol, very low gDNA input amounts were found sufficient for reliable detection of bacteria using 16S rRNA sequencing compared to standard protocols. The improved protocol includes an optimized amplification strategy by using a digital droplet PCR. We demonstrate how PCR products are generated even when using ultra-low concentrated gDNA, unable to be detected by using a Qubit. Importantly, the use of different 16S rRNA gene primers had a greater effect on the resulting taxonomical profiles, compared to using high or very low gDNA amounts.

Conclusion: Our improved protocol takes advantage of ddPCR and allows faithful amplification even of very low amounts of template. With this, samples of low bacterial biomass become comparable to those with high amounts of bacteria. Besides, it is imperative to state gDNA concentrations and volumes used and to include negative controls indicating possible shifts in taxonomical profiles. Despite this, results produced by using different primer pairs cannot be easily compared.

Keywords: ddPCR, ultra-low gDNA amounts, low-biomass samples, 16S rRNA gene sequencing

2.3.2 Background

In 1985, the 16S rRNA gene was described for the first time as a molecular tool for identifying microbes that were previously shown to be unculturable (Lane *et al.*, 1985). This ubiquitous bacterial gene possesses special features containing conserved regions that enable primer binding and thus amplification, as well as hypervariable regions allowing phylogenetic differentiation. Thus, sequencing of the 16S rRNA gene is the current method of choice to analyze taxonomical profiles of mixed bacterial communities (Bukin *et al.*, 2019, Vos *et al.*, 2012). An often applied, easy, time- and cost-efficient method nowadays is short-amplicon sequencing using second-generation sequencers such as the Illumina MiSeq. Several factors

affecting 16S rRNA gene sequencing results have been widely studied. Some of those are sampling and sample storage (Choo *et al.*, 2015, Flores *et al.*, 2015, Ma *et al.*, 2020, Penington *et al.*, 2018), the use of different variable regions or primers (Abellan-Schneyder *et al.*, 2021a, Fouhy *et al.*, 2016, Klindworth *et al.*, 2012, Thijs *et al.*, 2017, Tremblay *et al.*, 2015), sequence processing including the use of different denoising approaches, reference databases, and downstream analysis pipelines (Almeida *et al.*, 2018, De Filippis *et al.*, 2018, Marizzoni *et al.*, 2020, Nearing *et al.*, 2018, Park and Won, 2018, Sierra *et al.*, 2020).

In addition to the above, it was previously shown that the extracted genomic DNA (gDNA) can impact 16S rRNA analysis in two ways. Firstly, the use of different extraction methods or protocols influences the composition of a given sample (Fiedorová *et al.*, 2019, Hart *et al.*, 2015, Lim *et al.*, 2018, Wagner Mackenzie *et al.*, 2015, Wesolowska-Andersen *et al.*, 2014). More precisely, easy to lyse Gram-negative bacteria are favored by several extraction methods compared to hard to lyse Gram-positive bacteria (Costea *et al.*, 2017, McOrist *et al.*, 2002, Santiago *et al.*, 2014). Secondly, the gDNA concentration used for amplicon generation can influence the resulting taxonomical profiles (Multinu *et al.*, 2018). This becomes even more critical when low biomass samples are analyzed, because contaminations of those samples would more likely affect the resulting taxonomical profiles and, thus, could lead to misleading study results (Dahlberg *et al.*, 2019, Glassing *et al.*, 2016, Salter *et al.*, 2014). Lowering input amounts for 16S rRNA gene sequencing approaches are of particular interest for researchers investigating, e.g., the lower respiratory tract, preterm child microbiomes, stool samples of patients treated with antibiotics, milk samples, or any other sample type which is considered to be of low bacterial biomass (Claassen-Weitz *et al.*, 2020, Davis *et al.*, 2019, Douglas *et al.*, 2020, Saladié *et al.*, 2020, Stinson *et al.*, 2019). Only very few studies tried to find minimum input amounts that are needed to produce reliable results. Brandt & Albertsen (2018) defined a detection limit for bacteria in drinking water. They showed that if the bacterial input is 10^1 cells/ml or smaller, several contaminating Operational Taxonomic Units (OTUs) appeared, and thus, sample outcomes could not be counted as reliable data. Multinu *et al.* (2018) reported that a minimum concentration of 40 pg/µl and an ideal concentration of >200 pg/µl produce reliable 16S rRNA gene profiles of human stool samples. Velásquez-Mejía *et al.* (2018) showed that they needed at least 2 mg of fecal sample to extract sufficient gDNA for 16S rRNA gene sequencing and the lowest successful amount of gDNA used was 500 pg/µl in their study. Here, we wanted to assess whether we could decrease the minimum input amount of gDNA needed for reliable 16S rRNA gene sequencing of human stool samples even further. As a comparison, input amounts of 1-100 ng of gDNA are commonly used for PCRs designated to perform later 16S rRNA gene sequencing. Illumina suggests using 12.5 ng total gDNA input for a first step PCR (Illumina Inc., 16S Metagenomic Sequencing Library Preparation, Part #15044223 Rev.B). In our lab, we use 12 ng total gDNA in our standard 16S rRNA gene

sequencing approaches (Reitmeier *et al.*, 2020). To enable the use of lower input amounts, we used a digital droplet PCR (ddPCR) approach followed by standard short amplicon 16S rRNA gene sequencing.

Common PCRs take place in larger reaction volumes between about 20-50 µl. An advancement is the ddPCR, which splits the larger volume into about 20,000 droplets, in which independent reactions occur within each droplet. Dividing the PCR volume into thousands of droplets has certain advantages: ddPCR was shown to be less sensitive to inhibitors (Dreo *et al.*, 2014) and was able to allow for selective, and reproducible detection of rare alleles and the absolute quantification of targeted gene copy numbers (Hindson *et al.*, 2011). Other benefits of ddPCR protocols are a reduced PCR bias, by avoiding preference in the amplification of specific products over others by dividing the reaction mixture into small droplets, a simplified quantification compared to qPCR, and reduced consumable costs, as reaction volumes are really small (Demeke and Dobnik, 2018). Gobert *et al.* (2018) showed a quantification method for low amounts of *Lactobacilli* in fecal samples using a ddPCR approach. There, quantification was possible even though only low numbers of the target strains were present with high background noise. Wouters *et al.* (2020) stated that by using a ddPCR protocol, they could detect very low amounts of a pathogen's DNA in whole blood samples in as short as four hours. Nevertheless, to our knowledge, no protocol has been published which reuses a ddPCR product for sequencing, e.g., 16S rRNA genes. Thus, the use and limits of ddPCR were tested in this study in order to reliably obtain results for 16S rRNA gene sequencing with very low input amounts of gDNA.

2.3.3 Methods

Preparation of human gut samples

Stool samples were obtained from healthy volunteers of age after informed and written consent. An ethics approval is deemed unnecessary according to the statement given in the Drucksache 15/2849 of the German Bundestag about § 41 Abs. 2 Nr. 2 S. 1 and 2 Arzneimittelgesetz. Stool samples were collected in stool sample tubes as described previously by Abellan-Schneyder *et al.* (2021a).

Extraction of gDNA from stool samples

Genomic DNA was isolated using a modified protocol by Godon *et al.* (1997) as described previously by Reitmeier *et al.* (2020) and Abellan-Schneyder *et al.* (2021a).

Extraction of gDNA from mock communities

DNA of the Zymo mock community was purchased as a ready-to-use DNA mock (D6306, Zymo Research). Extraction and preparation of the ZIEL2 mock community was performed as

described in Abellan-Schneyder *et al.* (2021a). In brief, every 19 bacterial strains (18 different bacterial genera) of diverse taxonomy were cultured and afterward harvested by centrifugation. The extraction of genomic DNA (gDNA) was performed separately for each strain. For the ZIEL2 mock community DNA mixture, 12 ng of each bacterial gDNA was pooled.

Determination of concentration and dilution of gDNA input

Initial sample concentrations were measured in triplicates on a Qubit 4.0 (Thermo Fisher). According to the initial concentrations, stock solutions of 10 ng/µl and 1 ng/µl were set up and again measured in triplicates on a Qubit 4.0 (Thermo Fisher). The following dilution series, to reach the desired final concentrations (Table 2.3.1) were performed in 0.5 ml LoBinding Tubes (Eppendorf). After each dilution step, samples were briefly vortexed and spun down on a mini centrifuge.

Table 2.3.1: Concentrations and dilutions of gDNA input used for 1st-step PCR reaction.

Name	total input DNA (ng)	final DNA concentration in 50 µl PCR reaction (pg/µl)
60	60	1200
12	12	240
10	10	200
5	5	100
1	1	20
0.5	0.5	10
0.1	0.1	2
0.05	0.05	1
0.01	0.01	0.2

Amplicon preparation

For amplification of the variable regions and addition of adapters, a 1st-step PCR was performed in 50 µl volume containing 10 µl gDNA (total amounts are detailed in Table 2.3.1), 1x Phusion HF buffer, 0.2 mM dNTPs, 0.125 µM of each fw_primer and rv_primer, 7.5% (v/v) DMSO and 0.25 µl of Phusion HF II DNA polymerase (Thermo Fisher). PCR was performed as followed: 98°C for 40 s, followed by 15 cycles of 98°C for 20 s, V-region specific annealing temperature (Table 2.3.2) for 40 s and 72°C for 40 s, followed by a final extension step at 72°C for 2 min. The structure of the primers was 5' → 3': "overhang – $[N]_{15}$ – 16S specific sequence" for the 1st-step and "P5/P7 – 8 bp Barcode – overhang" for the 2nd-step PCR. To enable multiplexing, barcodes were added in a 2nd-step PCR. Here, a 100 µl PCR was prepared using 10 µl of the 1st-step PCR product, 1x Phusion HF buffer, 0.2 mM dNTPs, 0.125 µM of each fw_barcode and rv_barcode primer, 0.25% (v/v) DMSO, and 0.5 µl of Phusion HF II DNA polymerase. PCR conditions were 98°C for 40 s, 10 cycles of 98°C for 20 s, 55°C for 40 s and 72°C for 40 s as well as a final extension step at 72°C for 2 min. Further details, e.g., work time estimations, can be found in the work of Reitmeier *et al.* (2020).

Table 2.3.2: Variable region-specific forward and reverse primers and annealing temperature for 1st-step PCR.

Region	Forward primer	Reverse primer	Annealing Temperature	Reference
V1-V2	AGA GTT TGA TYM TGG CTC AG	GCT GCC TCC CGT AGG AGT	57°C	Salter *et al.* (2014)
V3-V4	CCT ACG GGN GGC WGC AG	GAC TAC HVG GGT ATC TAA TCC	55°C	Klindworth *et al.* (2012)
V7-V9	CAA CGA GCG CAA CCC T	GGT TAC CTT GTT ACG ACT T	51°C	Turner *et al.* (1999)

Library quality check

For validation and quality assurance, 8 µl of 2nd-step PCR product were loaded on a 1.5% (w/v) agarose gel to perform gel electrophoresis. The remaining 92 µl of the 2nd-step PCR were purified with 0.6x AMPure XP beads. Concentrations of the 2nd-step PCR product were measured in triplicates using a Qubit 4.0.

Digital droplet PCR

Amplicons generated in the two-step PCR (above) were amplified again using P5 and P7 primers in a ddPCR. At first, each sample was diluted to a concentration of approx. 20,000 copies/20 µl calculated by the Formula (2.3.1). The average library sizes were 486 bp for V1-V2, 602 bp for V3-V4, and 547 bp for V7-V9. Dilution series must be performed in LoBind tubes (Eppendorf) and in 1:10 steps to guarantee precise dilutions.

$$\frac{660\frac{\mathrm{g}}{\mathrm{mol}}\times\text{average library size [bp]}}{6.022\times10^{23}\ \mathrm{mol}^{-1}}\times 10^{9}\times\frac{20{,}000}{20\ \mu\mathrm{l}}=\text{concentration}\ [\frac{\mathrm{ng}}{\mu\mathrm{l}}] \qquad (2.3.1)$$

The composition of the reaction mixture for the ddPCR was as follows: 1× QX200™ EvaGreen® Supermix, 0.1 µM of P5 (forward) and 0.1 µM of P7 (reverse) primer, 2.5 µl DNA sample (1,000 copies/µl) and water up to 25 µl. These ingredients were mixed thoroughly by vortexing and 20 µl of the mixture was transferred into a DG8™ Cartridge for the QX200 Droplet generator. Next, 70 µl of QX200 Droplet Generator Oil for EvaGreen was transferred into the oil well of the cartridge. Then a gasket was spanned over the cartridge and droplets were produced by the droplet generator following the Droplet Generator Instruction Manual (BioRad). Droplets were then transferred to a 96-well plate. Before starting the PCR in a thermocycler, the plate was sealed with a pierceable PCR Plate Heat Seal Foil (BioRad) using a PX1 PCR Plate Sealer (BioRad). PCR was performed in a PeqStar thermocycler (PeqLab) using cycling conditions as described in Table 2.3.3.

Table 2.3.3: Cycling conditions for ddPCR using P5 and P7 primer amplifying amplicons.

Cycling Step	Temperature (°C)	Time	Ramp Rate	Number of Cycles
Enzyme Activation	95	5 min	2°C/s	1
Denaturation	95	30 s		40
Annealing	59	1 min		
Extension	72	1 min		
Signal Stabilization	4	5 min		1
	90	5 min		1
Hold	4	∞		1

The PCR products were recovered for further use of the amplicons. Each reaction was transferred into a clean 1.5 ml LoBind DNA tube (Eppendorf), and the lower oil phase was discarded by pipetting. After adding 20 µl 1x TE buffer and 70 µl chloroform to the remaining aqueous phase, mixtures were vortexed for 1 min at high speed in a 2 ml adapter for the Vortex-Genie 2 (Thermo Fisher) and centrifuged at 15,500×*g* for 10 min. The upper aqueous phase (volume approx. 25 µl), containing amplicons, was separated by pipetting. Samples were purified using 1x AMPure XP beads and eluted in 20 µl H_2O. The concentration was determined using the Qubit dsDNA HS Assay (Invitrogen). To analyze the size of the ddPCR product, agarose gel electrophoresis (1.5%, w/v) was performed with 4 µl of each sample.

Re-amplification of ddPCR using Q5U Polymerase

If not sufficient product for 16S rRNA gene sequencing could be extracted from the ddPCR, re-amplification was performed. Of note, re-amplification of the ddPCR product is only possible using a non-proofreading polymerase (e.g., *Taq* polymerase) or by using a polymerase that can read and amplify templates containing uracil (and inosine bases), e.g., Q5U or Phusion U DNA polymerase. The re-amplification reaction mix contained: 1× Q5U reaction buffer, 200 µM dNTPs (10 mM); 0.5 µM P5 primer (forward), 0.5 µM P7 primer (reverse), 5 µl ddPCR product (≤1 ng/µl), 0.02 U/µl Q5U Hot Start High-Fidelity DNA Polymerase, up to 50 µl nuclease-free H_2O. Cycling conditions were set as described in Table 2.3.4.

Table 2.3.4: Cycling conditions for re-amplification of ddPCR.

Cycling Step	Temperature (°C)	Time	Number of Cycles
Initial Denaturation	98	30 s	1
Denaturation	95	10 s	5
Annealing	55	20 s	
Extension	72	45 s	
Final Extension	90	5 min	1
Hold	4	∞	1

After the re-amplification, the PCR products were checked for the desired amplicon lengths via agarose gel electrophoresis. Samples showing bands at the desired size were purified by PAGE purification and eluted in 25 µl nuclease-free water. Concentrations were measured with the Qubit dsDNA HS Assay using 2 µl of the extracted amplicons.

Sequencing

Samples were adjusted to 0.5 nM and pooled. Samples were sequenced in paired-end modus on a cartridge v3 using PE300 of a MiSeq system (Illumina, Inc.) following the manufacturer's instructions and a final DNA concentration of 12 pM and 15% (v/v) PhiX standard library.

Data analysis using IMNGS and Rhea

Data were processed using the Integrated Microbial Next-generation sequencing (IMNGS) pipeline (Lagkouvardos *et al.*, 2016), an in-house developed pipeline based on UPARSE (Edgar, 2013). In the advanced IMNGS options, allowed mismatches were set to one. Demultiplexing was performed using a minimum read-length of 250 bp and a maximum read-length of 600 bp. Forward trim was set to 30 bp and reverse trim length was 60 bp. The abundance filter was set to 0.0025 (Clavel *et al.*, 2020) and the filter of low-read samples was set to off. Chimeric reads were removed using UCHIME (Edgar *et al.*, 2011) and zero-radius operational taxonomic units (zOTUs) were produced using UNOISE 2 (Edgar, 2016) and USEARCH v11.0.667. Further analysis was performed in Rhea (Lagkouvardos *et al.*, 2017). Rhea is a collection of R-scripts enabling comparison between samples. After normalization of data, alpha- and beta-diversities can be visualized. The script also performs taxonomic binning, enabling an insight on all known and unknown sequences of the microbial composition down to the genus level.

2.3.4 Results

Study overview

The influence of the initial input amount of gDNA was studied in detail. Towards this end, a general, published workflow for 16S rRNA gene amplicon sequencing was followed for the first part of the library preparation (Reitmeier *et al.*, 2020). Subsequently, after purifying the amplicons after the 2nd-step PCR (i.e., barcoding), a ddPCR was added, allowing processing very low gDNA input amounts (Figure 2.3.1).

To establish the new protocol, two different tests were performed. We prepared and sequenced, firstly, standard PCR products generated by using 12 ng gDNA input (standard amount for 16S rRNA gene sequencing approaches in our laboratory). The very same samples were diluted to different degrees and processed after dilutions by the additional ddPCR step. This shows whether the results are comparable or whether the ddPCR step introduces biases. Secondly, we performed dilution series and used decreasing amounts of initial gDNA input (60, 10, 5, 1, 0.5, 0.1, 0.05, and 0.01 ng total input) in the 1st-step PCR and evaluated whether we were able to produce reliable results even when gDNA input amounts below 500 pg were used. Taken together, we compared resulting taxonomical profiles and whether they are independent

of the gDNA input amount used in 1st-step PCR or independent of an additional ddPCR (Figure 2.3.1).

In brief, gDNA of stool samples were extracted, concentrations were measured, dilution series were performed, and 1st-step PCRs were set up. In the 1st-step PCR, primer amplifying different V-regions were used (e.g., V1-V2, V3-V4, and V7-V9). Products were cleaned and used as a template for the 2nd-step PCR. Primers used include barcodes and the Illumina sequencing primer (P5 and P7). The resulting amplicons were again cleaned up and checked whether the desired library size could be observed on agarose gels. The detection limit of the used GelRed dye is reported to be about 100 pg (Biotium, https://biotium.com/faqs/category/gelred-gelgreen/). However, it should be kept in mind that the actual limit depends on the used instrument's capability and exposure settings. Sharp and conclusive bands were observable with our equipment and settings for at least 2-5 ng DNA. For establishing the protocol, amplicons of the 2nd-step PCRs were diluted according to Formula (2.3.1). Next, ddPCR mixes were produced with different amounts of the above amplicons as input. The primers used were plain P5 and P7 primers, which allow the re-amplification of the templates generated thus far. The final ddPCR amplicons were extracted and checked for adequate concentrations allowing 16S rRNA gene sequencing. If concentrations were too low, re-amplification in a standard PCR but using a Q5U polymerase was performed. Afterward, all samples were sequenced on an Illumina MiSeq and compared.

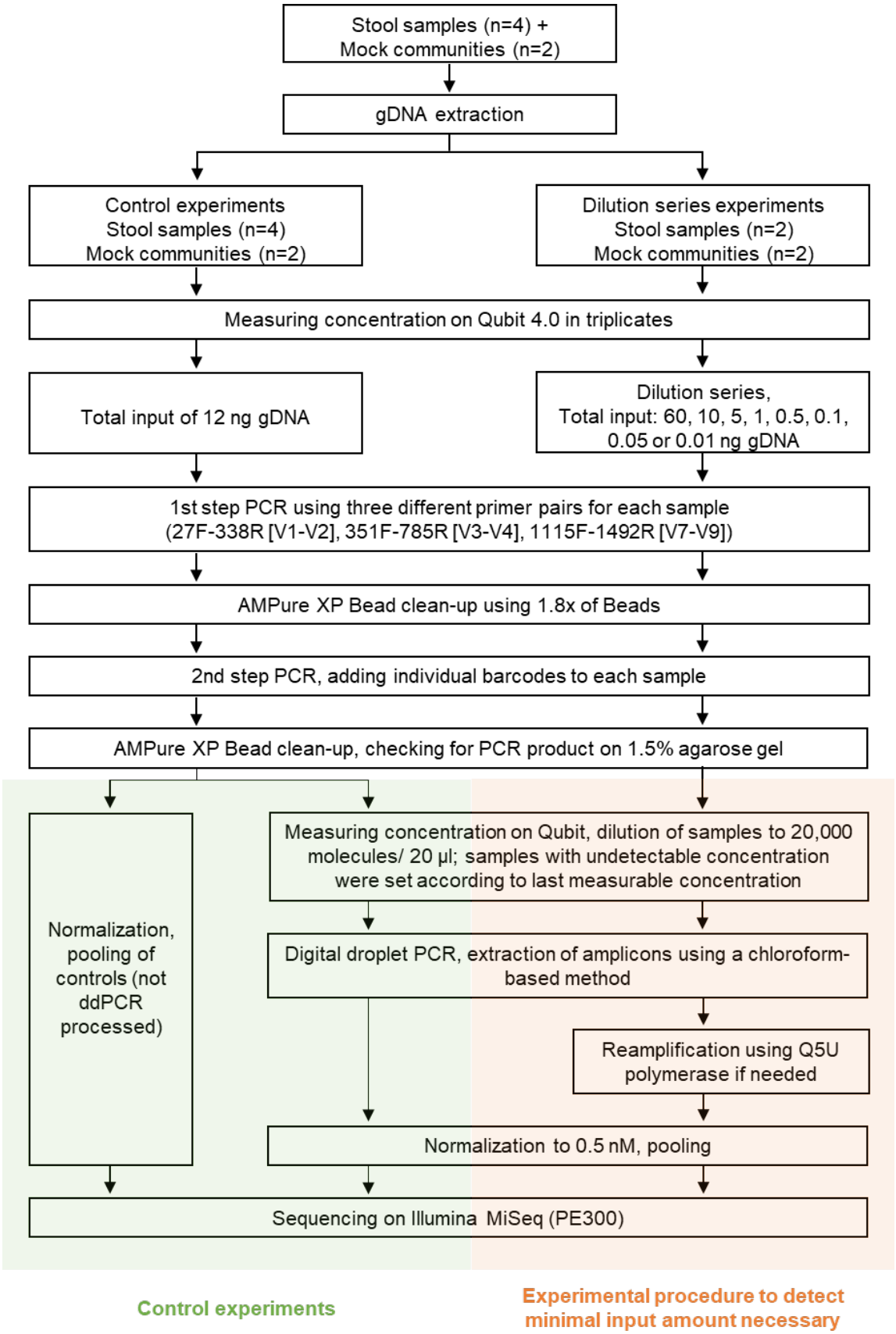

Figure 2.3.1: Overview of experimental procedure of this study. Experiments are divided into two parts. Left, control experiments (shaded green) were used for checking if the additional ddPCR step did not introduce bias in the resulting taxonomical profiles. Right, experimental procedures (shaded red) were described to detect the minimum input amount of gDNA input necessary to produce reliable 16S rRNA gene sequencing results.

Determination of detection limits in standard 16S rRNA gene sequencing approaches

For all samples of the control experiment, products were detectable using agarose gels after the 2nd-step PCR. For the dilution series experiment, bands corresponding to the desired product were only visible for gDNA inputs ≥5 ng total gDNA, irrespective of which primers were used for amplification. After the ddPCR, bands could be observed on agarose gels for all samples amplified with primers targeting V7-V9. For V1-V2 samples, clear bands were visible for input amounts ≥50 pg. The detection limit for V3-V4 samples was higher; bands could only be detected for input amounts ≥500 pg for all samples, while some samples produced products at 100 pg already (Table 2.3.5).

Table 2.3.5: Visibility of bands corresponding to desired PCR products observed on 1.5% (w/v) agarose gels. Green tick: bands were visible for all tested samples, yellow tick in brackets: bands were weak and/ or not visible for all tested samples. Red x: bands were not visible for none of the tested samples.

		gel after 2nd step PCR			gel after ddPCR		
total gDNA input	V-region amplified	V1-V2	V3-V4	V7-V9	V1-V2	V3-V4	V7-V9
60 ng		✓	✓	✓	✓	✓	✓
12 ng		✓	✓	✓	✓	✓	✓
10 ng		✓	✓	✓	✓	✓	✓
5 ng		✓	✓	✓	✓	✓	✓
1 ng		×	×	×	✓	✓	✓
500 pg		×	×	×	✓	✓	✓
100 pg		×	×	×	✓	(✓)	✓
50 pg		×	×	×	✓	×	✓
10 pg		×	×	×	(✓)	×	✓

If no band could be observed for a sample after ddPCR had been performed, re-amplification of the (invisible) product was conducted. Importantly, re-amplification was only possible when a uracil-tolerant polymerase, e.g., *Taq* polymerase or a U-tolerant proofreading polymerase such as Q5U (NEB) or Phusion U Hot Start DNA Polymerase (Thermo Fisher) was used. The QX200 EvaGreen supermix contains some amounts of dUTP, causing ddPCR products to contain uracil subsequently. dUTP is used to allow the destruction of carry-over products from previous PCRs using Uracil-N-Glycosylase in the PCR mixture (Pruvost *et al.*, 2005). In any case, re-amplification with normal proof-reading polymerases such as the Phusion (Thermo Fisher) is inhibited, and products can only be re-amplified with the mentioned U-tolerant polymerases.

The number of final sequenced reads, irrespectively of which approach was used, varied between 11,298 to 118,212 reads per sample, with an average of 42,754 reads. The average read number of the negative controls was 506.

Control experiment to assess the potential bias of the ddPCR step

In a first experimental setup, we assessed whether the integration of a ddPCR step after the 2nd-step PCR used for barcoding showed a bias on the β-diversity and the resulting

taxonomical profiles of the samples. Ideally, samples originating from the same human donor or the same mock community should not show any or only minor differences. We screened, therefore, four human samples (T1, T28, T29, T30) and two mock communities of known composition. The latter show different amounts of complexity as they are either composed of 8 different bacterial genera (Zymo mock community) or 18 different genera (ZIEL2 mock community). We further sequenced the samples using three different primer pairs amplifying V1-V2, V3-V4, and V7-V9.

Comparing the results, we found that the differences introduced by using different primer pairs for the different V-regions caused profiles to be more distinct from each other than differences introduced by either the preparation method (standard protocol, marked as Sample-C in Figure 2.3.2, vs. protocol with additional ddPCR, marked as Sample-D in Figure 2.3.2) or the donor (i.e., samples originating from the same donor do not cluster close to each other when amplified using different primer pairs). Concerning the latter, the difference between the three tested regions V1-V2, V3-V4, and V7-V9 (Figure 2.3.2A, in red, green, and blue, respectively) was significant with a p-value of ≤0.001 tested with PERMANOVA. More importantly, we could demonstrate that the additional ddPCR step did not lead to significant differences in final sample composition, as shown by only an insignificant difference in tree clustering (Figure 2.3.2B) and only very little shifts in the resulting taxonomical profiles of the samples (Figure 2.3.2C). The Zymo mock community performed overall well, regardless of which V-region was targeted. For the more complex ZIEL2 mock community, we could show by calculating the generalized UniFrac distances against the ideal composition that the most accurate representation was produced by targeting the V3-V4-region (see Supplementary Table 2.3.1).

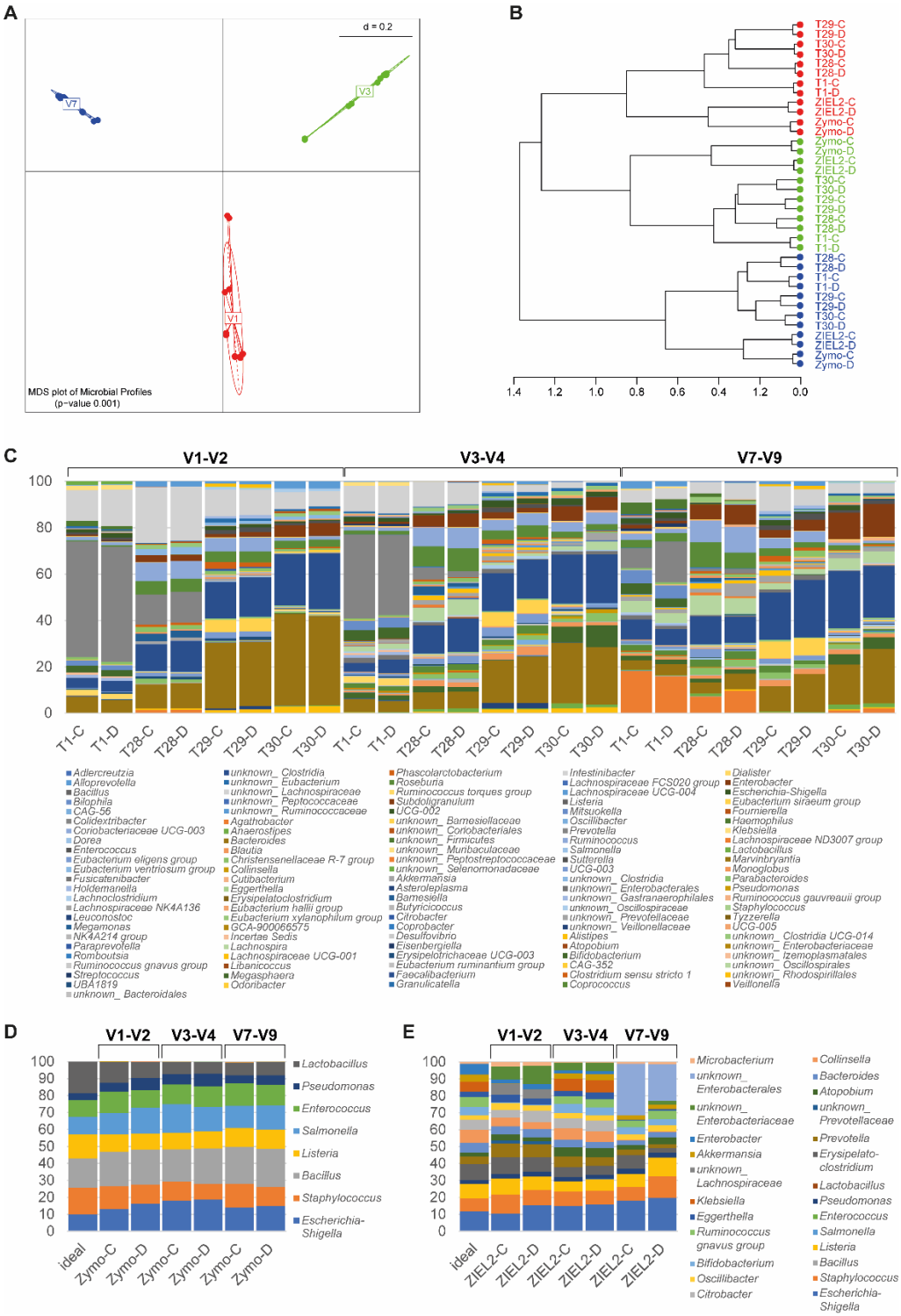

Figure 2.3.2: Control experiment to test for biases possibly introduced due to the extra ddPCR step after 2nd-step PCR. Samples processed with ddPCR (Samples marked "D" for ddPCR processed, 12 ng gDNA used) are compared to standard short amplicon controls which were not ddPCR processed (Samples marked "C" for Control, 12 ng gDNA used). Four human samples (T1, T28, T29, T30) and two mock communities (Zymo, ZIEL2) were sequenced using primer pairs amplifying different V-regions (V1-V2: red, V3-V4: green, V7-V9: blue). (**A**) Multi-Dimensional Scaling (MDS) shows that samples cluster significantly differently due to V-region targeted and not by preparation method or sample origin. (**B**) The dendrogram of all the samples cluster together inside the same V-region. (**C**) Taxonomic profiles at genus-level of Sample-C and Sample-D for human samples. (**D**) As before, for mock samples from Zymo and (**E**) and ZIEL2. Note, the taxonomic profiles at the genus level showed only minor differences when the same V-region is targeted.

Estimation of minimal gDNA input amounts for reliable 16S rRNA gene sequencing

In a second experimental set-up, dilution series of total gDNA input amounts were tested for the lowest gDNA input possible, producing reliable taxonomic profiles of human samples or mock communities. As before, samples were amplified using three different primer pairs. As shown in Tab. 2.3.5, detection limits varied for different V-regions. Thus, the taxonomic composition of each sample was checked for differences from either the actual composition in the case of mock communities or the taxonomical profiles achieved, amplifying high initial gDNA amounts in the case of the human samples. For the Zymo mock community, it was found that Gram-negative bacterial genera such as *Escherichia*, *Pseudomonas,* or *Salmonella* were increasingly overestimated with descending gDNA amounts, while Gram-positive bacteria genera, e.g., *Lactobacillus*, *Listeria,* or *Staphylococcus* were progressively underestimated. When analyzing the ZIEL2 mock, primer-dependent issues become more prominent, which has been observed before (Abellan-Schneyder *et al.*, 2021a). In contrast to the Zymo mock community, no clear tendencies concerning different genera could be observed for the human samples despite the increase of spurious sequencing reads arising in very diluted samples (combined in “other”).

Amounts ≤10 pg gDNA did not always produce reliable results when taxonomical profiles at the genus level were analyzed. For V3-V4 and V7-V9, deviations from the expected composition become more apparent with increasingly less gDNA used as initial input (Figure 2.3.3). For the highly diluted Zymo mock, we observed increasing numbers of reads not representing members of the original mock community. Overall, for Zymo DNA, the average amount of reads not corresponding to the expected bacteria was 0.9%. For V1-V2, the median amount of off-target reads for samples of 60 ng to 50 pg was 0.24%, and for 10 pg input, 1.27%. For V3-V4 and V7-V9, a drastically increased number of reads not matching the mock could be identified when using 10 pg input DNA. The average amount for off-target sequences was 0.19% and 0.24% for V3-V4 and V7-V9, respectively, of all reads concerning input amounts varying between 60 ng to 50 pg. The number of off-target reads for 10 pg samples reached 8.8% and 7.1% for V1-V2 and V7-V9, respectively. For human sample T1, 10 pg gDNA input was not sufficient when targeting V3-V4. Deviations from the expected taxonomic profile are apparent for this low amount of input DNA used (Figure 2.3.3). Thus, it seems that detection limits are not only V-region specific but also dependent on each sample. For instance, for T30, reliable profiles with just 10 pg input DNA targeting V3-V4 were produced, while T1 failed for the same combination and needed at least 50 pg input DNA.

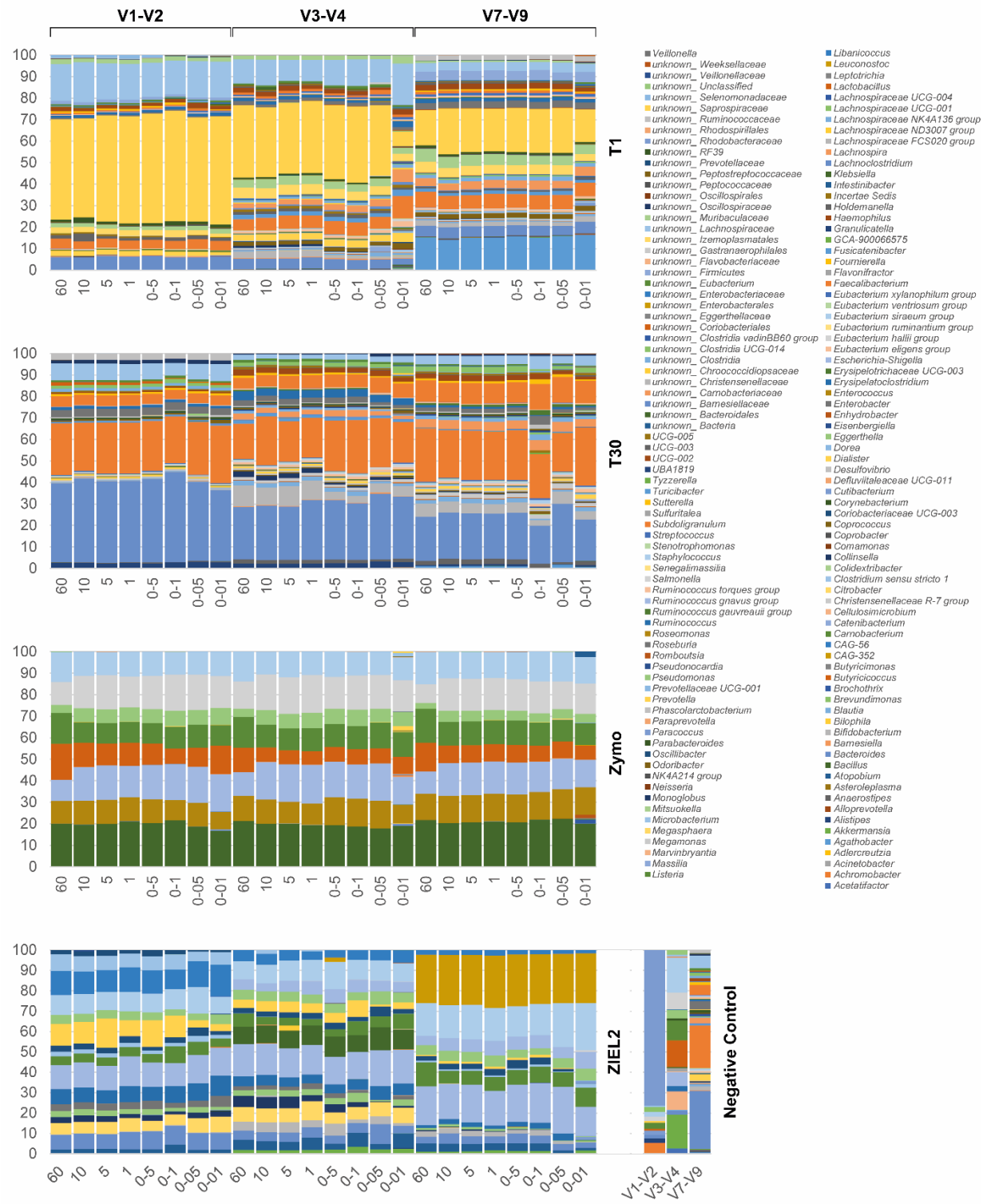

Figure 2.3.3: Taxonomical profiles at the genus level for two human samples (T1, T30) and two mock communities of known composition (Zymo, ZIEL2). For every sample, different initial gDNA amounts were used for 1st-step PCRs, and, further, different V-regions were targeted.

2.3.5 Discussion

In this study, it was investigated how to obtain reliable 16S rRNA sequencing-based taxonomies even with very low input amounts of gDNA. We found that the introduction of a

ddPCR step after standard PCR-based library production allowed using low-DNA concentrated samples. The ddPCR allowed to successfully and reliably re-amplify 16S rRNA amplicons from the foregone PCR steps, even if they were not detectable in gel electrophoreses nor measurable using a Qubit. The minimal gDNA input amount successfully used in this study for all samples was 50 pg total gDNA (equating to 1 pg/µl in the 1st-step PCR mix), while some samples were accessible with even lower input amounts. Nevertheless, 50 pg input DNA is by the factor 4 lower than in previously published studies, e.g., Multinu *et al.* (2018), who stated that concentrations <4 pg/µl should be interpreted with caution. Worse, Velásquez-Mejía *et al.* (2018) needed at least 500 pg/µl gDNA. Moreover, we could not confirm some of the other observations made by those groups. Multinu *et al.* (2018) described an overrepresentation of *Proteobacteria* and an underrepresentation of *Firmicutes* for low gDNA input samples. When using ddPCR, no general trend became obvious when analyzing human samples. For instance, for sample T1, the deviations between high and low DNA input were dependent on the targeted region rather than on the amount of input DNA. When sample T1 was amplified using 27F and 338R primer (V1-V2), we could identify a reduction in *Firmicutes* and *Proteobacteria* for samples with lower gDNA input amounts, whereas *Bacteroidota* (formerly *Bacteroidetes*) seemed to be overrepresented when using low gDNA amounts. When V3-V4 was targeted in T1, *Firmicutes* and *Proteobacteria* seemed to be overrepresented and *Bacteroidota* underrepresented. Concerning V7-V9, we saw no distinct change in *Firmicutes* or *Proteobacteria* amounts but an overrepresentation in *Bacteroidota* for T1 (data are summarized in Supplementary Table 2.3.2). Even more, for T30, the trends do neither follow the shifts we saw for T1 nor those described by Multinu *et al.* (2018). Generally, the phyla-level composition seems to be more inconsistent for T30 than for T1 when analyzing a dilution series within one targeted region (see Supplementary Table 2.3.3).

We conclude that the variation and shift in taxonomical compositions are mainly driven by using different primer pairs. This biasing factor was already intensively studied previously (e.g., Abellan-Schneyder *et al.*, 2021a, Fouhy *et al.*, 2016, Nelson *et al.*, 2014, Thijs *et al.*, 2017). As before, we could show that samples cluster mainly due to V-region(s) targeted and not due to sampling origin or limited input amount. Thus, not only detection limits but also a concentration-dependent association with certain taxa, under- or overrepresented at distinct starting gDNA concentrations, must be interpreted with caution. We suggest that for every sample type and primer pair, initial control experiments should be performed, evaluating minimal DNA amounts needed to produce reliable results.

The limitation of our study is the use of human stool samples with basically unknown composition. While, generally, studies analyzing human stool samples do not face problems concerning low initial gDNA concentrations, we needed a sample type for which we could test

high and low input amounts of gDNA. However, low DNA amounts were accessible after careful dilutions in low-bind tubes.

Interestingly, most of the commonly used protocols use high or even very high amounts of gDNA in their protocol. To list only some examples: the Zymo Quick 16S NGS Library Prep Kit (Zymo Research Europe GmbH, Freiburg, Germany) aims for about 40 ng gDNA that is free of PCR inhibitors; the QIAseq 16S kit (Qiagen, Hilden, Germany) recommends amounts of 12.5 ng, and the lowest amount usable is given with 1 ng, which is at least 20-fold higher compared to our protocol. Further, in the NEBNext Ultra DNA Library Prep kit (New England Biolabs, Ipswich, USA) for Illumina, 500 pg to 1 µg of input DNA is recommended. In this proof-of-principle study, we show that very-low initial gDNA concentrations, which are by far lower than the recommended input amounts listed above, can be successfully sequenced and reliably analyzed when implementing a ddPCR step. This is of special interest for low-bacterial biomass samples, such as milk, water samples, pathological or clinically relevant human samples, including sputum, infant stool, biopsies, and others. While several 16S rRNA gene sequencing optimization protocols for such samples were already published (e.g., Claassen-Weitz *et al.*, 2020, Davis *et al.*, 2019, Douglas *et al.*, 2020, Saladié *et al.*, 2020, Stinson *et al.*, 2019). Nevertheless, these studies aim at changing the parameters of existing protocols or try to reduce contamination sources in order to obtain taxonomic profiles of such low-biomass samples. In contrast, ddPCR, which has to our knowledge not been applied to improve sequencing of low biomass samples, has only to be added at the end of a commonly used 16S rRNA gene sequencing protocol. While ddPCR methods were already described for quantification of microbial species or communities (Dreo *et al.*, 2014, Gobert *et al.*, 2018, Manzari *et al.*, 2020, Pacocha *et al.*, 2019, Ziegler *et al.*, 2019), to our knowledge, resulting PCR products were never re-extracted from the oil-aqueous suspensions. Here we demonstrate that these products can be successfully sequenced, producing reliable taxonomic profiles. Even taking this a step further, these products can be re-amplified after ddPCR (e.g., in case of still too low concentrations), but uracil accepting polymerases must be used.

2.3.6 Conclusions

Taken together, ddPCR, which splits the reaction volume in about 20,000 droplets, allows faithful amplification even of low amounts of template. Thus, sequencing of samples of low bacterial biomass (e.g., of a sick person with low bacterial loads), currently not accessible, can now be sequenced and compared with control samples of healthy persons with high amounts of bacteria. Besides, in order to improve comparability between publications, it is important to always state the gDNA concentrations and volumes used. Negative controls indicating

possible shifts in taxonomical profiles are imperative. Finally, results produced by using different primer pairs cannot be easily compared.

2.3.6 Supplement

Supplementary Table 2.3.1: Generalized UniFrac dissimilarity matrix for ZIEL2 samples in comparison to the ideal composition at the genus level.

		ideal	V1-V2		V3-V4		V7-V9	
			ZIEL2-C	ZIEL2-D	ZIEL2-C	ZIEL2-D	ZIEL2-C	ZIEL2-D
ideal		0.00	0.88	0.88	0.39	0.38	1.00	1.00
V1-V2	ZIEL2-C	0.88	0.00	0.07	0.98	0.98	1.00	1.00
	ZIEL2-D	0.88	0.07	0.00	0.98	0.98	1.00	1.00
V3-V4	ZIEL2-C	0.39	0.98	0.98	0.00	0.04	1.00	1.00
	ZIEL2-D	0.38	0.98	0.98	0.04	0.00	1.00	1.00
V7-V9	ZIEL2-C	1.00	1.00	1.00	1.00	1.00	0.00	0.07
	ZIEL2-D	1.00	1.00	1.00	1.00	1.00	0.07	0.00

Supplementary Table 2.3.2: Phyla-level classification of dilution series for sample T1. In “other” the following phyla are combined*: Cyanobacteria*, *Desulfobacterota*, *Fusobacteriota*, *Verrucomicrobiota,* and unknown bacteria.

	T1-Samples, total gDNA input	*Actinobacteriota*	*Bacteroidota*	*Firmicutes*	*Proteobacteria*	other
V1-V2	60	0.11	57.17	41.54	1.13	0.05
	10	0.09	57.17	41.88	0.79	0.07
	5	0.14	60.20	38.77	0.83	0.06
	1	0.07	60.20	38.89	0.79	0.05
	0.5	0.09	61.61	37.32	0.91	0.08
	0.1	0.06	62.35	36.15	1.24	0.20
	0.05	0.01	58.50	40.98	0.45	0.06
	0.01	0.49	61.79	37.12	0.60	0.00
V3-V4	60	4.72	42.11	52.78	0.13	0.26
	10	4.77	42.12	52.84	0.04	0.23
	5	4.96	41.89	52.84	0.05	0.26
	1	4.74	42.77	51.99	0.24	0.26
	0.5	1.93	43.50	54.30	0.04	0.24
	0.1	1.67	43.91	54.26	0.03	0.13
	0.05	1.32	40.63	55.98	0.88	1.18
	0.01	1.83	13.68	79.76	3.71	1.01
V7-V9	60	2.65	23.82	72.46	0.95	0.12
	10	2.19	28.62	68.02	1.10	0.07
	5	2.10	27.41	69.50	0.89	0.10
	1	2.06	26.73	70.23	0.92	0.06
	0.5	1.93	27.26	69.83	0.91	0.08
	0.1	1.79	27.01	69.91	1.09	0.20
	0.05	1.66	27.36	69.94	0.83	0.22
	0.01	4.29	22.51	71.48	1.54	0.18

Supplementary Table 2.3.3: Phyla-level classification of dilution series for sample T30. In “other” the following phyla are combined*:* *Cyanobacteria*, *Desulfobacterota*, *Fusobacteriota*, *Verrucomicrobiota,* and unknown bacteria.

	T30-Samples, total gDNA input	*Actinobacteriota*	*Bacteroidota*	*Firmicutes*	*Proteobacteria*	other
V1-V2	60	1.88	41.32	55.32	1.14	0.33
	10	1.69	43.42	53.57	0.98	0.35
	5	1.84	42.24	54.71	0.90	0.32
	1	1.49	42.32	54.93	0.94	0.32
	0.5	1.45	43.14	54.13	1.01	0.27
	0.1	0.85	46.92	50.98	0.89	0.35
	0.05	0.92	40.98	54.88	2.64	0.58
	0.01	1.64	39.06	57.98	1.26	0.06
V3-V4	60	12.59	28.06	57.22	1.27	0.86
	10	10.23	29.01	58.97	1.14	0.65
	5	11.74	28.61	58.01	0.96	0.68
	1	11.27	31.66	55.37	1.09	0.62
	0.5	4.42	31.97	61.90	0.97	0.75
	0.1	3.77	30.31	64.08	1.15	0.68
	0.05	5.35	35.43	57.73	0.86	0.62
	0.01	6.34	33.56	58.13	0.89	1.08
V7-V9	60	6.58	22.18	69.69	1.13	0.43
	10	5.72	23.50	69.12	1.12	0.54
	5	5.28	23.18	70.05	1.02	0.47
	1	5.19	22.95	70.27	1.13	0.45
	0.5	5.26	23.39	69.94	0.93	0.48
	0.1	2.58	18.90	75.41	2.26	0.85
	0.05	6.07	29.13	62.43	1.99	0.38
	0.01	4.87	21.17	72.01	1.47	0.48

2.4 Cell-counting in human fecal samples comparing flow cytometry versus spike-in standards in 16S rRNA gene sequencing

Drafted manuscript

2.4.1 Abstract

Synthetic spike-in standards in 16S rRNA gene sequencing approaches are used to normalize bacterial abundances. Thus, quantification of bacterial abundances between samples can be achieved. This is especially of interest for clinically relevant samples, since changes in the relative and absolute abundances of various phyla can be an important medical parameter. Here, we wanted to assess how the synthetic spike-in approach (Tourlousse *et al.*, 2017) performs and compared it with quantification of actual cell counts using flow cytometry. Moreover, we evaluated whether the addition of the synthetic spike-in DNA led to a taxonomic shift or biased the subsequent 16S rRNA gene sequencing analysis. We could successfully demonstrate that the synthetic spike-in control produced conclusive results showing the same trends as observed in flow cytometry. Further, the synthetic spike-in does not influences the initial sample composition as comparison of samples processed without synthetic and with the spike-in standard produced highly correlated results, suggesting that the data originating from the same sample clustered in close proximity (i.e., insignificant differences). Summed up, we confirm that an easy-to-follow spike-in protocol produces results comparable to an independent method (flow cytometry) and allows convenient analysis of spike-normalized bacterial abundances in tested samples.

2.4.2 Introduction

The human gut microbiota is a diverse and complex microenvironment composed of 10^{14} microorganisms (Thursby and Juge, 2017). The composition of this dynamic and changing environment is mostly studied using 16S rRNA gene sequencing approaches. Synthetic spike-in standards are commonly used in RNA, or genome sequencing approaches as internal standards for measuring technical biases and as quantitative standards (Blackburn *et al.*, 2019, Hardwick *et al.*, 2018, Jiang *et al.*, 2011, Tembe *et al.*, 2014, Venkataraman *et al.*, 2018). Further, synthetic spike-in methods for 16S rRNA gene sequencing increase in popularity. Tourlousse *et al.* (2017) developed a set of synthetic 16S rRNA genes, which can be used as a universal standard for 16S rRNA gene sequencing experiments. By using such a set, they could show that spike-normalized OTU abundances could be determined, and therefore the dynamics within the microbiota can be analyzed. Despite the synthetic approach, Stämmler *et al.* (2016) described a method using uncommon (i.e., non-gut) bacterial species

Salinibacter ruber, *Rhizobium radiobacter*, and *Alicyclobacillus acidiphilus* as a spike-in marker instead of artificial 16S rRNA-like DNA. By pooling those bacteria in defined amounts and introducing it to the tested sample before DNA extraction, Stämmler *et al.* (2016) could show that microbial loads of the original sample can be determined. Spike-in standards, however, allow for determining spike-normalized bacterial abundances, and thus the source of taxonomic changes over time can be explained. Differences in bacterial loads could be due to antibiotic treatments or disease status (Martínez, 2017). Inflammatory bowel disease (IBD), which is characterized by chronic inflammation of the digestive tract (Rubin *et al.*, 2012), consists of two subtypes, Crohn's disease (CD) and ulcerative colitis (UC). The difference between those two is their clinical appearance (symptoms and disease location). While CD is associated with full-thickness inflammation and is located at any site of the gastrointestinal tract, UC is usually limited to the inflammation of the mucosal layer of the colon (Panaccione, 2013). IBD is characterized by phases of relapsing and remitting, where inflammation still persists and can further lead to epithelial injury, inducing lifelong morbidity (Atreya *et al.*, 2020). Therefore, the main goal is to treat IBD, mostly using optimized anti-inflammatory therapies and therapies that modulate the immune system. Often used therapies include, for example, the use of 5-Aminosalicylates, Corticosteroids, Thiopurines, Methotrexate, Janus kinase inhibitors, the use of different antibiotics or probiotics, enteral nutrition, or fecal transplantations (Bernstein, 2015). Even though IBD does not show a clear etiology and is difficult to diagnose, sometimes it seems to be associated with lower microbial α-diversities compared to healthy controls (Pascal *et al.*, 2017). Moreover, IBD was linked to a reduced species richness and evenness (Gevers *et al.*, 2014, Michail *et al.*, 2012). Here, we compared bacterial loads of human samples originating from donors diagnosed with either CD or UC and compared these to healthy controls by using synthetic spike-in 16S rRNA gene sequencing and cell counting by flow cytometry.

2.4.3 Material and Methods

Collection and storage of human stool samples

Human stool samples were collected as part of the Biotherapy cohort, described by Metwaly (2020). The test set for this study included samples from six human donors diagnosed with IBD (2 or 3 time-points per person) and six healthy controls (see Table 2.4.1). Fecal samples were stored in 50% (v/v) glycerol at -80°C.

Table 2.4.1: Overview of samples included in this study.

Sample	Disease phenotype	Number of samples at collected time-points (T)
CD1	CD	3
CD2	CD	3
CD3	CD	3
UC1	UC	2 (no T1)
UC2	UC	3
UC3	UC	2 (no T3)
Control 1	Healthy	1
Control 2	Healthy	1
Control 3	Healthy	1
Control 4	Healthy	1
Control 5	Healthy	1
Control 6	Healthy	1

Preparation of human fecal samples

About 100-300 mg per sample were transferred into a clean and sterile Eppendorf tube. This was performed in duplicates and one sample was used for flow cytometry and one for the spike-in experiment. Fecal samples were thawed, centrifuged, the supernatant discarded, and the pellet was resuspended in 2 ml PBS. Afterward, samples were thoroughly vortexed. Subsequently, they were then either processed for the spike-in 16S rRNA gene sequencing or prepared for flow cytometric analysis.

Preparation of spike-in control

The design and synthesis of the spike-in sequence were performed and adopted as described by Tourlousse *et al.* (2017). The Spike-in sequences Bv5501, Ca5501, Ga5501, Tb5501, Ec5001, Ec5002, Ec5003, Ec5004, Ec5005, Ec5501, Ec5502, and Ec6001 were ordered at BioCat GmbH. Plasmid vectors, including the genes, were introduced in *E. coli* via electroporation. Plasmid isolation from fresh overnight cultures was performed using the Sigma Gene Elute Plasmid Miniprep Kit. Extracted plasmids were verified using Sanger sequencing, DNA concentrations were recorded, and a pool containing the equimolar ratio of each of the twelve constructs was created. The pool was then digested using Hind III. For this, a reaction mixture containing 7 µg plasmid-pool, 70 U restriction enzyme, 10x NEBuffer, and water in a 700 µl was set up, incubated at 37°C for 1 h, and afterwards heat-inactivated at 80°C for 20 min. The digested pool was purified using 1.8x AMPure XP beads (Beckmann Coulter), the concentration was measured, and pools with a final concentration of 2 ng/µl were prepared and stored at -20°C.

Spike-in 16S rRNA gene sequencing

First, 600 µl of each sample resuspended in PBS were transferred into a clean Eppendorf tube. Then, 6 ng spike-in control was added to each sample. The samples were then transferred

into bead-beating tubes, and DNA extraction was performed as previously described by Reitmeier *et al.* (2020). Library preparation, targeting the V3-V4 region, and sequencing was performed as previously described by Abellan-Schneyder *et al.* (2021a).

Spike-in downstream analysis

Data without spike-in added were processed using an updated version of the Integrated Microbial Next-generation sequencing (IMNGS) pipeline, originally published by Lagkouvardos *et al.* (2016). This in-house developed pipeline is based on UPARSE (Edgar, 2013). In the advanced IMNGS options, allowed mismatches were set to one. Demultiplexing was performed using a minimum read-length of 250 bp and a maximum read-length of 600 bp. Forward trim was set to 30 bp, and reverse trim length was 60 bp. The abundance filter was set to 0.0025 (Clavel *et al.*, 2020), and the filter of low-read samples was set to “off”. Chimeric reads were removed using UCHIME (Edgar *et al.*, 2011), and zero-radius operational taxonomic units (zOTUs) were produced using USEARCH v11.0.667.

For the spike-in sequence analysis, the Namco tool was used (Dietrich *et al.*, 2021, unpublished). In a first step, the FASTQ files were mapped against the reference spike-in FASTA files. Here, Bowtie2 v2.3.4.3 was used (Langmead and Salzberg, 2012, Langmead *et al.*, 2018). The output FASTQ files after mapping included R1/R2 FASTQ files without the spike-in reads and R1/R2 FASTQ files of the spike-in sequences. OTU normalization is performed sample-wise and is based on the sum of spike reads for each sample. This allows the calculation of spike-normalized bacterial abundances.

Further analysis was performed in Rhea (Lagkouvardos *et al.*, 2017). Rhea is a collection of R-scripts enabling comparison between samples. After normalization of data, alpha- and beta-diversities can be visualized. The script also performs taxonomic binning, enabling an insight on all known and unknown sequences of the microbial composition down to the genus level.

Fecal flow cytometry

Two milliliter sample (resuspended in PBS) was filtered using a syringe filter (5 µm pore size) to remove solid fecal content. Then, the remaining sample was transferred into a Fast Prep Lysing Matrix D tube (MP Biomedicals) containing ceramic beads and incubated on ice for 1 h. Fecal samples were homogenized well using a vortexer for a few minutes. Centrifugation was carried out at 50×*g*, for 15 min at 4°C to remove large particles. The complete supernatant was transferred into a fresh sterile, pre-labeled Eppendorf tube. Then, 900 µl staining buffer (sterile PBS containing 2% FCS) was added. The SYTO 9 nucleic acid stain (Thermo Fisher Scientific Inc.) was defrosted on ice. Samples were well vortexed and centrifuged for 5 min (8,000×*g*, 4°C). The supernatant was discarded, and pellets were resuspended in 1 ml staining buffer. Samples were washed (8,000×*g*, 10 minutes, 4°C) and resuspend afterwards, according to

the pellet size, in 500-1,000 µl staining buffer. The OD_{600} was recorded in the Cell Culture mode using a Nanodrop (Thermo Fisher Scientific Inc.). Based on the recorded concentrations, sample volumes were adjusted. Staining solution including SYTO 9 was prepared (1:1,000 dilution in staining buffer), and 100 µl of this staining solution were added to a sterile V-shaped 96-wells plate. Next, 100 µl of the bacterial suspension (OD_{600} around 0.1) was added and the plate was incubated for 10 min at room temperature in the dark (the plate was covered with aluminum foil). Samples were then washed with 200 µl staining buffer (2,250×*g,* 10 min, 4°C), and pellets were resuspended in 300 µl staining buffer before flow cytometric analysis. The filters were set accordingly: FSC: 550 nm, SSC: 250 nm, FITC: 455 nm, PE: 500 nm, APC: 450 nm. Flow cytometry was performed using a BD LSR II machine. Samples were recorded for 120 s, measuring all events. Analysis of experiments was performed using FlowJo v10.
To calculate the number of bacteria per ml present in the analyzed samples, reference beads of 6 µm size was used (Sphero blank calibration beads, BD Biosciences). The Formula 2.4.1 used to calculate the concentration of bacteria detected was adapted from Ou *et al.* (2017):

$$\text{concentration of bacteria} = \frac{(\#\text{ number of events in bacterial region})}{(\#\text{ of events in bead region})} * \text{concentration of beads} * \text{dilution factor} \qquad (2.4.1)$$

2.4.4 Results

Comparability of standard 16S rRNA gene sequencing and spike-in 16S rRNA gene sequencing

In a first experiment, it was assessed whether the 16S rRNA sequencing profiles produced by standard 16S rRNA gene sequencing (without spike-in control) are comparable to those spiked-in and analyzed after spike-in removal. Here, we could show that clustering is due to sampling origin, i.e., samples originating from the same human donor cluster together and not samples prepared using the same protocol (Fig 2.4.1A). Moreover, comparability is given, as differences in the non-metric multi-dimensional scaling (NMDS) plot are small and insignificant (Fig 2.4.1B).

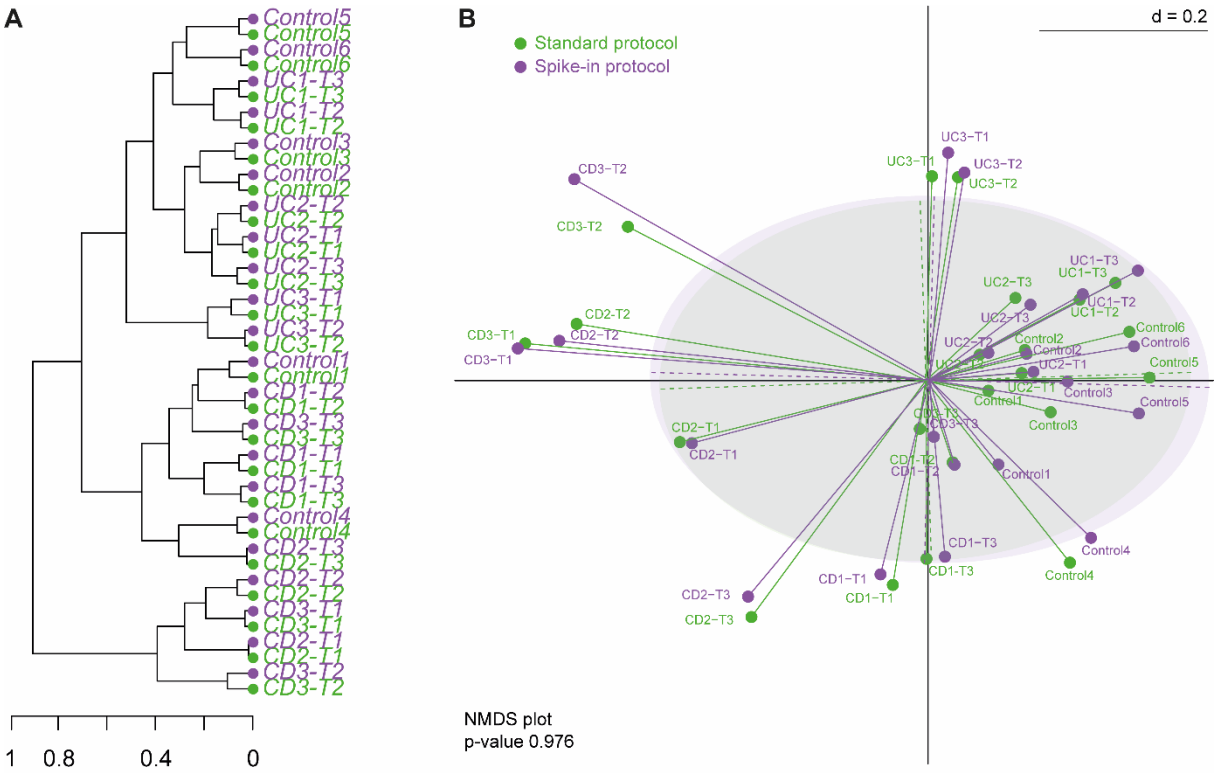

Figure 2.4.1: Beta-diversity analysis of human samples either processed by the standard 16S rRNA gene sequencing protocol (Standard protocol, green) or the spike-in 16S rRNA gene sequencing protocol (Spike-in protocol, violet). Hierarchical clustering of the samples in a phylogenetic tree (**A**) and meta NMDS plot calculated from the generalized UniFrac dissimilarity matrix (**B**) show that clustering is due to sampling origin (human donor) and not due to the method used.

Comparability of spike-in 16S rRNA gene sequencing and bacterial cell count numerations via flow cytometry

Afterward, it was estimated whether cell counts determined by using either a flow cytometry approach or the spike-in 16S rRNA gene sequencing approach correlate. Therefore, a total of 20 samples (6 clinical fecal samples measured at two or three different time points (T1, T2 and T3) and 6 healthy control samples) were processed by flow cytometry and sequenced after synthetic spike-in (Fig. 2.4.2). The estimation of bacterial cells per ml using the flow cytometry approach was facilitated by using counting beads, allowing an indirect measurement of bacterial cells. For the spike-in approach, the number of bacteria was indirectly determined through the defined amount of spike-in gDNA added in comparison to the total sample volume and the generated read output corresponding either to the synthetic spike-in or the real gDNA content.

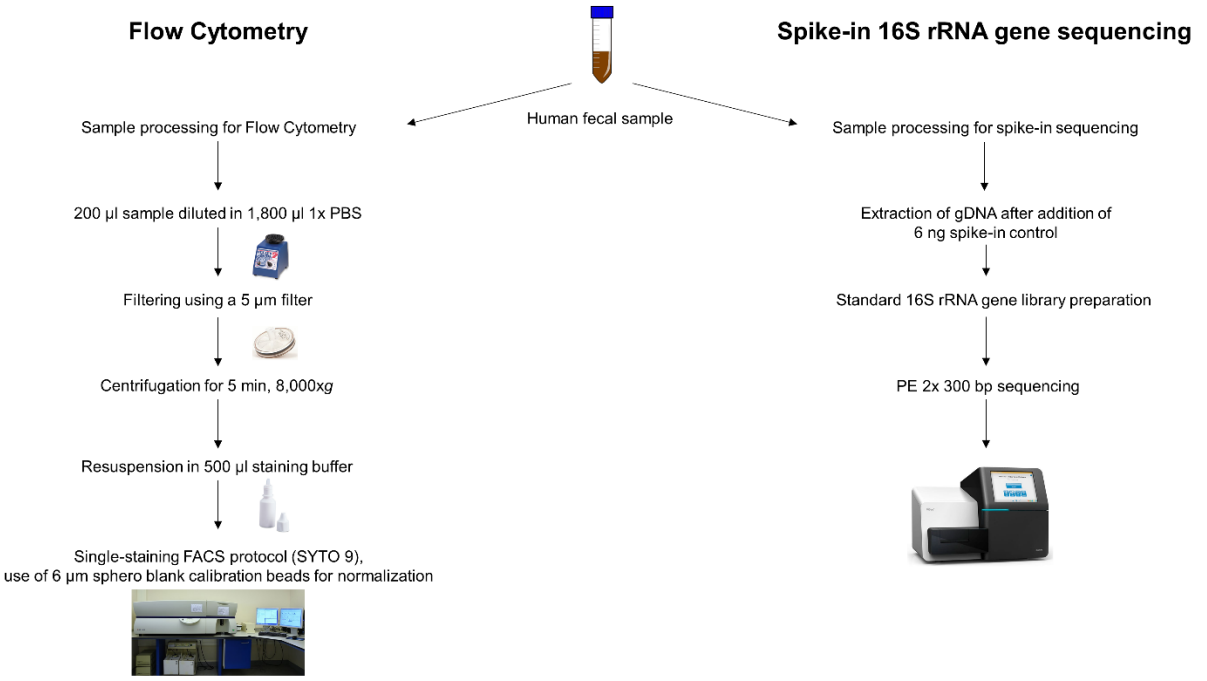

Figure 2.4.2: Overview of the workflows for flow cytometry and spike-in 16S rRNA gene sequencing. Methods were compared, and it was assessed whether a similar concentration of bacteria in the samples could be detected.

In a first step, control samples were assessed via flow cytometry, and it was checked whether bacterial counts could be determined (Fig 2.4.3). Flow cytometric measurements are based on different properties in the recorded light scatters and fluorescent properties. Forward (FSC) and side scatters (SSC) allow discrimination of populations due to cell size. Thus, counting beads, which had a size of 6 µm cluster apart from bacterial cells which are much smaller as, e.g., *E. coli* has an average size of 2 µm. Staining with SYTO 9 allows differentiating between cell-debris and intact cells. SYTO 9 has a similar emission spectrum to fluoresceine (FITC) when it is bound to nucleic acids and can therefore be easily recorded with most machines (see Fig 2.4.3C unstained vs. 2.4.3D stained).

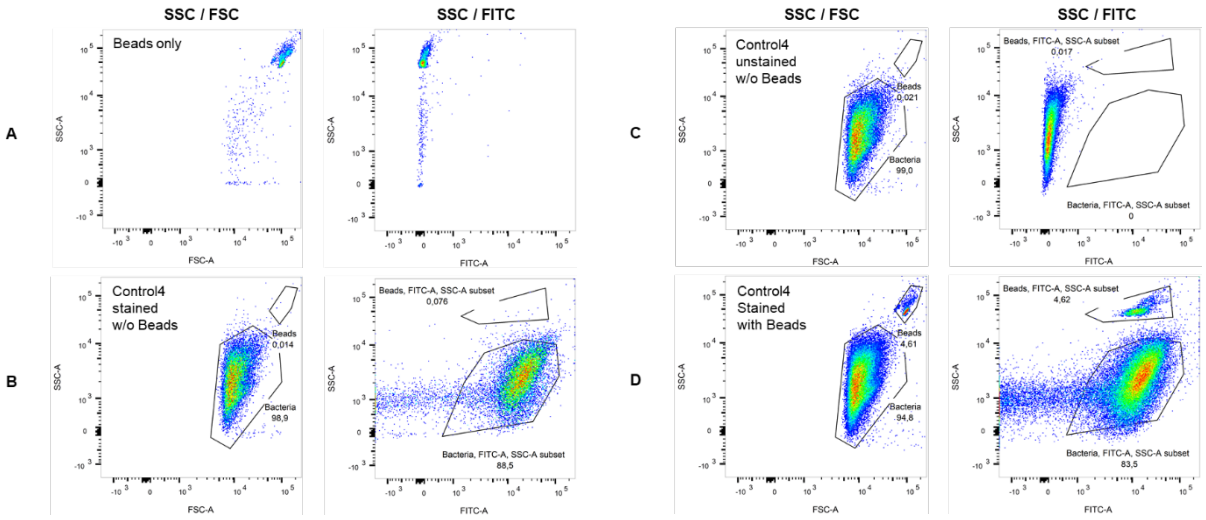

Figure 2.4.3: Control experiments for the flow cytometry approach. The number of bacteria/ml was indirectly measured through the defined amount of added counting beads. Counting beads with PBS (**A**), and several control experiments of Control sample 4 were used for the initial gating. First, stained samples without beads were used to form the bacterial gate (**B**), the unstained sample was used to gate the FITC bacterial section (**C**), and the stained and bead added sample was used for later control of correct gating (**D**).

After measuring those initial control samples, gating was performed, and all other regular samples and controls were measured via flow cytometry (Fig. 2.4.4 - 2.4.7) for 120 s while recording all events. All samples were later analyzed using FlowJo, the number of bacteria detected was noted, and the amount of bacteria/ml was calculated using Formula 2.4.1 (see Table 2.4.2).

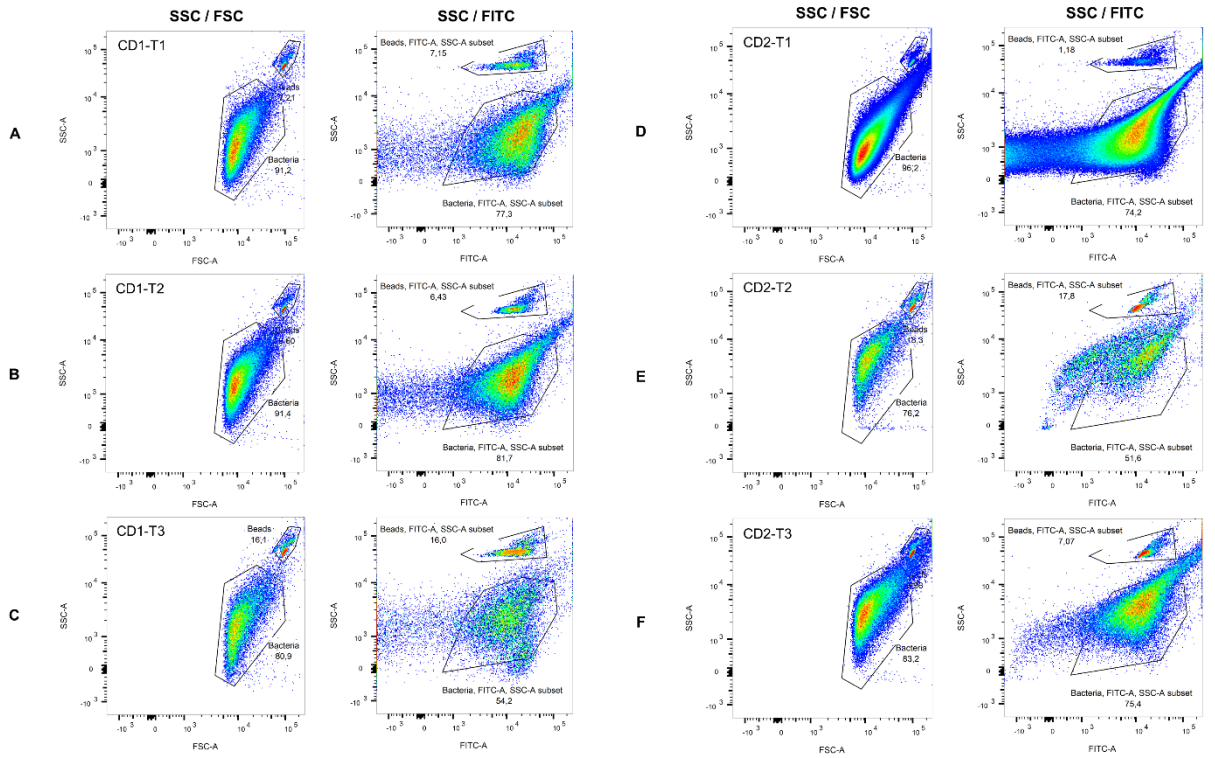

Figure 2.4.4: Analysis of the number of bacterial cells detected within human samples via flow cytometry. **A**: Human sample CD1-T1, **B**: Human sample CD1-T2, **C**: Human sample CD1-T3, **D**: Human sample CD2-T1, **E**: Human sample CD2-T2, **F**: Human sample CD2-T3.

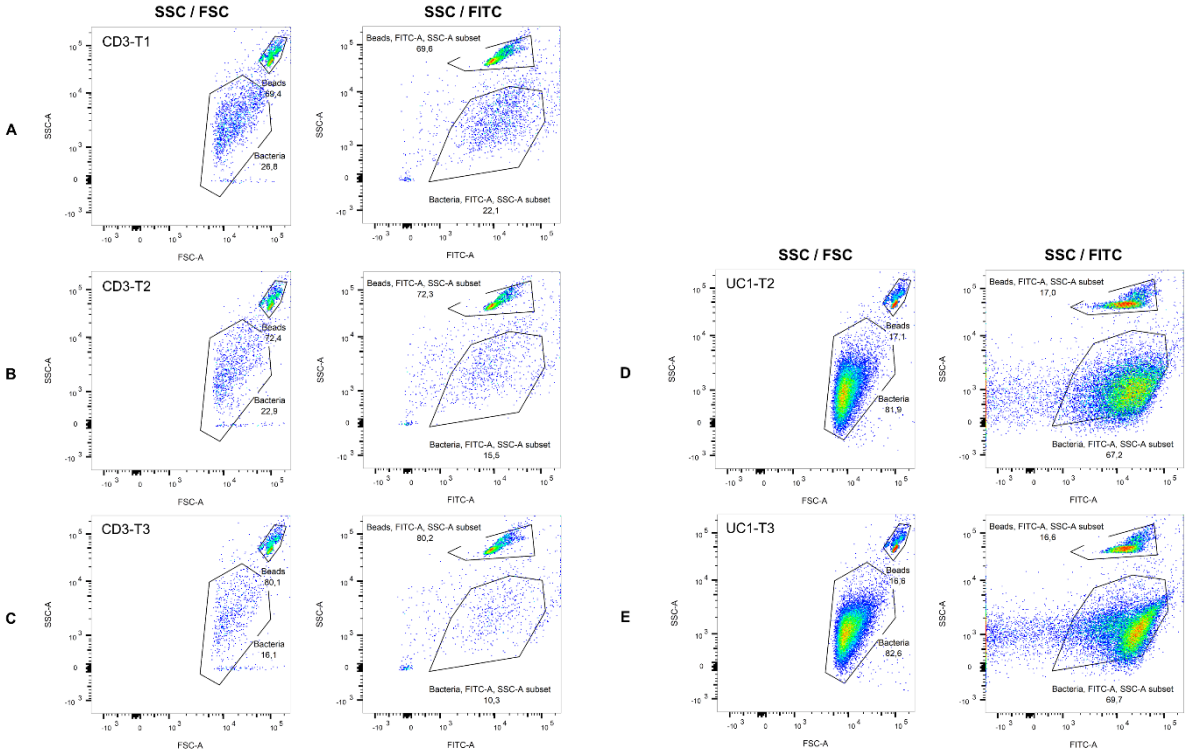

Figure 2.4.5: Analysis of the number of bacterial cells detected within human samples via flow cytometry. **A**: Human sample CD3-T1, **B**: Human sample CD3-T2, **C**: Human sample CD3-T3, **D**: Human sample UC1-T2, **E**: Human sample UC1-T3.

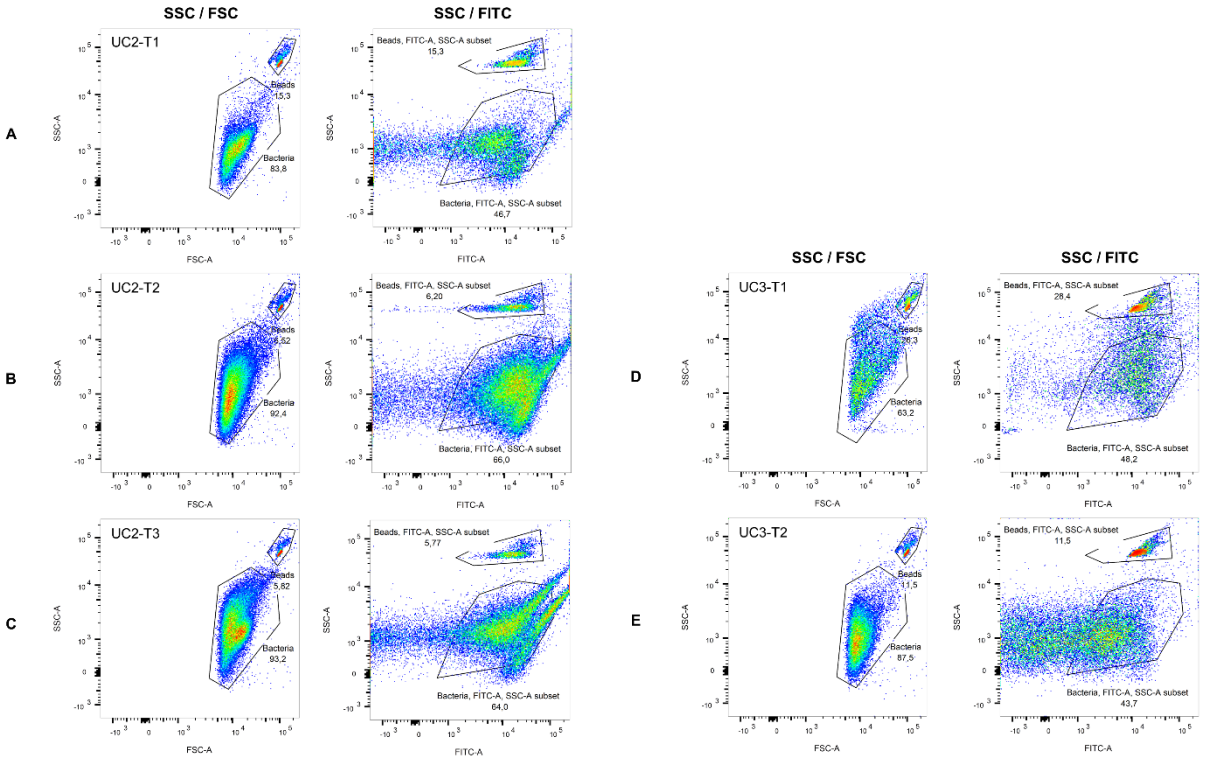

Figure 2.4.6: Analysis of the number of bacterial cells detected within human samples via flow cytometry. **A**: Human sample UC2-T1, **B**: Human sample UC2-T2, **C**: Human sample UC2-T3, **D**: Human sample UC3-T1, **E**: Human sample UC3-T2.

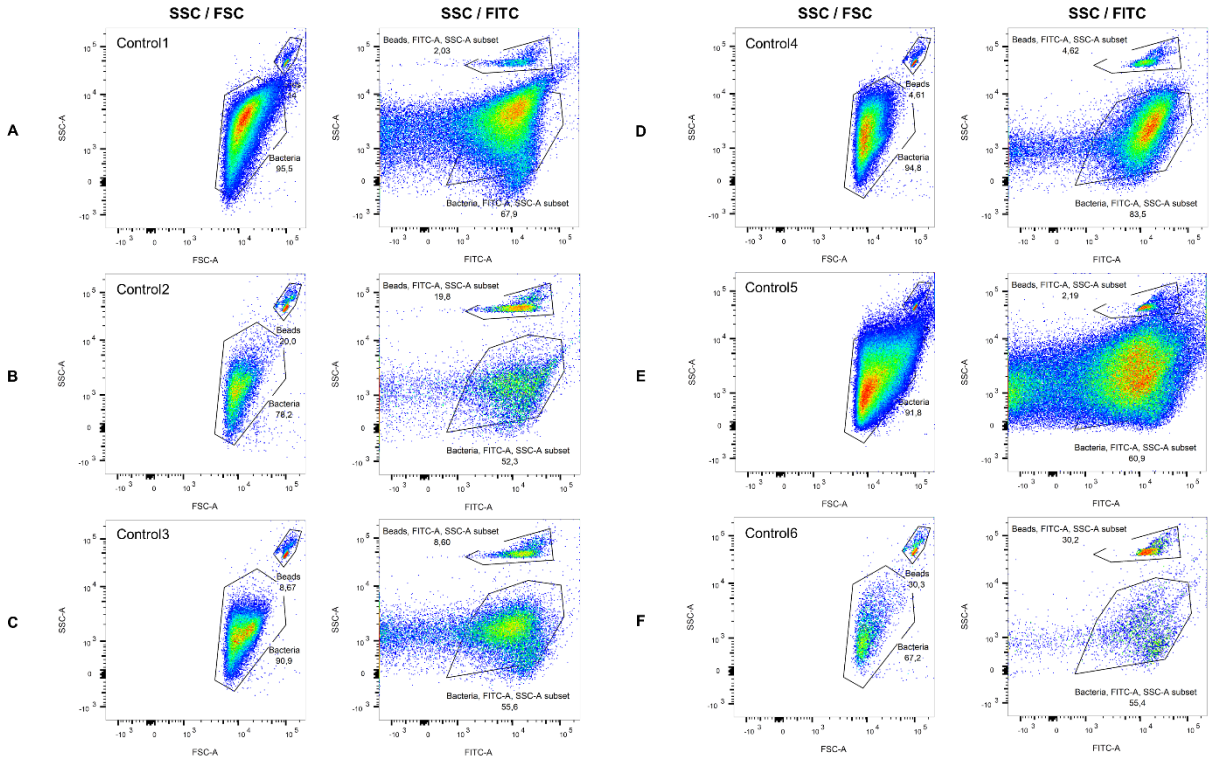

Figure 2.4.7: Analysis of the number of bacterial cells detected within human samples via flow cytometry. **A**: Human control sample 1, **B**: Human control sample 2, **C**: Human control sample 3, **D**: Human control sample 4, **E**: Human control sample 5, **F**: Human control sample 6.

Table 2.4.2: Overview of the cell count measures using flow cytometry. The total numbers of counts were detected for 120 s (all events).

Sample	Counts	Bacteria, FITC-A, SSC-A (Counts)	Beads, FITC-A, SSC-A (Counts)	Dilution	Bacteria/ml
CD1-T1	43,102	33,322	3,081	1.0	1.0E+09
CD2-T2	49,979	40,816	3,216	1.8	2.2E+09
CD1-T3	19,914	10,794	3,178	1.0	3.2E+08
CD2-T1	278,346	206,549	3,293	1.0	5.9E+09
CD2-T2	17,458	9,006	3,105	1.0	2.7E+08
CD2-T3	52,555	39,614	3,717	1.0	1.0E+09
CD3-T1	6,106	1,349	4,248	1.0	3.0E+07
CD3-T2	4,362	677	3,152	1.0	2.0E+07
CD3-T3	4,353	447	3,493	1.0	1.2E+07
UC1-T2	19,628	13,184	3,331	2.1	7.8E+08
UC1-T3	23,566	16,437	3,919	1.3	5.2E+08
UC2-T1	19,468	9,087	2,969	1.0	2.9E+08
UC2-T2	49,818	32,889	3,091	1.6	1.6E+09
UC2-T3	37,927	24,288	2,188	2.1	2.2E+09
UC3-T1	12,711	6,124	3,615	1.0	1.6E+08
UC3-T2	24,056	10,521	2,760	2.3	8.3E+08
Control1	86,158	58,532	1,750	3.5	1.1E+10
Control2	15,851	8,290	3,141	7.5	1.9E+09
Control3	31,786	17,673	2,734	10.0	6.1E+09
Control4	51,543	43,033	2,382	10.0	1.7E+10
Control5	128,725	78,448	2,822	58.8	1.5E+11
Control6	6,044	3,346	1,823	15.0	2.6E+09

Interestingly, the number of bacteria per ml was on average higher for the control samples with an average of 3.1×10^{10} bacteria/ml versus 1.1×10^{9} for the clinical samples. Control sample 5 showed the overall highest numbers of bacteria/ml with 1.5×10^{11} bacteria/ml. In contrast, sample CD3-T3 was recorded with the least number of bacteria corresponding to approximately 1.2×10^{7} bacteria/ml. In a second step, it was assessed whether the same trends observed by measuring the bacteria/ml via flow cytometry could be confirmed by spike-in sequencing.

Again, we found that Control sample 5 showed the overall highest spike-normalized bacterial abundances when spike-in samples were analyzed (Fig 2.4.8B). Even though different overall numbers were calculated and the resulting measures (bacteria/ml vs. 16S rRNA gene copy/ng) are not comparable *per se*, we could show that both flow cytometry and spike-in results have a strong Pearson correlation of R=0.95 (R^2=0.90) (Fig 2.4.8). Nonetheless, it should be noticed that this strong correlation is significantly dependent on Control sample 5. If this sample would be excluded, a Pearson correlation of R=0.82 (R^2=0.67) is observed.

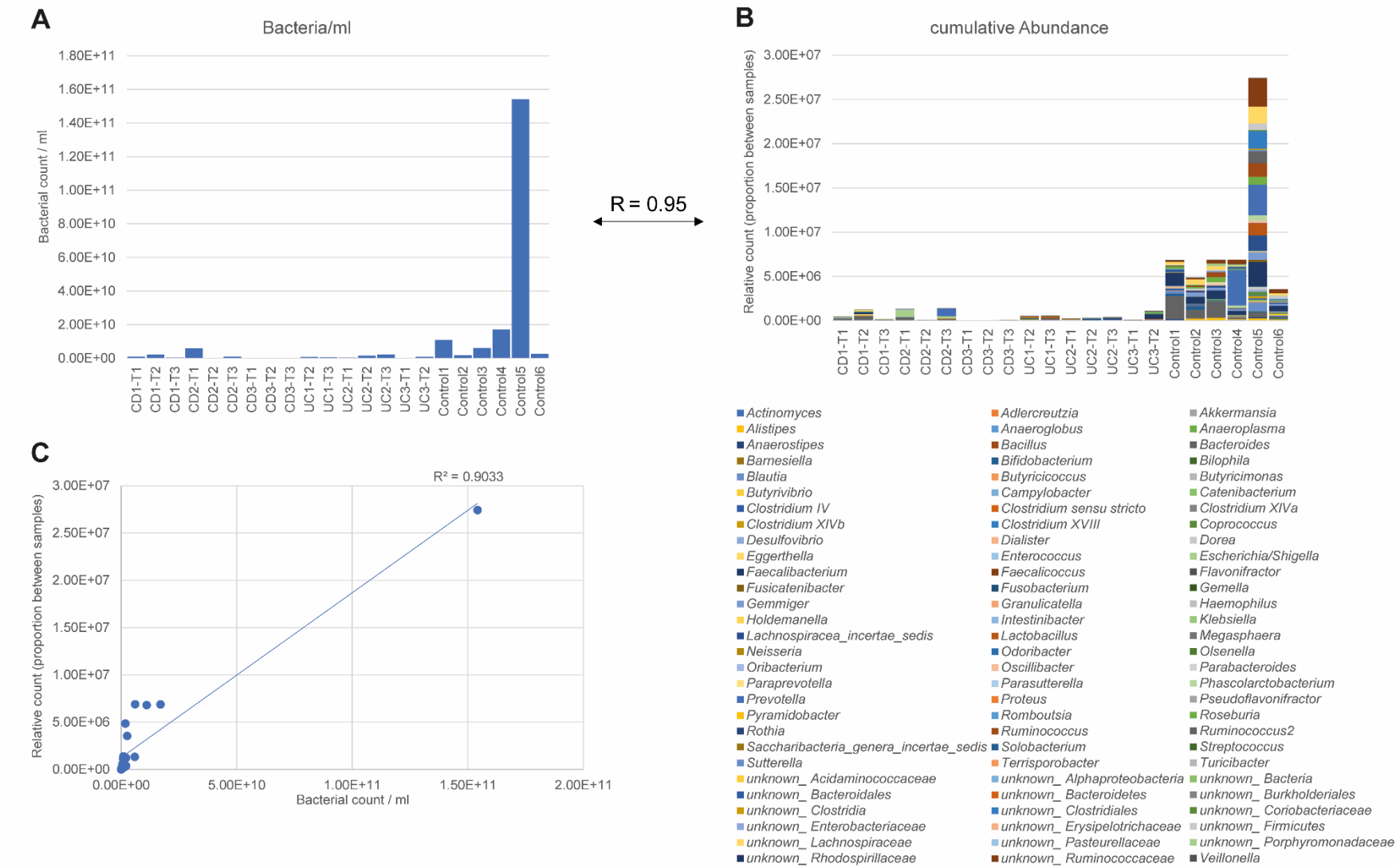

Figure 2.4.8: Comparison of the number of bacteria detected by flow cytometry (**A**) or spike-in sequencing analysis (**B**). Pearson correlation coefficient was found to be R=0.95, and R^2=0.90 was calculated (**C**).

When quantitative profiles are compared to standard 16S rRNA gene sequencing abundances, markedly different genera abundance became noticeable (Fig 2.4.9). For example, when analyzing sample CD3-T2, 52% of all detected reads could be accounted for *Haemophilus*.

2.4.5 Discussion

In this proof-of-principle study, we showed that synthetic spike-in sequencing allows users a convenient analysis of spike-normalized bacterial abundances within different samples. Thus, it can be assessed whether the bacterial load differs significantly from one sample to another, which can be of great interest, especially when clinically relevant samples (e.g., after fecal transplantation, antibiotic use, in case of IBD or other intestinal diseases) are compared to control groups. The spike-in sequencing approach for 16S rRNA gene sequencing, published by Tourlousse *et al.* (2017), was used here to assess whether microbial loads of given samples could be estimated. Further, the resulting spike-normalized bacterial abundances were compared to the results gained through flow cytometric cell counting.

Flow cytometric estimations of bacterial loads are based on the ability to count cells through light scattering and fluorescent-based detection (Wilkinson, 2018). Several concrete protocols for the preparation of environmental samples for flow cytometry were previously described, e.g., Ou *et al.* (2017), Brown *et al.* (2019), Frossard *et al.* (2016), or Bellali *et al.* (2019). Nevertheless, to our knowledge, synthetic spike-in sequencing for 16S rRNA gene sequencing was not yet compared to flow cytometry. On this account, it should be checked whether the same trends, e.g., samples showing the highest / lowest bacterial count, could be identified independently by those two methods. Indeed, we could show that trends generated by both protocols are similar. This is of interest, as spike-in 16S rRNA gene sequencing is easy to perform and does not bias the general 16S rRNA gene sequencing analysis, as no divergence from samples prepared with or without spike-in could be detected (see Fig 2.4.1). Moreover, flow cytometry is a highly complex and work-intensive measuring technique that needs a trained scientist to be performed. However, flow cytometry can be used to isolate certain bacterial populations if cell sorting is applicable, and specific dye combinations can differentiate between viable and non-viable bacteria, e.g., the Live/Dead® BacLight™ system by Thermo Fisher Scientific (Duquenoy *et al.*, 2020, Emerson *et al.*, 2017),which is not possible using spike-in controls. Vandeputte *et al.* (2017) described a parallelization and thus a combined approach of amplicon sequencing and flow cytometry that allowed them to study if a disease's phenotype was associated with a reduced bacterial load. Indeed, they found that patients with Crohn’s disease showed a reduced bacterial load, which was not noticeable when only simple 16S rRNA gene sequencing was performed.

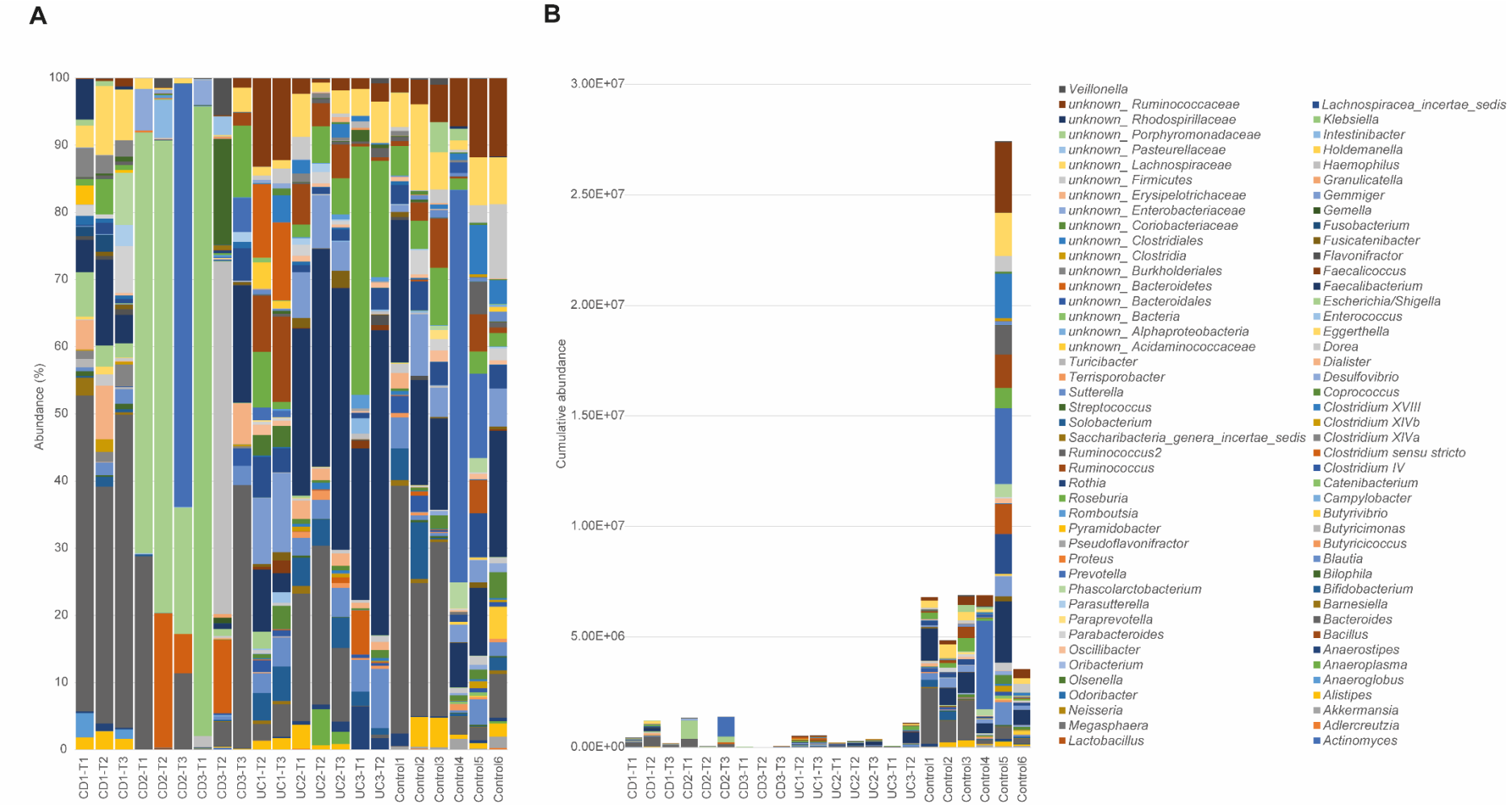

Figure 2.4.9: Relative (**A**) versus cumulative abundance (**B**) of 16S rRNA gene sequencing profiles at the genus level. Relative microbiota profiles were generated after spike-in sequencing removal. Cumulative abundances were created by normalization based on spike-in reads.

However, we could demonstrate that spike-in sequencing allowed us to show the reduced bacterial load in IBD patients compared to the healthy control groups (see Fig 2.4.9) without performing flow cytometric analysis, even though flow cytometric analysis confirmed the results. Thus, a reduction in hands-on time was achieved, which will probably allow a convenient analysis of spike-normalized bacterial abundances in a broader range of study set-ups. Nevertheless, drawbacks, e.g., careful handling and precise laboratory practice (Harrison *et al.*, 2021) must be taken into account. Furthermore, caution must be taken when low biomass samples are processed, as then the spike-in amount must be decreased to prevent drowning of the target reads in too many spike reads in the sample output.

2.4.6 Conclusion

Here, we compared two independent analysis strategies to assess the microbial load of a given sample. On the one hand, we evaluated the number of bacteria present within a given sample through flow cytometry and, on the other hand, by synthetic spike-in sequencing. By the latter, spike-normalized bacterial abundances can be compared between samples which will be interesting, especially if clinically relevant samples are analyzed. Moreover, as differentially abundant taxa can be analyzed and compared between samples with different overall microbial loads, this analysis strategy will facilitate to study microbial dynamics within the samples analyzed. As the method is easy to be performed and resulted in highly correlated taxonomic profiles when compared to the standard 16S rRNA gene sequencing approach, we suggest that spike-in sequencing should be used regularly.

2.5 Synthetic versus long-read 16S rRNA sequencing approaches – benefits, drawbacks, and feasibility

2.5.1 Abstract

Full-length sequencing approaches are advertised for allowing an improved species-level classification compared to standard short amplicon sequencing. In this study, it was tested whether the species-level classification indeed is improved by full-length 16S rRNA gene sequencing approaches. Up to date, several different such approaches exist, which are either categorized as long-read sequencing approaches or as synthetic full-length approaches. In this study, we compared both a long-read sequencing approach and two synthetic full-length approaches to standard short amplicon sequencing. For this, two mock communities and a human fecal sample were used.

Overall, it was assessed whether (i) different sequencing approaches using the same fecal samples from the same human donor could generate comparable taxonomical profiles, (ii) if the species-level classification was improved by the full-length sequencing approaches, and (iii) if either long-read or synthetic full-length approaches performed comparably better.

We found that if the LoopSeq synthetic full-length sequencing approach was compared to short amplicon sequencing approaches, comparable taxonomical profiles could be generated, and the species-level classification was improved for the full-length approach. Nevertheless, the MinION long-read sequencing approach and the in-house synthetic full-length sequencing approach need further improvements to be competitive and efficient.

2.5.2 Introduction

Sequencing of the complete 16S rRNA gene has the benefits of the higher taxonomical resolution, potentially even sub-species classification, e.g., as shown by Johnson *et al.* (2019). Further, a reduced primer bias is found when compared to short amplicon sequencing, were typically one, two, or three adjunct variable regions (V-regions) are sequenced.

Generally, full-length 16S rRNA gene sequencing can be either performed by long-read sequencing approaches typically using Oxford Nanopores Technologies (ONT) MinION or by Pacific Biosciences (PacBio) circular-consensus sequencing (CSS) approach. Earl *et al.* (2018) showed, for example, that by using the CSS technology, they could greatly improve their taxonomic and phylogenetic resolution for their human sinonasal microbiome samples. Shortly afterward, Callahan *et al.* (2019) showed that combining the PacBio CSS

sequencing approach with the DADA2 sequence processing pipeline (Callahan *et al.*, 2016) resulted in a single-nucleotide resolution and a near-zero error rate. MinION sequencing is less expensive than CSS strategies. Nevertheless, the error rates, even though they could be reduced dramatically in the last years, are still considerably high (Kono and Arakawa, 2019). Nevertheless, several recent publications demonstrated that by using ONTs long-read sequencing approach for 16S rRNA gene sequencing, the taxonomical resolution was improved compared to short amplicon sequencing (e.g., Matsuo *et al.*, 2021, Nygaard *et al.*, 2020). Nonetheless, caution must be taken and results have to be evaluated carefully as Winand *et al.* (2019), for example, found that up to 40% of their mock community reads produced by MinION long-read strategies were misclassified at the species level. To improve and guarantee a precise and reliable species-level classification, Karst *et al.* (2021) described a method that combines unique molecular identifier (UMIs) and long-read sequencing approaches. Using this approach, they were able to produce single-molecule consensus sequences with a low chimera rate (Karst *et al.*, 2021).

Besides long-read sequencing approaches, synthetic full-length 16S rRNA gene sequencing approaches gained interest within the last years. The first protocols, described by Burke & Darling (2016) and Karst *et al.* (2018), produce full-length 16S rRNA amplicons that are later fragmented, sequenced on a short read sequencer (Illumina), and then *de novo* assembled to the full-length. Important for this technique is the use of UMIs by which the fragments are identified and sorted. Due to the use of a second-generation sequencing device, sequencing error rates are comparably low, and this methodology is applicable to a wide variety of laboratories, as the Illumina sequencers are still the most used and widespread sequencers.

In this project, long-read, synthetic full-length, and short amplicon 16S rRNA gene sequencing strategies are tested and compared. For the long-read approach, we used ONTs MinION device. For the synthetic approach, the commercially available LoopSeq Kit (LoopGenomics) and an in-house sequencing strategy were applied; and for comparison, short amplicon sequencing of the V3-V4 and the V1-V2 region was performed. Here, we used human fecal samples of healthy donors and mock communities of different complexity.

2.5.3 Material and Methods

2.5.3.1 General Material & Methods

Preparation of human gut samples

Stool samples were obtained from healthy volunteers of age after informed and written consent. An ethics approval is deemed unnecessary according to the statement given in the Drucksache 15/2849 of the German Bundestag about § 41 Abs. 2 Nr. 2 S. 1 and 2 Arzneimittelgesetz. Stool samples were collected in stool sample tubes as described previously by Abellan-Schneyder *et al.* (2021a).

Extraction of gDNA from stool samples

Genomic DNA was isolated using a modified protocol by Godon *et al.* (1997), as described previously by Reitmeier *et al.* (2020) and Abellan-Schneyder *et al.* (2021a).

Extraction of gDNA from mock communities

DNA of the Zymo mock community was purchased as a ready-to-use DNA mock (D6306, Zymo Research). For the more complex ZIEL2 mock community, preparation and extraction were performed as described in Abellan-Schneyder *et al.* (2021a). In brief, 19 bacterial strains (18 different bacterial genera) of diverse taxonomy were cultured and afterward harvested by centrifugation. For each bacterial strain, a separate extraction of genomic DNA (gDNA) was performed. For the ZIEL2 mock community, 12 ng of each bacterial gDNA was pooled.

2.5.3.2 Short amplicon 16S rRNA gene sequencing

Amplicon preparation

Short amplicons were prepared as previously described by Reitmeier *et al.* (2020) and Abellan-Schneyder *et al.* (2021a). For amplification of the variable regions and addition of adapters, a 1st-step PCR was performed in 50 µl volume targeting either V1-V2 or V3-V4 (Tab 2.5.1). To enable multiplexing, barcodes were added in a 2nd-step PCR.

Table 2.5.1: Variable region-specific forward and reverse primers and annealing temperature for 1st-step PCR.

Region	Forward primer	Reverse primer	Annealing Temperature	Reference
V1-V2	AGA GTT TGA TYM TGG CTC AG	GCT GCC TCC CGT AGG AGT	57°C	Salter *et al.* (2014)
V3-V4	CCT ACG GGN GGC WGC AG	GAC TAC HVG GGT ATC TAA TCC	55°C	Klindworth *et al.* (2012)

Library quality check

For validation and quality assurance, 8 µl of 2nd-step PCR product were loaded on a 1.5% (w/v) agarose gel to perform gel electrophoresis. The remaining product was purified with 0.6x AMPure XP beads. Concentrations of the 2nd-step PCR product were measured in triplicates using a Qubit 4.0.

Sequencing on Illumina MiSeq

Samples were adjusted to 0.5 nM, pooled, and sequenced in paired-end modus (PE300) on a MiSeq system (Illumina, Inc.) following the manufacturer's instructions. Loading was performed using 12 pM of the pool and 15% (v/v) PhiX standard library.

Data analysis

For the downstream analysis, samples were processed with the DADA2 pipeline v1.18.0 (Callahan *et al.*, 2016). First, primer and remaining adapter sequences had to be removed. Therefore, short amplicon reads were trimmed using cutadapt (Martin, 2011). DADA2 was run. The following parameters were set individually: paired-end (PE) mode was used, for V1-V2, a truncation length of 200/180 bp was set, and for V3-V4, it was set to 260/220 bp. Further settings were: maxN = 0, maxEE = 2/2, and trunQ = 2. As a reference database, Silva v132 (https://www.arb-silva.de/documentation/release-132/, accessed on 2/12/2020) was chosen. Downstream analysis was performed in Rhea (Lagkouvardos *et al.*, 2017). Rhea is a collection of R-scripts enabling comparison between samples. After normalization of data, α- and β-diversities can be visualized. The script also performs taxonomic binning, enabling an insight on all known and unknown sequences of the microbial composition down to the genus level.

2.5.3.3 Full-length amplicon Preparation

For the MinION and the in-house synthetic full-length approach, amplicons of about 1,700 bp were produced in a three-step set-up (Figure 2.5.1).

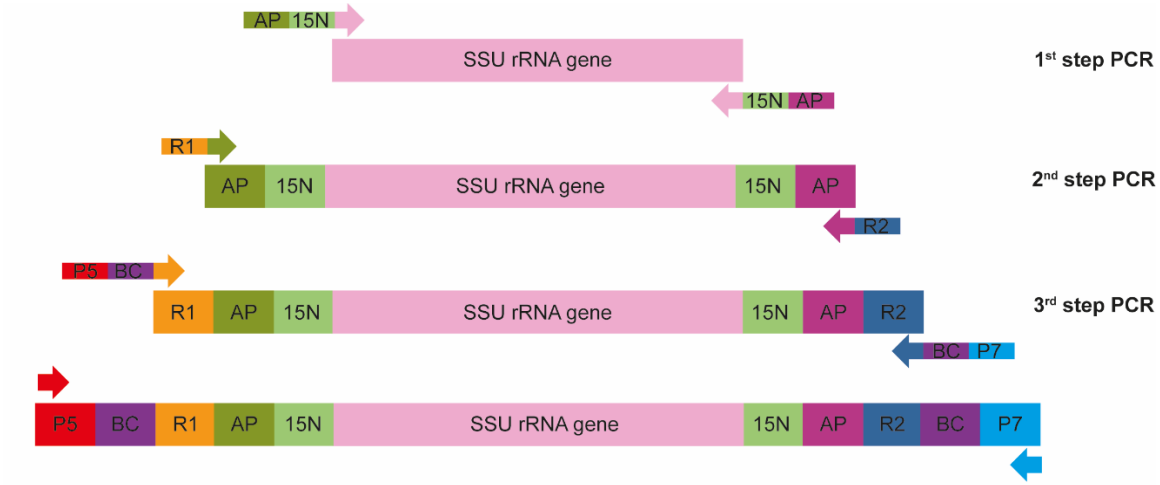

Figure 2.5.1: Overview of the three-step PCR procedure to produce full-length SSU rRNA gene amplicons. In a 1st-step PCR, primers annealing to the SSU rRNA gene are used that include a 15N-UMI tag and an adapter sequence (AP). This AP is then targeted in a 2nd-step PCR, where the Illumina overhangs (R1/R2) are added. Barcodes (BC) and the Illumina adapter (P5/P7) are added in the 3rd step PCR. If, after the 3rd step, insufficient amounts of full-length amplicons were produced, re-amplification with P5/P7 primers was performed.

In short, a 1st-step PCR was performed using 10 ng gDNA, 1x Phusion HF buffer, 0.2 mM dNTPs, 0.2 µM forward primer(s), 0.2 µM reverse primer(s), 5% (v/v) DMSO, and 1.5 U of Phusion Hot Start II DNA Polymerase. In a thermocycler, denaturation took place at 98°C for 30 s followed by five cycles of denaturation at 98°C for 10 s, primer annealing at 57°C for 30 s, elongation at 72°C for 90 s, and completing the reaction with a final elongation step at 72°C for 5 min. 1st-step PCR products were purified with 1.8 AMPure XP beads, washed twice with 80% EtOH and eluted in 10 µl water. This 10 µl were then used for the 2nd-step PCR, which contained 1x Phusion HF buffer, 0.2 mM dNTPs, 0.125 µM forward primer, 0.125 µM reverse primer, 5% (v/v) DMSO, and 1.5 U of Phusion Hot Start II DNA Polymerase. In a thermocycler, denaturation took place at 98°C for 30 s followed by ten cycles of denaturation at 98°C for 10 s, primer annealing at 57°C for 30 s, elongation at 72°C for 90 s, and completing the reaction with a final elongation step at 72°C for 5 min. PCR products were purified with 1.8x AMPure XP beads and eluted in 20 µl water. A third step PCR was performed to enable multiplexing. The third step PCR was performed accordingly to the second step, with the difference of using barcoding primer for amplification. Products were again cleaned up afterward with 1.8x AMPure XP beads, and the concentration of the final product was assessed using a Qubit. Amplicons were assessed using agarose gel electrophoresis, and gel extraction was performed if side products were observed. If insufficient material was produced to process, re-amplification with P5/P7 primers was performed.

2.5.3.4 Long-read sequencing using ONTs MinION

For MinION sequencing, the direct cDNA sequencing Kit (SQK-DSC109) was used. Starting-point was the end-prep reaction, as amplicons were already produced. Thus, samples were further processed as described by the manufacturer. Pooling of 12 samples per run was performed, where every sample was pooled using 200 fmol DNA amplicon. Sequencing adapters were added as described by the manufacturer, and then the final library was loaded onto an FLO-Min106 (R9.4.1) flow cell. Sequencing was performed until 1-1.2 million reads were generated, which took on average 4 hours. The generated FASTQ files were further analyzed using workflows provided by the EPI2ME platform (Metrichor, ONT).

2.5.3.5 Synthetic full-length sequencing using an in-house protocol

Droplet digital PCR (ddPCR) and re-amplification

In a first step, the previously produced full-length amplicons (1st-step PCR products) were processed by ddPCR. The advantage of ddPCR is the partition of the PCR reaction into thousands of individual reaction compartments and thereby an equal amplification of each target in the vessels, which function as microreactors. In our case, this will allow containment of UMIs, which are later needed to have enough reads bearing the same UMI for *de novo* assembly.

Samples were diluted to approximately 20,000 copies/20 µl, which lead to final dilutions ranging between 10^{-5} to 10^{-7} of the original starting material. The ddPCR mixture containing 10 µl of 2x ddPCR EvaGreen Supermix (Bio-Rad), 100 nM of forward and reverse primers, and the targeted PCR product in a total reaction volume of 20 µl. Then, the 20 µl reaction mixture and 70 µl Droplet Generation Oil for EvaGreen were loaded into a clean droplet cartridge (Bio-Rad). After droplet generation, the emulsion was transferred into a 96 well plate. The plate was sealed using the PX1 PCR Plate Sealer (Bio-Rad), and PCR was performed in a peqSTAR (VWR) cycler with following conditions: 95°C for 5 min, 40 cycles of 95°C for 30 s, 57°C for 1 min and 72°C for 3 min, and afterward signal stabilization steps at 4°C for 5 min, 90°C for 5 min and 4°C hold step, temp rate was set to 2°C/s for all steps. PCR reactions were transferred into fresh Eppendorf tubes, and extraction of the PCR products was performed by chloroform extraction. Recovery was performed as described in the ddPCR application guide provided by Bio-Rad with chloroform and 1x TE buffer. In brief, the oil phase of each sample was discarded, and 20 µl 1x TE buffer and 70 µl chloroform were added to the remaining aqueous phase. The mixtures were vortexed for 1 min at high speed in a 2 ml adapter for the Vortex-Genie 2 (Thermo Fisher) and centrifuged at 15,500×*g* for 10 min. The upper aqueous phase (volume approx. 25 µl),

containing amplicons, was separated by pipetting. Samples were purified using 1x AMPure XP beads and eluted in 20 µl H_2O. Recovered products were then loaded on a 1.5% (w/v) agarose gel and checked for successful amplification of the desired product. Afterward, concentrations were assessed on a Qubit. Then, the 2nd- and 3rd-step PCRs were performed as described in section 2.5.3.3 with the exception that the Q5U polymerase and the corresponding Q5U buffer were used. The Q5U polymerase was used as we found that a non-proofreading polymerase (e.g., *Taq* polymerase) or a polymerase that can read and amplify templates containing uracil (and inosine bases), are needed since the ddPCR Supermix contains dUTPs. After the 3rd-step PCR, concentrations were assessed on a Qubit and re-amplification was performed if less than 4 ng total material was available. The re-amplification reaction mix contained: 1× Q5U reaction buffer, 200 µM dNTPs (10 mM); 0.5 µM P5 primer (forward), 0.5 µM P7 primer (reverse), 5 µl ddPCR product (≤1 ng/µl), 0.02 U/µl Q5U Hot Start High-Fidelity DNA Polymerase, up to 50 µl nuclease-free H_2O. PCR was performed using the following conditions: 98°C for 30 s, 10/15 cycles of 98°C for 10 s, 57°C for 20 s and 72°C for 90 s, and a final extension for 2 min at 72°C before samples were stored at 8°C.

Tagmentation of full-length amplicons

Tagmentation, is a portmanteau word of “tagging” and “fragmentation” and refers to using the transposase Tn5 with preloaded DNA-fragments (i.e., adaptors) to cut double-stranded DNA and ligate specific adaptors to both ends of the opened DNA strand (Di *et al.*, 2020). This was performed using the previously produced full-length amplicons as template to enable sequencing on a short-read sequencer. Here, the Nextera DNA tagmentation kit (Illumina Inc.) was used and performed, as illustrated in Fig 2.5.2. In brief, a reaction mix containing 1x Tris-DMF buffer, 1.2 µl of a 1:10 diluted tagmented DNA enzyme 1 (Illumina Inc.) and 4 ng purified full-length amplicon were prepared in a total volume of 20 µl. Samples were incubated in a thermocycler for 5 min at 55°C before cooling them to 10°C. Complete termination of the reaction was accomplished by purification with 0.6x AMPure XP beads and elution in 22 µl H_2O. Then, 10 µl of the purified sample were used for the tagmentation PCR containing 1x Q5U reaction buffer, 0.2 mM dNTPs, 0.25 µM forward barcode primer, 0.25 µM reverse barcode primer, and 0.5 U of Q5U DNA polymerase in a total volume of 25 µl. The PCR reaction was set up with an initial temperature at 72°C for 3 min and an initial denaturation at 98°C for 30 s followed by 15 cycles of denaturation at 98°C for 10 s, primer annealing at 59°C for 30 s, and elongation at 72°C for 2 min. Final elongation was taking place at 72°C for 5 min, followed by sample storage at 8°C.

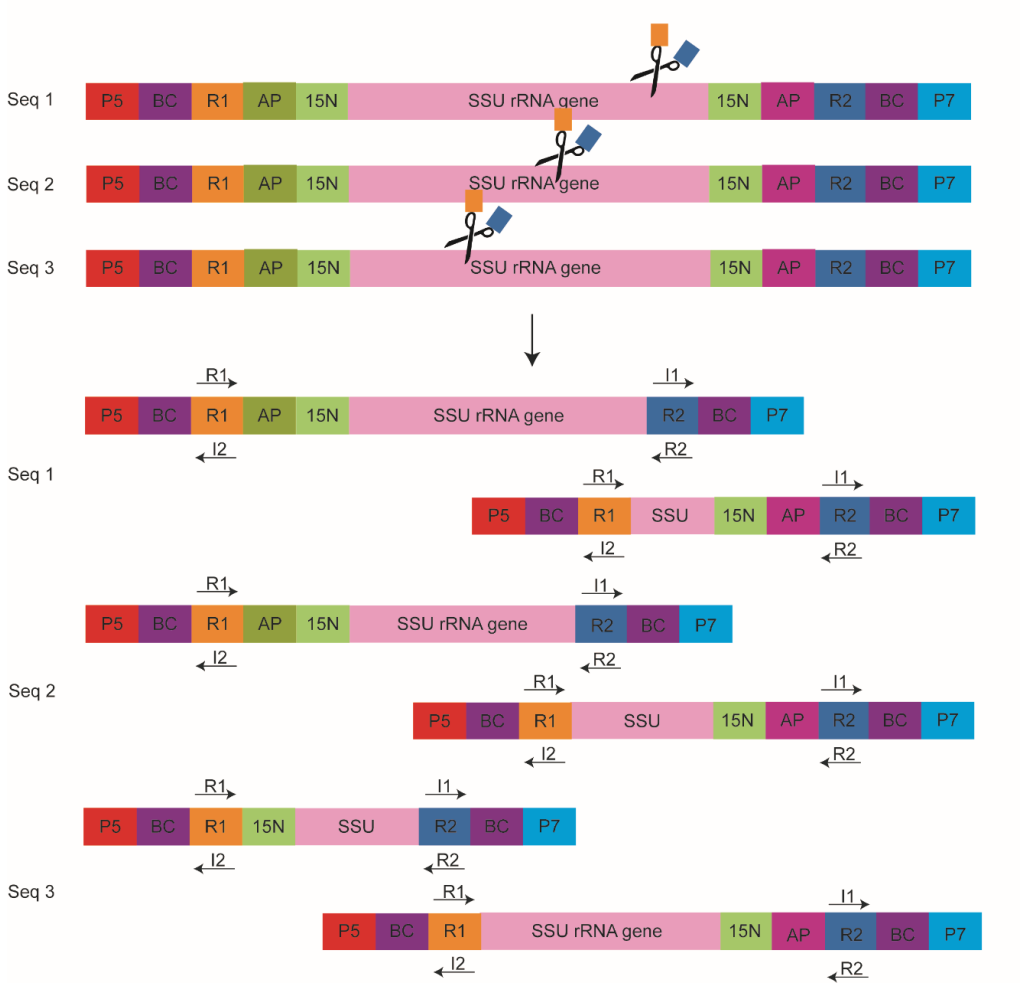

Figure 2.5.2: Tagmentation of the full-length amplicons. The Tn5 transposase cut the amplicon and ligates an Illumina overhang (R1/R2) which is later targeted in the tagmentation PCR. The arrows labeled with R1, R2, I1, and I2 show how Illumina sequencing is performed in the paired-end mode.

After the amplification PCR took place, 1x TriTrack loading dye was added to samples before loading 18 µl of them on a 6% PAA/TBE gel, which was run for 50 min at 145 V. The gel was then stained with 8 µl SYBR Gold in 50 ml 1xTBE buffer for 10 min. Bands from 600 bp to approximately 1,800 bp length were cut out from the gel. Gel pieces were transferred into a gel-breaker tubes and centrifuged at 13,000×*g* for 5 min. Gel debris was incubated overnight in water. Afterward, debris was removed by filtering the samples through Corning Costar Spin-X centrifuge tube filters (Sigma Aldrich). The remaining sample was cleaned up using 0.6x AMPure XP beads and eluted in 20 µl H_2O. The concentration of the final product was measured on the Qubit, and library size distribution was assessed on the Bioanalyzer (Agilent).

Self-circularization for binning 15N-tags of both ends

For adapter assignment (see Fig 2.5.3), 2 ng of the 1st-step ddPCR processed full-length amplicon sample was used. First, all samples were end-repaired. Towards this end, for

each sample, a reaction including 2 ng amplicon product, 1x NEB End repair reaction buffer, and 0.75 µl NEB End repair enzyme mix was set up in a total volume of 20 µl. The reaction was incubated at 20°C for 30 min.

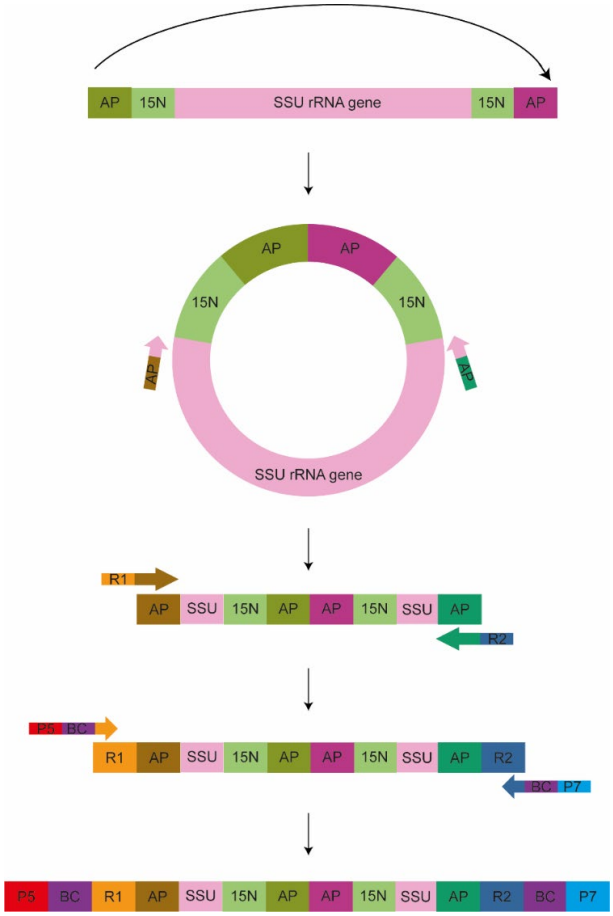

Figure 2.5.3: Overview of the adapter assignment process. First, samples are end-repaired, and a blunt-end self-ligation is performed. Then products are amplified by using reverse-complement primers of the SSU rRNA gene sequences. In the last step, sequencing adapters are added to allow sequencing on an Illumina MiSeq.

Afterward, a clean-up was performed using 1x AMPure XP beads. A subsequent blunt-end ligation was performed for 1 h at 16°C. The 200 µl reaction included 10 µl of the end-repaired product, 16% of a 40% (w/w) PEG 8000 solution, 1x T4 DNA ligase buffer, and 3,200 U T4 DNA Ligase (New England Biolabs). The ligation reaction was purified with 1x AMPure XP beads, and an amplification PCR was performed. The reaction contained 15 µl of the purified circulated product, 1x Q5U reaction buffer, 200 µM dNTPs, 0.5 µM forward primer, 0.5 µM reverse primer, and 2.5 U Q5U polymerase. PCR was performed at 98°C for 30 s, followed by 15 cycles of 98°C for 10 s, 59°C for 20 s, and 72°C for 30 s, and then a final extension at 72°C for 5 min. Samples were purified using 1x AMPure XP beads and subsequently barcoded in a two-step barcode reaction. The 1st-step was performed in a total volume of 50 µl. The reaction contained 8 µl PCR product from above, 1x Q5U reaction buffer, 200 µM dNTPs, 0,5 µM forward R1-adapter primer, 0,5 µM reverse R2-

adapter primer, and 2.5 U Q5U polymerase. PCR was performed at 98°C for 30 s, followed by 10 cycles of 98°C for 10 s, 59°C for 20 s, and 72°C for 30 s, and then a final extension at 72°C for 5 min. PCR products were purified using 1x AMPure XP beads and eluted in 20 µl water. For the 2nd step, 10 µl sample was mixed-up with 1x Q5U reaction buffer, 200 µM dNTPs, 0.5 µM forward barcode primer, 0.5 µM reverse barcode primer, and 2.5 U Q5U polymerase. PCR was performed at 98°C for 30 s, followed by 10 cycles of 98°C for 10 s, 57°C for 20 s, and 72°C for 30 s, and then a final extension at 72°C for 5 min. Then, 5 µl of each of the final barcoded products were loaded onto a 1.5% (w/v) agarose gel, and gels were screened for positive bands at about 300 bp. If bands were observed, samples were cleaned up using 1x AMPureXP beads and eluted in 30 µl water. The concentration of the barcoded products was determined on a Qubit in triplicates.

Sequencing of the tagmented full-length amplicons on Illumina MiSeq

Pooling of 0.5 nM of each sample (for each initial sample, two libraries were pooled: the tagmented sample (RT) and the self-circularized sample (LT)) was performed. Afterward, the pool was denaturized and prepared for sequencing. Paired-end sequencing was performed on an Illumina MiSeq following the manufacturer's instruction (Illumina Inc.). A final DNA concentration of 12 pM after the addition of 15% (v/v) PhiX standard library were used.

Data analysis

The quality of tagmented reads was controlled using FastQC (Andrews, 2010). Files were processed using an in-house protocol. The tagmen_suite was used. In short, tagmented sample reads are screened for their 15N-tag, and 15N sequences are extracted using the readtag.py script. Then an output folder is created, for which the CD-HIT-EST script is run. Clusters containing the same 15N-tags are created by running the parse_cluster script. The corresponding 15N-tags are identified through the reads emerging from the self-circularization step. 15N are therefore extracted using the linktag.py script. In a final step, results are aggregated through the aggerate.py script. The *de novo* assembly is performed using the A5_miseq assembly pipeline (Tritt *et al.*, 2012), and FASTA output is generated by the in-house get_fasta.py script. FASTA files are then classified using the SINA ACT: Alignment, Classification and Tree Service (Pruesse *et al.*, 2012). All scripts are available online through https://github.com/TUM-Core-Facility-Microbiome/tagmen (last accessed 06/24/2021).

2.5.3.7 Synthetic full-length sequencing using the 16S/18S LoopSeq kit (Loop Genomics)

The LoopSeq 16S & 18S Long Read Kit (Version 2.1, Loop Genomics) was used generating *de novo* assembled long reads, which were sequenced on a short-read sequencer before. This kit provides primers that can bind to 16S and 18SrRNAs, allowing the parallel estimation of bacteria, archaea, and eukaryotic microorganisms. Sample processing was performed as described by the manufacturer. In short, 5 µl 1:10-diluted gDNA was used for the enrichment PCR. The enrichment PCR was performed for 20 cycles of denaturation, annealing, and elongation as described by the manufacturer. For barcode assignment, 1.5 ng PCR product of each sample were used. The barcode assignment was performed as described in the manual. qPCR was performed using 10 µl reaction volumes for each sample instead of 20 µl as described by the manufacturer. Afterward, samples were adjusted to 8,000 barcoded sequences/sample. Next, the barcode distribution step was performed, and the concentrations were checked afterward on a Qubit. Based on the measured concentrations, pooling was performed in equimolar amounts, and all further steps were performed as suggested by the manufacturer. Further details and the sequencing procedure were previously described by Abellan-Schneyder *et al.* (2021b). After successful sequencing, the resulting short raw reads were uploaded to the analysis platform provided by LoopGenomics. After the analysis was finished, a zip file including the stats and sequences for the assembled full-length sequences, and taxonomy files, were downloaded from the website.

2.5.4 Results

Species-level classification

In a first step, it was assessed and compared how accurate species-level classification was performed for each of the four approaches (short amplicons, MinION long-reads, LoopSeq, and in-house synthetic full-length approach) when mock communities were analyzed. Towards this end, we used, on the one hand, the simple Zymo mock community, which consists of 8 different bacterial species, and the more complex ZIEL2 mock community, which was created out of 19 different bacterial species belonging to 18 different genera. Overall, the LoopSeq method performed best for both the Zymo and the ZIEL2 mock community (Fig 2.5.4). Regarding the Zymo mock community (Fig 2.5.4A), the V3-V4 amplicon approach performed worst, followed by the in-house synthetic full-length approach, the short amplicon V1-V2 approach, and the ONT MinION long-read approach. For the ZIEL2 mock community (Fig 2.5.4B), again, the LoopSeq approach performed best. Nevertheless, the short-read amplicons performed comparably well with

V1-V2 performing better than the long-read MinION approach and the in-house synthetic full-length approach.

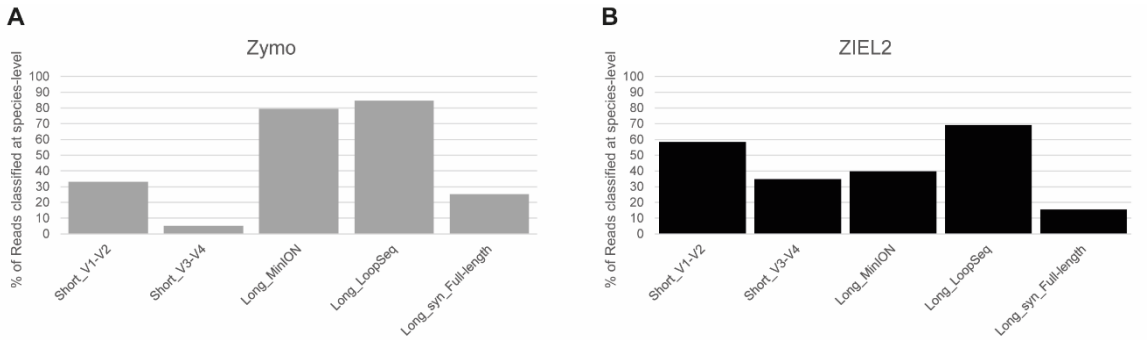

Figure 2.5.4: Percentage of all reads that could be classified down to species level. The Zymo mock community (**A**) performed well when long-reads were recorded using either ONTs MinION or the LoopSeq kit. The ZIEL2 mock community (**B**) was best classified at the species level by the LoopSeq kit, followed by the short amplicon V1-V2 approach.

Performance of different sequencing approaches on mock communities

As the species-level classification varied strongly between the different methods applied, the taxonomic composition was analyzed at the genus level (Fig 2.5.5). When the less complex Zymo mock community is analyzed, taxonomical profiles vary only slightly (Fig 2.5.5A). Of note, the MinION full-length approach drastically underrepresents *Escherichia/Shigella* (marked in green in Fig. 2.5.5A). A non-metric multidimensional scaling (NMDS) plot was created to assess how the different approaches performed compared to the ideal (theoretical composition). For the Zymo mock community, the NMDS plot (Fig 2.5.5B) showed moderate distances for both short amplicon approaches (triangles) and the LoopSeq approach. The MinION and in-house approach, on the other hand, showed larger distances. In any case, when analyzing the more complex ZIEL2 mock community, taxonomical profiles vary more strongly from method to method (Fig 2.5.5C). Several taxa are not identified or dramatically under-/ overrepresented within the different methods applied. For example, the short amplicon V1-V2 approach drastically underrepresented *Akkermansia* and *Bifidobacterium*. The V3-V4 approach could not classify *Ruminoccus* at the genus level. All full-length approaches suffered to classify or dramatically underrepresents *Akkermansia*, *Bifidobacterium,* and *Eggerthella*. The MinION and in-house approach markedly overrepresented *Staphylococcus,* whereas the LoopSeq, for example, overrepresented *Escherichia*/*Shigella*. The NMDS plot (Fig 2.5.5D) demonstrated that overall, at the genus level, the short amplicons performed more similar to each other and to the “ideal” (i.e., theoretical) profile. Thus, taxonomical profiles that are closer to this ideal composition were observed, while the full-length approaches cause cluster to be more distant from the ideal composition.

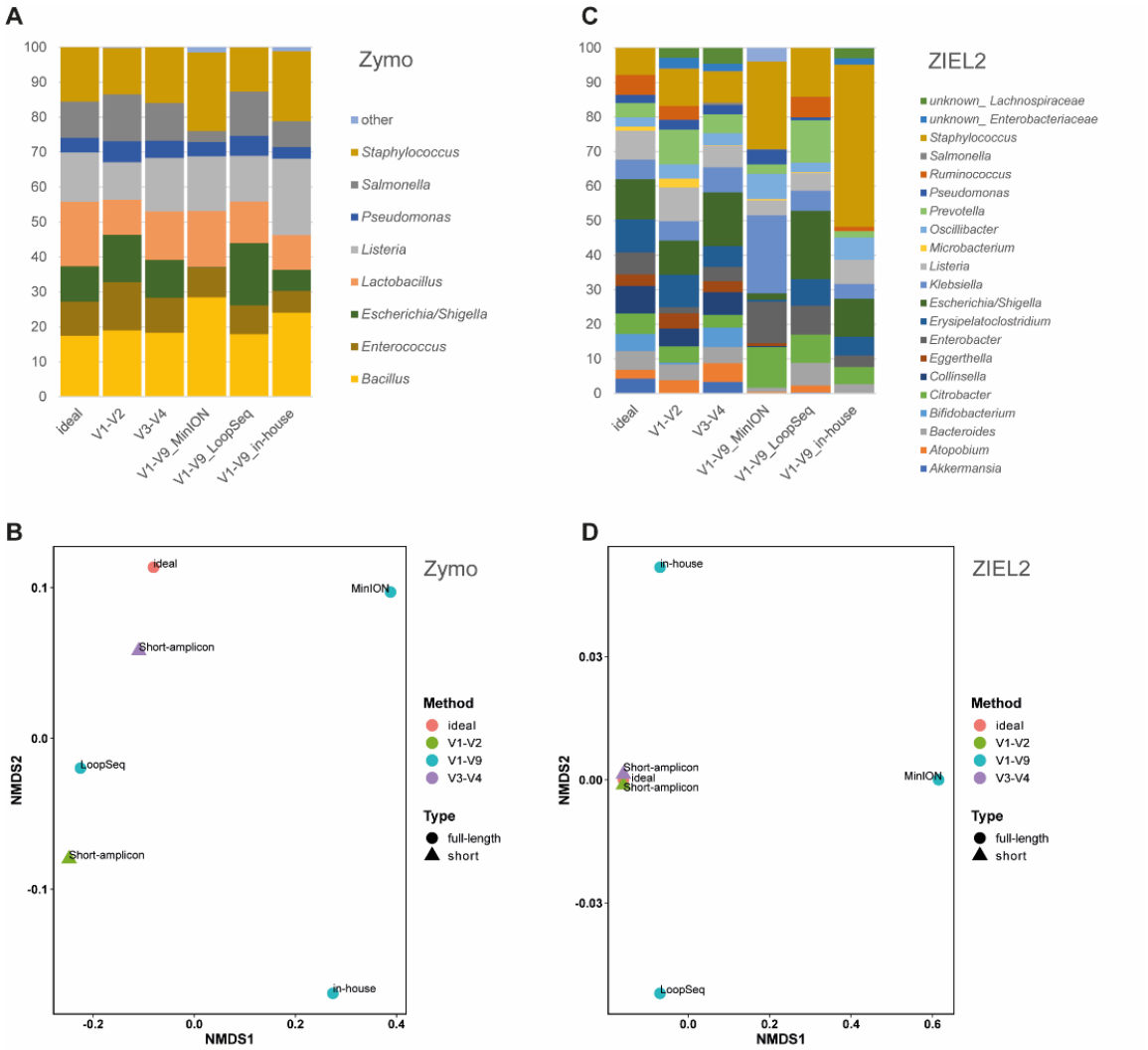

Figure 2.5.5: Performance of the different sequencing methods using two different mock communities: Zymo (A, B) and ZIEL2 (C, D). Plotting of relative abundances of bacteria included in the Zymo mock (**A**). The ideal (theoretical) composition (first row) was compared to taxonomical profiles gained for the different sequencing approaches tested. Distances for the tested approaches were compared in an NMDS plot for the Zymo mock community. Resulting compositions of short amplicons V1–V2 (green), V3–V4 (purple), or full-length approaches, V1–V9 (blue), were compared to the ideal composition (red) at the genus level (**B**). Plotting of relative abundances of bacteria included in the ZIEL2 mock community (**C**). The ideal (theoretical) composition (first row) was compared to taxonomical profiles gained for the different sequencing approaches tested. Distances for the tested approaches were compared in an NMDS plot for the ZIEL2 mock community (**D**). Resulting compositions of short amplicons V1–V2 (green), V3–V4 (purple), or full-length approaches, V1–V9 (blue) compared to the ideal composition (red). Please note the differences between the data points indicated by the scale compared to panel B.

To further assess, which approach performed best for the ZIEL2 mock community, the generalized UniFrac distance was calculated (Tab 2.5.2). Principally, the generalized UniFrac distance matrix is calculated to detect compositional changes of abundant but also rare taxa. The lower the calculated values are, the smaller the difference between the two sets that were compared is expected. Here, we found that the V3-V4 short amplicon data for the ZIEL2 mock showed the smallest value and was, therefore, the most

comparable to the ideal mock. The short amplicon V1-V2 approach performed second best and thus superior to the LoopSeq, in-house, and MinION sequencing strategy.

Table 2.5.2: Generalized UniFrac matrix for ZIEL2 samples in comparison to the ideal composition at the genus level.

	ideal ZIEL2	Short V1-V2	Short V3-V4	MinION V1-V9	LoopSeq V1-V9	in-house V1-V9
ideal ZIEL2	0.00	0.12	0.10	0.43	0.19	0.28
Short V1-V2	0.12	0.00	0.14	0.45	0.19	0.27
Short V3-V4	0.10	0.14	0.00	0.43	0.21	0.29
MinION V1-V9	0.43	0.45	0.43	0.00	0.38	0.34
LoopSeq V1-V9	0.19	0.19	0.21	0.38	0.00	0.26
in-house V1-V9	0.28	0.27	0.29	0.34	0.26	0.00

Performance of different sequencing approaches on a human fecal sample

It must be noted that the different sequencing strategies cannot be directly compared. On the one hand, all besides the MinION strategy were sequenced on an Illumina MiSeq and, thus, have an overall lower error rate compared to MinION. Moreover, the total number of sequencing reads vary dramatically (Tab 2.5.3).

Table 2.5.3: Overview of the sequencing reads produced for the Zymo and ZIEL2 mock community as well as for human sample T1. Five different sequencing strategies were used. For the synthetic full-length approach, the number of *de novo* assembled full-length sequences is noted as well.

	Reads sequenced					Assembled full-length sequences	
Region	V1-V2	V3-V4	Full-length (V1-V9)			Full-length (V1-V9)	
Method	Short	Short	MinION	LoopSeq	in-house	LoopSeq	in-house
Zymo	65,590	49,483	60,899	1,457,754	28,588	7,652	1,738
ZIEL2	85,086	92,888	13,472	1,360,029	23,299	7,086	1,430
T1	24,424	43,446	79,446	1,333,162	57,806	6,766	240

Nonetheless, in this proof-of-principle it was analyzed whether taxonomical profiles show recurrent characteristics. Here, we compared the taxonomical profiles (Fig 2.5.6) of a human sample (T1), which was kindly provided by a healthy human donor. Overall, taxonomical profiles (Fig .2.5.6A) show similarities, e.g., *Prevotella* (dark blue) being highly abundant and larger proportions of *Bacteroides* (brown) and *Ruminococcus* (yellow). Nonetheless, also concerning this recurrent feature, changes are observable. The amount of *Prevotella* varies drastically, with about 50% of all reads clustering to *Prevotella* in the MinION dataset, while in the in-house full-length approach, the proportion of *Prevotella* is much smaller, with about 15% of all reads. The short amplicon reads look very similar, which is reinforced in the NMDS plot (Fig 2.5.6B), where only smaller distances of the short amplicons are observed. The LoopSeq approach had smaller distances to the short

amplicons and, therefore, is considered to produce more similar data points than the MinION and in-house full-length sequencing approaches.

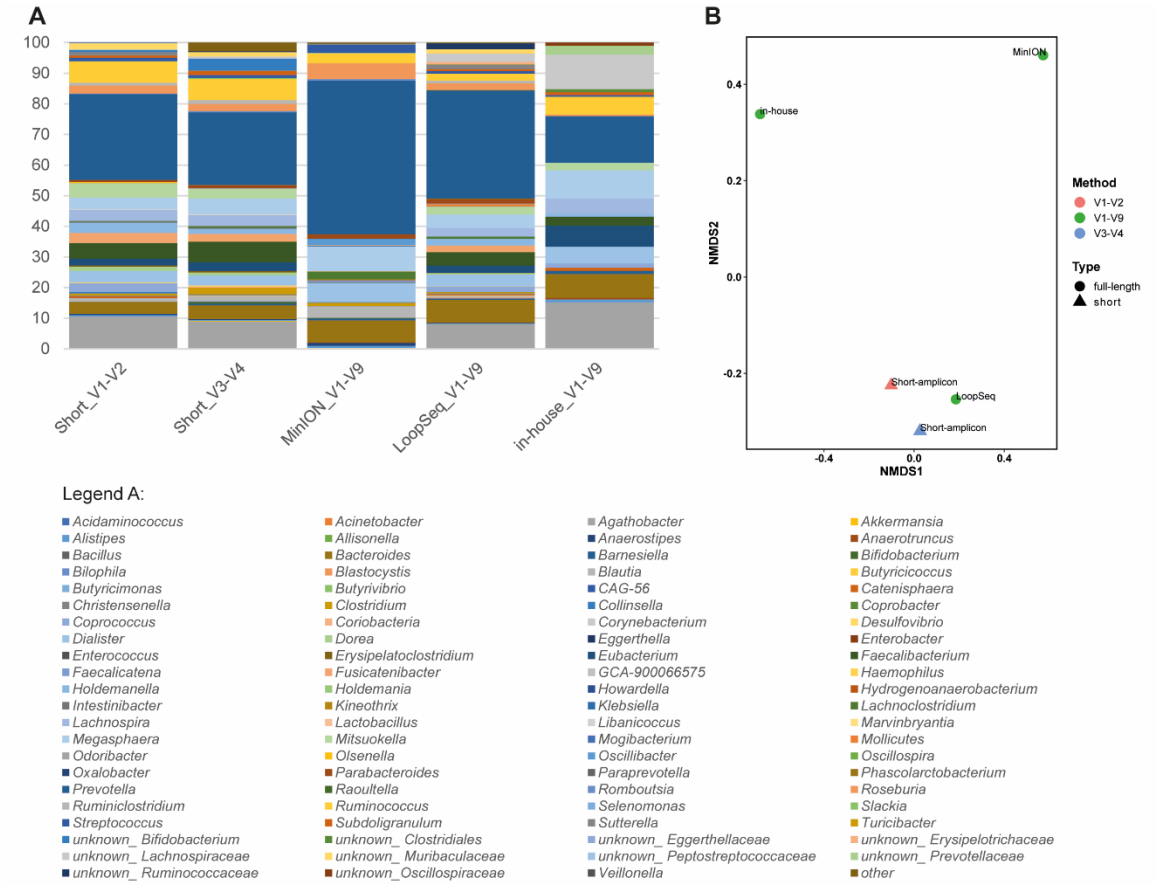

Figure 2.5.6: Performance of the different sequencing methods using a human fecal sample. The relative abundance of bacteria at the genus level is compared (**A**). NMDS plot showing the results for the human sample sequenced using short amplicons V1–V2 (red), V3–V4 (blue), or full-length approaches, V1–V9 (green) are compared (**B**).

Benefits, drawbacks, and feasibility of the different sequencing approaches

To assess which sequencing strategy should be used for different purposes, we aimed to compare the most prominent benefits and drawbacks. Moreover, we determined the feasibility and time-effort for each approach. The results are summed up in Table 2.5.4. Generally, one must consider whether to use short amplicon sequencing of a full-length approach. Short amplicon sequencing is overall cheap, easy to conduct, and can be standardized. Nevertheless, it is subjected to primer bias depending on which V-regions are targeted and sequenced and which primers are used. Full-length sequencing approaches are overall more difficult and time-consuming to conduct and should, therefore, only be used if species-level information is needed. For species-level classification, the LoopSeq approach performed overall best and was the suggested method of choice here. Still, it should be noted that this method allows sequencing only in a more than ten-fold lower throughput than short amplicons. The in-house methodology

had a sequencing depth similar to what is used for short amplicons and, thus, was not sequenced in sufficient depth. Samples should be re-sequenced, and analysis should be repeated to allow better comparability of the results.

For the MinION approach, an average raw read accuracy of 89% was detected, accordingly the error rate is much higher compared to all other approaches. Therefore, sequencing using the MinION is, at the moment, only recommended if time is critical, e.g., to detect pathogens in a clinical environment. Further investigations and newer kits of ONT should be tested, and it should be assessed whether the sequencing errors can be reduced, for instance using the $1D^2$ approach as suggested by Calus *et al.* (2018) or by applying the Rolling Circle Amplification to Concatemeric Consensus (R2C2) method (Volden *et al.*, 2018). With this, the technology would be of greater interest as it is easy to conduct and space and time saving.

2.5.5 Discussion

This study compared four different sequencing strategies, three full-length 16S rRNA gene sequencing strategies, and a short amplicon sequencing approach using two different primer sets. For this comparison, two mock communities of known composition with increasing complexity and a human fecal sample were used. We found that the LoopSeq approach performed best in terms of reads that could be classified at the species level. At genus level for the ZIEL2 mock, the V3-V4 approach performed best, i.e., showing the smallest deviation from the ideal composition. Next to this, V1-V2 performed second best, LoopSeq third, the in-house full-length approach is on the fourth position, and the MinION approach performing worst. Even though we do not know for the human samples which method cluster best, as the “ideal” composition is not known we could observe that the short amplicon sequencing profiles of the human sample clustered in close proximity to the LoopSeq results. On contrast, longer distances were observed to the MinION and in-house processed reads.

Thus, we believe, when using the settings of the analysis pipelines as described, short amplicon approaches perform best as long as genus-level taxonomic information is sufficient. If species-level information is of interest, we would suggest applying the LoopSeq approach. Nevertheless, it must be noted that our results might be biased according to several different factors: (1) the use of different sequencing types of machinery (ONT MinION vs. Illumina MiSeq), (2) the sequencing depth of the different approaches varied considerably, (3) different downstream and analysis pipelines (including filtering) and reference databases were used.

Table 2.5.4: Overview of the main benefits, drawbacks, and characteristics for the four tested different sequencing approaches.

	Benefits	Drawback	Hands-on Time	Feasibility / Know-how need	References/resources and similar approaches (last accessed 06/15/21)
Short amplicon sequencing	+ Cost-efficient and high-throughput (350 samples on one cartridge) + Widely available/accessible + Standardized protocols + Easy-to-use analysis tools available + Automation possible	- Primer dependency and primer biased - Different reference database and pipelines add bias	○ DNA extraction: 3 h ○ Library construction: 3 h ○ Sequencing preparation: 1.5 h ○ Sequencing: 48 h	- Easy - Availability of hands-on protocols - Downstream pipelines can be used without intense bioinformatical training	Abellan-Schneyder *et al.* (2021a), Reitmeier *et al.* (2020), Reitmeier *et al.* (2021), Allaband *et al.* (2019), http://www.human-microbiome.org/index.php https://hmpdacc.org/
MinION full-length sequencing	+ Fast + Easy to conduct (kit-based) + Small device, no extensive room/ cooling needed + ONT specific analysis solutions	- High input amounts of DNA needed - Analysis pipelines vary in classification and accuracy - Standard primers in 16S kit system not optimized - No “best practice” - Limited number of samples at the same time (96 samples/run) - Long run time for base-calling	○ DNA extraction: 3 h ○ Library construction: 2 h ○ Sequencing preparation: 1 h ○ Sequencing: 4-12 h	○ Easy ○ Protocol provided by the manufacturer ○ Downstream pipelines can be used without intense bioinformatical training (but EPI2ME pipeline was tested to classify error-prone)	Benitez-Paez & Sanz (2017), Matsuo *et al.* (2021), Heikema *et al.* (2020), Karst *et al.* (2021), Shanmuganandam *et al.* (2019) https://nanoporetech.com/
LoopSeq full-length sequencing	+ A kit-based system, highly standardized and easy to conduct + Best species-level classification + Easy and quick downstream analysis-pipeline with automated *de novo* assembly	- High in costs - Exact methodology/ strategy not published - Further equipment, e.g., qPCR machinery and Bioanalyzer, needed - Lower throughput (24 samples on one Illumina cartridge)	○ DNA extraction: 3 h ○ Library construction: 9 h ○ Sequencing preparation: 2 h ○ Sequencing: 48 h	○ Easy ○ Protocol provided by the manufacturer ○ Downstream analysis performed by LoopGenomics, easy and convenient	Chung *et al.* (2020), Callahan *et al.* (2021), Jeong *et al.* (2021), Abellan-Schneyder *et al.* (2021b), https://www.loopgenomics.com/
in-house full-length sequencing	+ Amplification of very low amounts of gDNA possible + Possible to start with RNA instead of DNA as starting material (primer-free) + Primers can be easily adapted/ changed, kit-independent + A mixture of different primers (e.g., targeting 18S or ITS) can be used	- low throughput - high sequencing depth needed - difficult protocol steps (e.g. ddPCR, PAGE extraction, tagmentation) - ddPCR step is limiting (performed in 8-sample stripes) - Tagmentation difficult to be scaled (DNA concentration-dependent)	○ DNA extraction: 3 h ○ Library construction: 15 h ○ Sequencing preparation: 2 h ○ Sequencing: 48 h	○ Difficult ○ Protocol existing, can be improved and adapted ○ Downstream pipelines without a graphical interface ○ Long-run time of downstream analysis	Burke & Darling (2016) Karst *et al.* (2018), Deutscher *et al.* (2018)

For example, it was previously shown that the ONT EPI2ME workflow for the MinION sequencing strategy does not allow the best possible and most precise species-level classification and is, further, not a convenient downstream analysis pipeline (Santos *et al.*, 2020, Winand *et al.*, 2019). However, in addition to the above, (4) different primers were used, and (5) different PCR cycling protocols and library purification steps were applied.

Use of different sequencing devices

While the commonly used Illumina MiSeq is based on the sequencing by synthesis approach, the MinION sequencing strategy relies on the measurement of changes in the electric current, which is used to differentiate between the four bases and, perhaps, also to detect base modifications (Ansorge *et al.*, 2017, Liu *et al.*, 2019b). Even though the error rate decreased for the MinION device throughout the last couple of years, the error rates are still high (Ciuffreda *et al.*, 2021). In our analysis, we had average accuracies of 89-90% (i.e., 10-11% erroneous bases), while the MiSeq system produces highly accurate reads with an average error rate reported to be <0.5% (Pfeiffer *et al.*, 2018). Thus, correct identification and classification of MinION data at genus and species-level with identity thresholds typically being ≥97% is difficult, as the error rates of the technology applied are higher than the identity threshold (Winand *et al.*, 2019). Nonetheless, it must be noted that recently methods were described that could show significantly increased accuracies for long-read amplicon sequencing approaches using ONTs MinION. Karst *et al.* (2021) published a method sequencing the complete rRNA gene operon using ONTs MinION. They showed that by using unique molecular identifiers (UMIs), sequencing of the whole *rrn* operon, and drastically increased read coverage, they were able to lower the error rate to values <0.1%. To achieve that, they created UMIs that contain an internal pattern, which their provided downstream analysis pipeline can easily identify. Using those UMIs and long-read sequencing, they showed that single-molecule consensus sequences could be produced that showed an overall low chimera rate. Thus, by using such a precise method, species-level classification becomes possible.

Differences in sequencing depth

As mentioned above, Karst *et al.* (2021) had to, besides the use of the UMI strategy, drastically increase the read coverage to be able to reduce the error rates substantially. They stated that a coverage rate of 40x would be ideal per UMI sequence amplified. This exceeds by far standard protocols, and increases cost drastically. In our approaches, sequencing depths varied (Tab 2.5.3). The LoopSeq approach was sequenced in the highest depth and performed overall best. Thus, it should be evaluated if this would still be the case when samples sequenced using the MinION approach or the in-house protocol

would need to be sequenced as deep as the LoopSeq samples. For such testing's, an increase of 15-60x in sequencing depth would be required.

Use of different downstream analysis pipelines and reference databases

For short amplicon 16S rRNA gene sequencing, it was previously shown that the use of different clustering approaches, analysis pipelines, and references databases could influence the taxonomical profiles of samples analyzed (e.g., Abellan-Schneyder *et al.*, 2021a, Almeida *et al.*, 2018, De Filippis *et al.*, 2018, Nearing *et al.*, 2018, Park and Won, 2018). All data, except the MinION approach, used Silva as a reference database. The MinION reads were processed using the ONT EPI2ME tool, which is based on the NCBI 16S RefSeq. It was previously shown that EPI2ME is not performing best, as Cuscó *et al.* (2019) showed that Minimap2 outperformed EPI2ME and Winand *et al.* (2019) showed that the program Mothur performed superior when compared to EPI2ME. Nevertheless, EPI2ME is the most convenient analysis pipeline as no bioinformatical knowledge is needed. Regardless, improved pipelines, which are user-friendly, i.e., having a graphical interface allowing easy handling and changing settings (e.g., cut-offs in clustering approach or the used reference database) are needed. All of which currently is not available in EPI2ME. Some alternative approaches were published recently, including the 16S_ppm pipeline (Marino, 2020), which uses BLAST as an aligner and the NCBI 16S database as a reference database, or NanoCLUST (Rodríguez-Pérez *et al.*, 2020), which is based on Uniform Manifold Approximation and Projection (UMAP) and classifies based on BLAST. Besides, some studies use already existing and accepted approaches developed for short amplicon sequencing and adapted those for using it with long-reads. Examples include QIIME (Quan *et al.*, 2019, Sheahan *et al.*, 2019) and Mothur (Winand *et al.*, 2019). Such approaches facilitate comparability with short amplicon sequencing data, which is of great interest for the scientific community.

Use of different primers

The use of different primer sets cause bias, and the comparability of different approaches is decreased. For short amplicon sequencing, this factor was studied intensively before (e.g., Abellan-Schneyder *et al.*, 2021a, Clooney *et al.*, 2016, Fouhy *et al.*, 2016, Tremblay *et al.*, 2015). For short amplicon sequencing, primer pairs that were previously shown to perform overall well were chosen (Abellan-Schneyder *et al.*, 2021a). The LoopSeq kit includes a mixture of four different forward and two reverse primers targeting not only bacteria but also archaea and eukaryotes. For better comparability, we used the same microorganism-specific targeting sequences in our in-house protocol as were used for the LoopSeq kit. Nevertheless, it was shown previously (Abellan-Schneyder *et al.*, 2021b) that

this primer mixture should be re-evaluated. It includes, for example, the 27F-CM primer instead of the improved 27F-YM version.

Concerning MinION sequencing, several different approaches exist. The easiest and most convenient would be to simply purchase the 16S Barcoding Kit (SQK-16S024), which already includes 16S Primer. Again, the inferior 27-CM primer is used. Matsuo *et al.* (2021) demonstrated that the use of 27F-YM improves taxonomical classification, especially concerning *Bifidobacterium*, since this species cannot be targeted using the 27F-CM but can with the improved 27F-YM primer.

Different library preparation and PCR cycling approaches

In addition to the use of different primers, it was also shown that different library preparation protocols could influence the downstream analysis. For short amplicon sequencing, the effect of different PCR cycles or library preparation in one vs. two steps was formerly studied in detail (e.g., Ahn *et al.*, 2012, Drengenes *et al.*, 2021, Mallott *et al.*, 2019, Siebert *et al.*, 2021, Wu *et al.*, 2010b). For the MinION approach only a few different protocols (e.g., Cuscó *et al.*, 2019, Karst *et al.*, 2021, Matsuo *et al.*, 2021) were described that do not use of the standard ONT 16S library preparation kit.

In our study, we used an approach similar to the one by Karst, *et al*. where we first generated the full-length amplicons (Fig 2.5.1) and then ligated the ONT adapters using adapter ligation components of the Direct cDNA Native Barcoding Kit (SQK-DCS109) sequencing kit. For our approach, we detected roughly 50% of library loss after the adapter ligation step. Thus, an approach that does not need a ligation step, such as the protocol described by Matsuo *et al.* (2021) would be beneficial.

Cuscó *et al.* (2019) and Karst *et al.* (2021) could show that sequencing the whole *rrn* operon further improved the ability to classify bacteria down to the species level. Therefore, in further approaches, sequencing of the whole operon should be considered.

Besides sequencing the *rrn* operon, Leggett *et al.* (2019) presented an interesting ONT-based approach. They performed shotgun metagenomic sequencing of both mock communities and fecal samples. With this approach, not targeting the *rrn* genes directly, they demonstrated that several pathogenic bacteria such as *Klebsiella pneumoniae*, can be identified. Besides, we see the potential for our in-house approach, which is independent of any kit-based components. Thus, primers, cycling steps, even actual sequencing can be adapted and changed. Furthermore, primers allowing to reverse transcribe rRNA to cDNA were created, such that sequencing of the actual rRNA is possible. This approach is inspired by Karst *et al.* (2018) and allows studying not only the microorganisms present, but target those which are metabolically active. Disadvantages

are a further increase in library preparation time, more complex handling, and the need for an increased sequencing depth.

2.5.6 Conclusion

Here, we presented and compared three different full-length SSU rRNA gene sequencing approaches. The LoopSeq approach produced results most comparable to short amplicon sequencing (V1-V2 and V3-V4) and showed the best species-level resolution. Our in-house method failed to perform superior in species-level classification compared to the short amplicon approaches, but sequencing depth was low and should be improved. The MinION approach has an increased species-level classification for the simple Zymo mock community but failed to perform better at species-level for more complex samples. Nevertheless, we believe that the MinION and in-house approach could be drastically improved in their performance by choosing better primer, a higher sequencing depth, and/ or improve the analysis pipelines. Summarizing, all presented sequencing approaches are useful to determine a simple mock community at the genus level well. Nevertheless, resolution decreased for complex samples. Future work must show on how to improve library preparation, sequencing, and the downstream analysis for any of the full-length approaches improving performance.

2.6 Full-length SSU rRNA gene sequencing allows the species-level detection of bacteria, archaea and yeasts present in milk

Summary

In this study, it was analyzed whether full-length SSU rRNA gene sequencing could improve species-level classification of microorganisms present in milk samples when compared to standard short amplicon 16S rRNA gene sequencing. Further, it was assessed if putative mastitis pathogens could be detected on species-level.

We showed that using synthetic full-length sequencing instead of short amplicon sequencing improved species-level classification of microorganisms present in randomly chosen bulk-tank milk samples as well as in mock communities of known composition. Moreover, by using a primer mixture of primers targeting both 16S rRNA but also 18S rRNA, bacterial, archaeal, and eukaryotic microorganisms (i.e., yeasts) could be identified.

In a proof-of-concept, we demonstrated that putative mastitis-causing bacteria can be identified at the species level. Out of a list of 25 most commonly found mastitis pathogens, we identified 17 in the raw milk samples and, overall, 19 in the whole data set.

Some of those were, for example, *Streptococcus uberis*, *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, *Escherichia coli*, or *Staphylococcus aureus*.

Furthermore, we could also identify several *Candida* species, such as *Candida boidinii*, *Candida metapsilosis*, *Candida intermedia*, *Candida zeylanoides*, or *Candida pararugosa*.

Pathogenic or mastitis-associated archaeal species are not known. Nevertheless, we identified archaea at species level, which might be of interest, as archaea are not commonly targeted by standard short amplicon 16S rRNA gene primer pairs. In our data set, we detected several *Methanosarcina* species such as *Methanosarcina soligelidi*, *Methanosarcina mazei*, *Methanosarcina horonobensis*, as well as *Methanobrevibacter* species, e.g., *Methanobrevibacter millerae*.

In summary, the advantage of the presented full-length SSU rRNA gene sequencing approach over standard short amplicon 16S rRNA gene sequencing was shown to be an improved species-level classification, the concurrent analysis of bacteria, archaea, and eukaryotic microorganisms present in the analyzed samples, as well as the advantage that a kit-based methodology was used which is convenient in terms of standardization and reproducibility. Finally, the method was time-efficient, which is an important factor in many settings.

microorganisms

MDPI

Article

Full-Length SSU rRNA Gene Sequencing Allows Species-Level Detection of Bacteria, Archaea, and Yeasts Present in Milk

Isabel Abellan-Schneyder [1], Annemarie Siebert [2,†], Katharina Hofmann [2,†], Mareike Wenning [2,3] and Klaus Neuhaus [1,*]

[1] Core Facility Microbiome, ZIEL—Institute for Food & Health, Technische Universität München, 85354 Freising, Germany; isabel.abellan-schneyder@tum.de
[2] Chair of Microbial Ecology, ZIEL—Institute for Food & Health, Technische Universität München, 85354 Freising, Germany; annemarie.siebert@tum.de (A.S.); katharina.hofmann@tum.de (K.H.); mareike.wenning@lgl.bayern.de (M.W.)
[3] Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit LGL, 85764 Oberschleißheim, Germany
* Correspondence: neuhaus@tum.de
† These authors contributed equally to this work.

Abstract: Full-length SSU rRNA gene sequencing allows species-level identification of the microorganisms present in milk samples. Here, we used bulk-tank raw milk samples of two German dairies and detected, using this method, a great diversity of bacteria, archaea, and yeasts within the samples. Moreover, the species-level classification was improved in comparison to short amplicon sequencing. Therefore, we anticipate that this approach might be useful for the detection of possible mastitis-causing species, as well as for the control of spoilage-associated microorganisms. In a proof of concept, we showed that we were able to identify several putative mastitis-causing or mastitis-associated species such as *Streptococcus uberis*, *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, *Escherichia coli* and *Staphylococcus aureus*, as well as several *Candida* species. Overall, the presented full-length approach for the sequencing of SSU rRNA is easy to conduct, able to be standardized, and allows the screening of microorganisms in labs with Illumina sequencing machines.

Keywords: full-length sequencing; SSU rRNA gene sequencing; milk microbiota; LoopSeq

Citation: Abellan-Schneyder, I.; Siebert, A.; Hofmann, K.; Wenning, M.; Neuhaus, K. Full-Length SSU rRNA Gene Sequencing Allows Species-Level Detection of Bacteria, Archaea, and Yeasts Present in Milk. *Microorganisms* **2021**, 9, 1251. https://doi.org/10.3390/microorganisms9061251

Academic Editor: Chao-Nan Lin

Received: 17 May 2021
Accepted: 5 June 2021
Published: 9 June 2021

1. Introduction

Sequencing has become a reliable and fast method over the years, allowing a time-efficient long-term screening perspective of bacterial communities from many different habitats, allowing the detection of potential pathogens. The 16S rRNA gene comprises nine variable regions (V-regions, V1–V9) that are separated by constant regions [1]. The more stable evolutionary constant regions are used for primer binding. The variable regions, evolved under varying evolutionary processes, are enclosed within a PCR product and are used for taxonomic classification and differentiation [2]. Today, the most frequently used method to study the microbiota of a given sample is short amplicon 16S rRNA gene sequencing, where one, two, or three adjacent variable regions of the 16S rRNA gene are sequenced on a short-read sequencer, e.g., Illumina's MiSeq. The benefits of this technology are the low overall costs, the standardization of the protocols, and the ability to sequence in a high-throughput manner [3]. The drawbacks, on the other hand, are the read-length limitations of 600 bp maximum (due to the short-read sequencers) and the comparability issues of taxonomic profiles when using different short amplicon sequencing protocols or processing pipelines after the sequencing.

Ideas on how to sequence full-length 16S rRNA genes using short-read sequencers like Illumina, together with assembly procedures, have been presented. For instance, Burke and Darling [4] and Karst et al. [5] described methods producing synthetic long-reads after the *de novo* assembly of fragmented 16S rRNA genes sequenced on a short-read sequencer.

A similar approach was presented by Loop Genomic, a Silicon Valley-based company. The reconstruction of full-length molecules is possible after sequencing on a short-read sequencer. The key to their technology is to attach a unique molecular identifier (UMI) to each initial sequence at first, and to subsequently distribute this UMI intramolecularly. The latter is possible, most likely by the usage of a transposase (e.g., [6]) or circularization, followed by enzymatic digestion.

Here, we used the Loop Genomic technology to improve the taxonomical classification down to the genus and species level, and compared this to short amplicon sequencing, either using the V1–V2 or V3–V4 regions. In short, the amplicon sequencing taxonomic resolution is more or less limited to the genus level. The sequencing of milk samples using short amplicon sequencing was performed intensively before, e.g., by Porcellato et al. [7], Taponen et al. [8], Metzger et al. [9], Cremonesi et al. [10], Metzger et al. [11], Pang et al. [12], Doyle et al. [13], Oultram et al. [14], Sokolov et al. [15] and Li et al. [16]. Nevertheless, full-length sequencing approaches are rare and, if published, are mostly performed using long-read sequencers, e.g., Catozzi et al. [17]. In a proof of concept, we assessed whether we could identify putative mastitis pathogens at the species level in random bulk-tank milk samples. We believe better species resolution to be of interest for several other cases, e.g., detecting potential pathogens in milk, or detecting contamination with specific spoilage-related bacteria in dairy products.

As said, we used bovine bulk tank milk samples, but also two mock communities of known composition. The mock communities were included in order to precisely assess the known taxonomic profile of a given sample, and to allow us to draw conclusions about each method. Using the milk samples, we showed the feasibility of the detection of potential pathogens at the species level. In the past, different genera or bacterial species were found to be associated with mastitis, which is defined as an inflammation of the mammary gland [18]. The species most frequently associated with mastitis are non-aureus *Staphylococci*, *Streptococcus uberis*, *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, and also *Staphylococcus aureus*, *Corynebacterium bovis*, *Escherichia coli* and *Klebsiella pneumoniae* [19–23]. Thus, the objectives of this study were (i) to compare short amplicon with full-length 16S rRNA gene sequencing concerning species identification and (ii), as proof-of-concept, to further assess whether we could detect some of these bacteria at the species level in our data set from the milk samples. Furthermore, due to the use of a primer mixture that targets all small subunit (SSU) rRNAs, bacteria, archaea and yeasts were detected at the same time. This enables the collection of additional information of archaeal and eukaryotic microorganisms in the milk microbiota, as these groups of microorganisms are typically not targeted by short amplicon sequencing, which is directed mainly towards bacterial 16S rRNA genes.

2. Materials and Methods

2.1. Milk Samples

Between May and June 2020, the bulk-tank raw milk of different German farms was collected by two dairies and conserved with azidiol (0.33 % *v*/*v*) automatically by the collection trucks. Ten milk samples, each representing a different farm, were selected randomly from farms in Southern Germany participating in a raw milk research project supported by the German Federal Ministry for Food and Agriculture (project number: 281A105616). The dairies providing the milk samples also provided information on the total bacterial counts (measured as colony-forming units, CFU/mL) and the average values of the somatic cell count (SCC) recorded within the last 12 months for each farm. The individual bacterial counts (IBC/mL) were detected via flow cytometry using a BactoCount IBC (Bentley Instruments EU, Maroeuil, France). Sample vials, containing 30–40 mL conserved milk, were shipped and stored refrigerated for a maximum of three days until processing. Of the above-mentioned project, we used the milk sample numbers 796, 797, 798, 879, 880, 978, 979, 980, 982 and 983. Further characteristics of the used raw milk samples were recorded beforehand (Table 1).

Table 1. Characteristics of the bovine raw milk samples used in this study.

Sample Name	Total Bacterial Count (CFU/mL)	Individual Bacterial Count (IBC/mL)	Somatic Cell Count (SCC) per mL (Average Values Detected for Last 12 Months)
796	6.00×10^4	2.23×10^5	not available
797	1.23×10^5	4.69×10^5	not available
798	1.55×10^5	5.97×10^5	not available
879	5.30×10^4	1.95×10^5	247,000
880	2.70×10^4	9.80×10^4	149,000
978	2.90×10^4	1.02×10^5	146,000
979	2.30×10^4	8.20×10^4	90,000
980	2.00×10^4	7.10×10^4	185,000
981	3.20×10^4	1.15×10^5	91,000
983	9.00×10^3	3.20×10^4	96,000

2.2. DNA Extraction of the Bovine Milk

Cells were harvested and DNA was extracted as previously described by Siebert et al. [24]. In brief, 30 mL bovine raw milk were treated with 1.8 mL 0.3 M EDTA. After the cell harvesting by centrifugation (20 min at 4 °C) and the removal of the milk fat and skimmed milk in the supernatant, the selective lysis of the somatic DNA was performed using proteinase K (20 mg/mL, AppliChem GmbH, Darmstadt, Germany) and DNase I (Thermo Fisher Scientific, Waltham, MA, USA). The DNA extraction was performed with the PowerFood Microbial DNA isolation kit (Qiagen, Hilden, Germany), modified by an additional enzymatic lysis step. Towards this end, lysozyme (25 µg/mL, Carl Roth) and mutanolysin (100 U, Sigma-Aldrich, St. Louis, MO, USA) were added to the cell suspensions together with the MBL solution of the DNA isolation kit, followed by an incubation at 37 °C and 350 rpm for 30 min. After an additional treatment with proteinase K (12.5 mg/mL, AppliChem), the remaining bacterial cells were disrupted in tubes with silica beads using a FastPrep-24TM instrument (MP Biomedicals, LLC, Irvine, CA, USA). The subsequent DNA isolation followed the manufacturer's protocol, i.e., that of the PowerFood Microbial DNA isolation kit. The DNA was finally eluted in 2 × 24 µL of PCR-grade water (preheated to 55 °C).

2.3. DNA Extraction of the Mock Communities

The Zymo mock community was purchased as ZymoBIOMICS Microbial Community Standard (D6300, Zymo Research Europe GmbH, Freiburg, Germany) and the DNA was extracted by an adapted protocol originally described by Godon et al. [25], slightly modified. The details can be found in Reitmeier et al. [3]. The ZIEL2 mock community was prepared and extracted as described in Abellan-Schneyder et al. [26]. Briefly, 19 bacterial strains (18 different bacterial genera) of diverse taxonomy were cultured and harvested afterward by centrifugation. A genomic DNA (gDNA) extraction was performed separately for each strain. The ZIEL2 mock community was constructed by pooling 12 ng of each bacterial gDNA.

2.4. Short Amplicon 16S rRNA Gene Library Preparation

For the amplification of the V1–V2 and V3–V4 regions of the 16S rRNA genes, 2-step PCRs of 20 and 10 cycles for the milk samples and 15 and 10 cycles for the mock communities were performed as described in Table S1 (first step PCR), Table S2 (second step PCR) and Abellan-Schneyder et al. [26].

For V1–V2, primers 27F and 338R [27] and for V3-V4, primers 341F and 785R [28] were used (Table S3). Further details and work time estimations can be found in Reitmeier et al. [3].

2.5. Library Quality Check and the Sequencing of the Short Amplicons

The concentrations of the final PCR products were measured in triplicates using a Qubit 4.0 (Thermo Fisher Scientific, Waltham, MA, USA). Each sample was adjusted to 0.5 nM. All of the samples were pooled and sequenced in the paired-end mode for 2 × 300 bp (PE300) using a MiSeq system (Illumina, Inc., San Diego, CA, USA), following the manufacturer's instructions. The final DNA concentration of the library was 12 pM, and 15% (*v*/*v*) PhiX was added.

2.6. Synthetic Long-Read Sequencing Using the LoopSeq 16S & 18S Microbiome Kit

The library preparation was performed as described by LoopGenomics for the LoopSeqTM 16S & 18S Long Read Kit (Version 2.1, Loop Genomics, San Jose, CA, USA). This kit enables us to generate de novo assembled long reads which are able to be sequenced on a short-read sequencer. The de novo assembly is possible due to the intramolecular distribution of a specific unique molecular identifier (UMI), which is unique for every initially tagged full-length molecule. At the same time, the full-length PCR products are split into shorter fragments that can be sequenced using short reads. The exact mechanism of the LoopSeq protocol is proprietary, but possibly relies on a transposase-like enzymatic function. Furthermore, the LoopSeq kit contains primers which can bind to 16S and 18S rRNAs, allowing the estimation of bacteria and archaea together with eukaryotic microorganisms. In brief, 5 µL 1:10-diluted gDNA (in 10 mM TRIS buffer, pH 8.5) was used for the 'Enrichment'. The enrichment PCR was performed for 30 cycles of denaturation, annealing and elongation, as described in Table S4. Instead of using the pre-mixed enrichment primer from the supplier, a custom 2 µM primer mix was used. We ordered four forward primers and two reverse primers, HPLC purified, in 10 µM stock concentrations (biomers.net GmbH, Ulm, Germany). Every primer was diluted with water to 2 µM and added to the final primer mix in equimolar ratios. The sequences of the six different primers were taken from the Loop Genomics primer file 'Genome Amplification Oligonucleotide Sequences', Version 1.0, which is available from Loop Genomics. Concentrations of the 1:10 diluted PCR products (i.e., 'Enrichment') were measured using a Qubit. For all of the samples, concentrations >0.1 ng/µL could be detected after the PCR. Barcodes were assigned to each sample as described by the manufacturer. The barcode calibration was performed in a total volume of 10 µL instead of 20 µL. Based on the calculated sample concentration, the samples were diluted to about 8,000 barcodes per sample. The barcode distribution was performed according to the manual, but concentrations of each sample were assessed using a Qubit before the pooling. The samples were pooled in equimolar amounts based on the PCR-product concentrations. Afterward, the pool was purified using AMPure XP Beads according to the manual. Next, the barcode distribution and library preparation were performed as described in the LoopSeq manual. In this work, Index Primer 1 was used for the pool. The final pool was loaded on an Agilent HS DNA Chip to assess its quality. An average library size of 600 bp was determined and deemed adequate for sequencing. The pool was adjusted to a concentration of 2 nM and prepared for sequencing on an Illumina MiSeq as described by the sequencer's manual. The final library concentration was 12.5 pM, and 5% PhiX was added. The raw short reads were uploaded to the LoopGenomics platform. The results, including the assembled full-length sequences, stats and taxonomy files, could be downloaded from the website after a run time of about 2 h.

2.7. Data Analysis of the Short Reads Using DADA2

The primer sequences of the short amplicon reads were trimmed using cutadapt [29]. Afterward, the samples were processed with the DADA2 pipeline v1.18.0 [30]. The following options were used: paired-end mode, a truncation length of 200/180 bp for V1–V2 and 260/220 bp for V3–V4, maxN = 0, maxEE = 2/2, and trunQ = 2. As the reference database, SILVA v132 (https://www.arb-silva.de/documentation/release-132/, accessed on 2 December 2020) was chosen.

3. Results

3.1. Protocol Overview for the Short and Full-Length 16S rRNA Gene Sequencing

In the present study, we sequenced ten bovine raw milk samples and two mock communities of known composition (Figure 1). The isolated DNA was sequenced on the one hand using short amplicon primers spanning the V-regions V1–V2 and V3–V4 of the 16S rRNA genes (highlighted blue in Figure 1). On the other hand, the samples were processed by using the LoopSeq™ 16S & 18S Long Read Kit (Loop Genomics, San Jose, California, USA, highlighted orange in Figure 1).

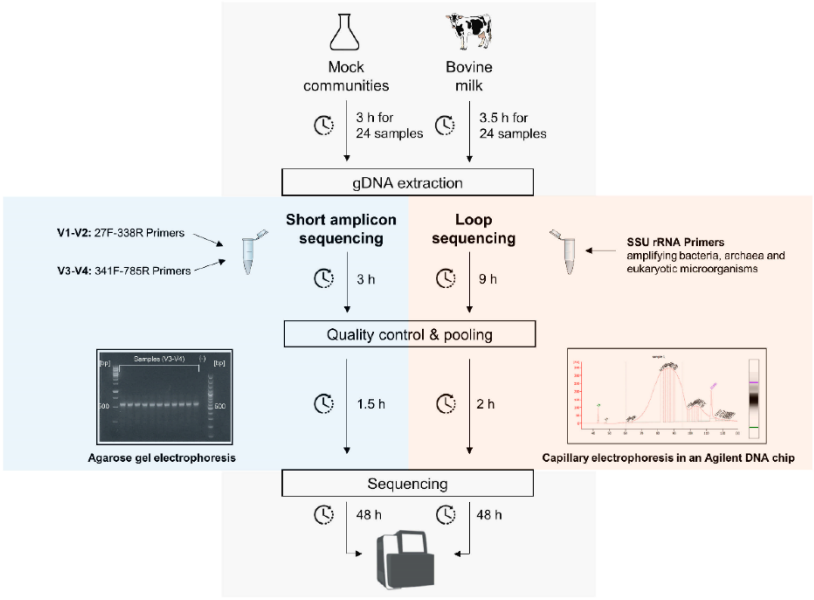

Figure 1. Overview of the experimental procedure. Mock communities of known composition and bovine raw milk samples were used for the microbial gDNA extraction (grey above). The samples were prepared for short amplicon 16S rRNA gene sequencing (blue) and full-length sequencing using the 16S/18S kit targeting the SSU rRNA (orange). After the libraries were prepared, cleaned, and attested to be of good quality, the sequencing was performed on an Illumina MiSeq (grey, below). The execution time estimations in hours are shown for all steps.

3.2. Performance on Mock Communities

Two short amplicon procedures (V1–V2 and V3–V4) and the full-length 16S & 18S long read method (basically including V1–V9 concerning 16S rRNA genes and, thus, designated as such from hereon) were compared using mock communities at first (Figure 2). When analyzing only the less-complex Zymo mock community, which consists of eight different bacterial genera, the distances in the non-metric multidimensional scaling (NMDS) plots were relatively low (Figure 2a, compare the scale to panel d). Generally, distances are considered to be excellent (e.g., in the sense of being negligible) when they are <0.05. Otherwise, they are considered good at <0.1, usable at <0.2, and not acceptable at ≥0.2, i.e., in the sense that the results are 'too different' [31]. These low distances observed here for the Zymo mock were also reflected in the resulting taxonomical profiles (Figure 2b), showing no major deviations between any method used and the ideal theoretic composition. Overall, the Zymo mock performed satisfactorily for all of the sequencing approaches. However, naturally occurring communities are normally poly-species. Thus, the more complex ZIEL2 mock community, which consists of 18 different genera, was used. Here, the results between

the different methods deviated more. Even though V3–V4 and the ideal composition showed short distances, the distances to the ideal composition were larger for V1–V2 and V1–V9 (Figure 2d). This finding was reinforced when the taxonomic profiles were analyzed and compared (Figure 2e). For V1–V2, *Akkermansia*, *Bifidobacterium* and *Enterobacteria* were underrepresented compared to the ideal composition. For V3–V4, *Ruminococcus* and *Microbacterium* were extremely reduced in proportion compared to the ideal mix. Concerning V1–V9, no identification of *Bifidobacterium* and *Collinesella* was possible at all, and a dramatic underrepresentation of *Akkermansia*, *Eggerthella* and *Microbacterium* was observed. Nevertheless, one of the key features of full-length approaches is a species-level classification. Therefore, we analyzed the precision of the species-level classification for each of the different sequencing approaches. For the Zymo mock community, we found that nearly 85% of all reads could be identified down to the species level when using V1–V9. In contrast, only 33% and 5% for V1–V2 and V3–V4, respectively, could be classified at the species level (Figure 2c).

The picture changes when we use the more complex ZIEL2 mock. For this mock community, V1–V9 also performed best, allowing 69.2% of the communities to be identified down to the species level. Interestingly, V1–V2 also did relatively well, allowing up to 58.4% of the bacteria to be identified to the species level. As before, V3–V4 performed worst, i.e., only 34.8% of the species could be identified here (Figure 2f).

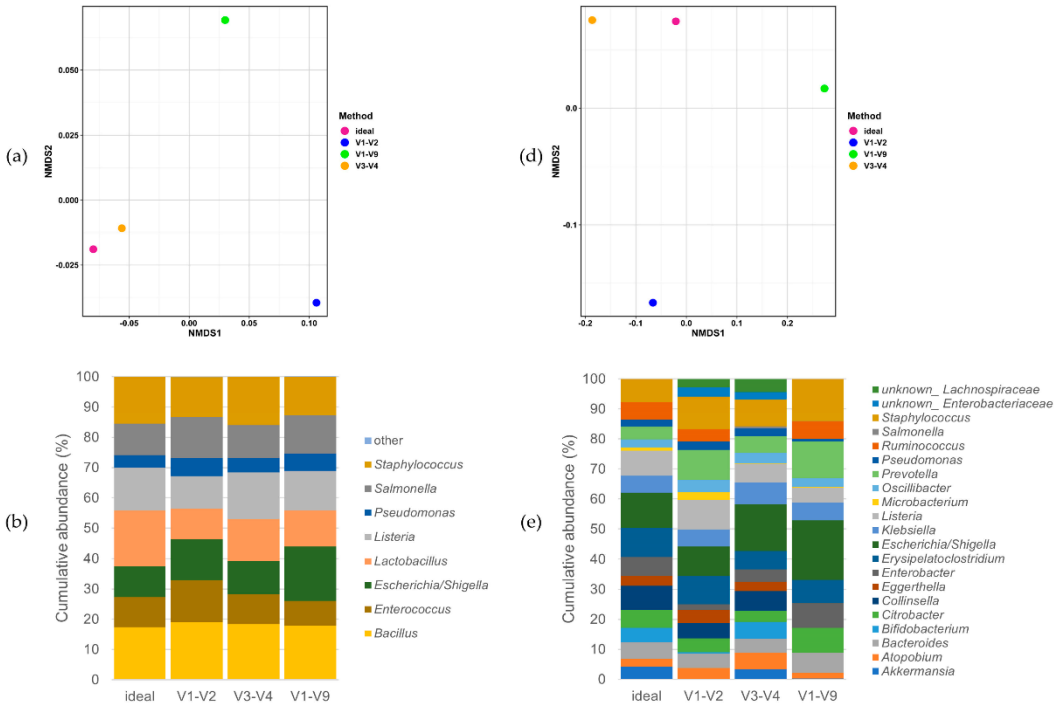

Figure 2. *Cont.*

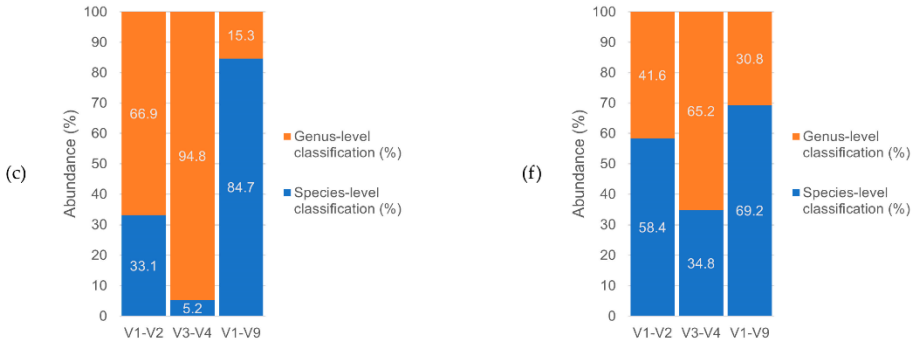

Figure 2. Global performance of three different sequencing procedures using two different mock communities: Zymo (**a**–**c**) and ZIEL2 (**d**–**f**). (**a**) NMDS plot showing the results for the Zymo mock community sequenced with V1–V2 (blue), V3–V4 (orange), and V1–V9 (green). Furthermore, the 'ideal', i.e., the theoretical composition of the Zymo mock is added as an additional data point (violet). (**b**) The relative abundance of bacteria included in the Zymo mock, consisting of eight bacterial genera (and two eukaryotic, not shown). The first column shows the ideal (theoretical) composition, while the following columns show the data gained for V1–V2, V3–V4 and V1–V9. (**c**) Overview for the percentage of the bacteria of the Zymo mock, which could be classified down to the species- (blue) or only to the genus-level (orange) for the three sequencing procedures (V1–V2, V3–V4, and V1–V9). As can be seen, the full-length approach (V1–V9) outperformed the short amplicon approaches (V1–V2 and V3–V4) in species classification. (**d**) NMDS plot showing the results for the ZIEL2 mock community sequenced with V1–V2 (blue), V3–V4 (orange) and V1–V9 (green). Further, the 'ideal', i.e., the theoretical composition of the Zymo mock is added as an additional data point (violet). Please note the larger differences between the data points indicated by the scale compared to panel **a**. (**e**) Relative abundance of bacteria included in the ZIEL2 mock, consisting of 18 bacterial genera. The first column shows the ideal (theoretical) composition, while the following columns show the data gained for V1–V2, V3–V4, and V1–V9. The different sequencing approaches lead to wider differences when comparing the relative abundance in each case to the 'ideal' composition. This was already reflected in the larger distances in the NMDS plot in panel **d**. (**f**) Overview for the percentage of the bacteria of the ZIEL2 mock community, which could be classified down to the species- (blue) or only to the genus-level (orange) for the three sequencing procedures (V1–V2, V3–V4 and V1–V9). The full-length approach classified more reads correctly down to the species-level than the short amplicon sequencing (V1–V2 and V3–V4) approaches.

3.3. Performance on the Bovine Milk Samples for Bacteria Identification

Comparisons of the LoopSeq results with short amplicon profiles were only possible for bacteria, as the short amplicon primers which were used were not optimized for the targeting of archaea and were not suitable for eukaryotes. The five genera/species with the cumulative highest amount of reads for bacteria in the raw milk samples sequenced using V1–V9 (LoopSeq) were *Romboutsia* sp., *Turicibacter* sp., *Paeniclostridium* sp., *Clostridium* sp. and *Streptococcus uberis*. For comparison, in V1–V2, the five bacterial genera with the most reads were *Corynebacterium*, *Romboutsia*, *Streptococcus*, *Turicibacter*, and *Staphylococcus*, and for V3–V4, these were *Romboutsia*, *Streptococcus*, *Corynebacterium*, *Turicibacter* and *Paeniclostridium*.

When analyzing the 50 most prevalent bacterial genera, clustering was not significantly dependent on the targeted V-region or sample origin (Figure 3a). However, the V1–V9 sequenced samples seemed to cluster apart from V1–V2 and V3–V4, which partially overlapped in the clustering and showed overall smaller clustering distances (Figure 3b). Nevertheless, when focusing on each sample analyzed with different methods (e.g., V1–V2, V3–V4 and V1–V9) in the NMDS plot, it becomes clear that changing the method leads to sometimes extensive intra-sample distances (Figure 3c).

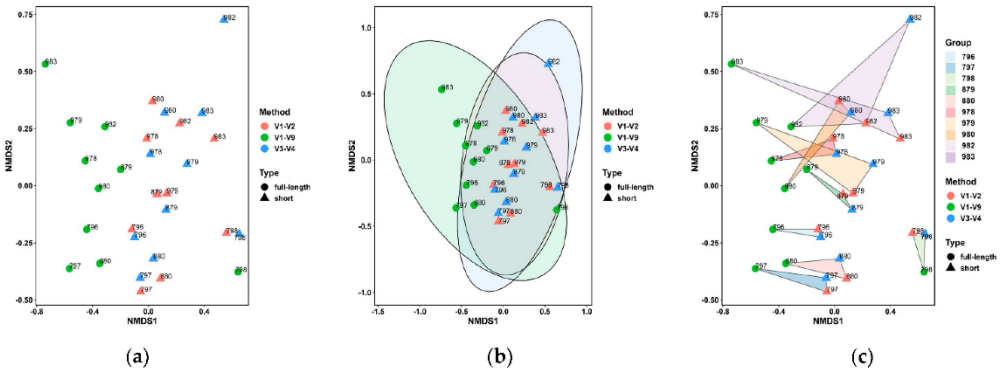

Figure 3. Non-metric multidimensional scaling (NMDS) plots for the comparison of the sequencing results of V1–V2, V3–V4 and V1–V9. The NMDS plots highlight that full-length sequenced samples cluster a little apart from V1–V2 and V3–V4 sequenced samples (**a**). This becomes more evident when samples are grouped by targeted region (**b**). However, differences between sequencings of the same sample using different primer pairs are large and are not only explainable through the targeted region (**c**).

Regardless of the above, taxonomic profiles showed recurrent similarities when the same sample was analyzed (Figure 4). For example, milk samples 798 and 880 had a higher amount of *Streptococcus* (rich orange) compared to the other samples.

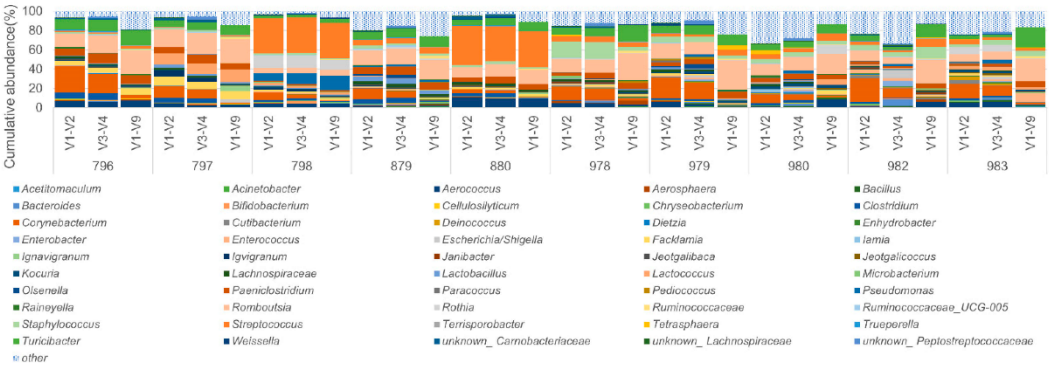

Figure 4. LoopSeq and short amplicon sequencing compared at the genus level for the top 50 bacterial genera. The relative abundances (%) at the genus level of the 10 raw milk samples sequenced by each method (V1–V2, V3–V4 and V1–V9) vary from each other (between samples from different origins) when the 50 most abundant taxa are analyzed. All other remaining taxa are shown in hatched blue.

3.4. Identification of Archaea and Eukaryotes in Bovine Milk Samples Using the LoopSeq Protocol

Most interestingly, when analyzing the samples using the 16S & 18S Long Read Kit, referred to as V1–V9 before, not only bacteria but also archaea and eukaryotes could be identified. When analyzing the taxonomical composition at the kingdom level, sample 979 was shown to have the highest number of reads mapping to eukaryotes, and sample 796 had the highest number of archaeal reads (Figure 5). Nevertheless, the amount of eukaryotic and archaeal contribution to the raw milk microbiota was mostly low except for samples 796, 879 and 979. Thus, the milk samples were shown to be diverse and have unique taxonomic profiles.

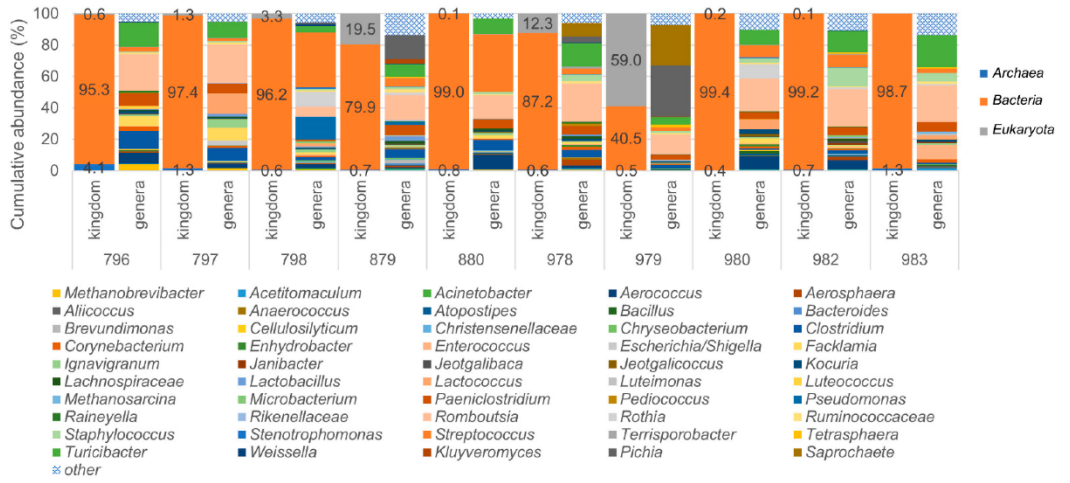

Figure 5. Relative abundance (%) at the kingdom and genus levels after the 16S/18S rRNA (SSU) gene sequencing analysis using the LoopGenomics kit. Taxonomic classification regarding the kingdom (**left**) and genus-level of the top 50 taxa (**right**) of each LoopSeq-sequenced milk sample. All of the other remaining taxa are shown in hatched blue.

In total, we were able to identify 41 different eukaryotic genera and 52 eukaryotic species, as well as seven archaeal genera and seven different archaeal species in the milk samples tested. For eukaryotes, *Pichia* (54.2%), *Saprochaete* (39.5%) and *Kluyveromyces* (2.9%) were by far the most dominant genera, whereas for archaea, these were *Methanobrevibacter* (80.6%) and *Methanosarcina* (16.7%). The overall top five hits in either kingdom are listed in Table 2.

Table 2. Top five genera/species detected in the raw milk samples for archaea and eukaryotes.

Top Hit	*Archaea*	*Eukaryotes*
Top 1	*Methanobrevibacter* sp.	*Pichia scutulata*
Top 2	*Methanobrevibacter millerae*	*Saprochaete clavata*
Top 3	*Methanosarcina soligelidi*	*Pichia cactophila*
Top 4	*Methanosarcina mazei*	*Kluyveromyces marxianus*
Top 5	*Methanosarcina horonobensis*	*Pichia pseudocactophila*

3.5. Identification of Putative Mastitis-Causing Pathogens

Concerning archaea, *Methanobrevibacter*, *Methanocorpusculum*, *Methanosphaera* and *Methanomassiliicoccus* are the most common archaeal bovine-associated genera [32,33]. Of those, we could identify *Methanobrevibacter*, *Methanocorpusculum*, and *Methanosphaera*. For archaea, no association with mastitis is known. In contrast, for eukaryotes, *Cryptococcus neoformans*, *Candida albicans* and other *Candida* species are known to be potential mastitis-causers [34]. A cultivation-based study from 2008, which analyzed Brazilian subclinical mastitis milk samples, found that *Candida*, *Pichia*, *Cryptococcus* and *Rhodotorula* were the most frequent genera. However, they could not be directly linked to the subclinical mastitis state in the mentioned publication [35]. Regarding *Cryptococcus* and *Candida*, we were unable to identify *Cryptococcus* reads in any one of the milk samples, but for *Candida*, we identified *Candida boidinii*, *Candida metapsilosis*, *Candida intermedia*, *Candida zeylanoides* and *Candida pararugosa*.

Next, it was investigated whether putative mastitis-causing bacteria could be identified down to the species level. Towards this end, it was checked whether the species listed as mastitis-causing pathogens by Cobirka et al. [20] could be found in our dataset

(Table 3). From the 25 genera/species listed by Cobirka et al. [20], we could identify 19 here. Admittedly, five species (*Enterococcus durans*, *Enterococcus faecium*, *Klebsiella oxytoca*, *Serratia* spp. and *Staphylococcus simulans*) were only identified in very low amounts of <10 reads.

Table 3. Detection of species suspected to be mastitis-causing. The species are adapted from Cobirka et al. [20].

Species	Found in Raw Milk Samples	Detected on Genus Level in Raw Milk Samples	Found in Mock Communities
Arcanobacterium/ Trueperella pyogenes	yes	*Trueperella*	no
Corynebacterium bovis	no	*Corynebacterium*	no
Enterobacter aerogenes	no	no	no
Enterococcus durans	yes	*Enterococcus*	no
Enterococcus faecalis	yes		no
Enterococcus faecium	yes		no
Escherichia coli	yes	*Escherichia/Shigella*	*Escherichia coli*
Klebsiella oxytoca	yes	*Klebsiella*	no
Klebsiella pneumoniae	no		*Klebsiella pneumoniae*
Mycoplasma bovis	no	no	no
Proteus spp. (*)	no	no	no
Pseudomonas aeruginosa	no	*Pseudomonas*	*Pseudomonas aeruginosa*
Serratia marcescens (*)	yes	*Serratia*	no
Staphylococcus aureus	yes	*Staphylococcus*	*Staphylococcus aureus*
Staphylococcus chromogenes	yes		no
Staphylococcus epidermidis	yes		*Staphylococcus epidermidis*
Staphylococcus haemolyticus	yes		no
Staphylococcus sciuri	yes		no
Staphylococcus simulans	yes		no
Streptococcus agalactiae	yes	*Streptococcus*	no
Streptococcus bovis	no		no
Streptococcus dysgalactiae	yes		no
Streptococcus equinus	yes		no
Streptococcus uberis	yes		no
Yersinia spp. (*)	no	no	no

* As, e.g., all *Serratia* spp. are listed according to Cobirka et al. [20], all of the species of this genus were considered to be possible mastitis-causing bacteria.

4. Discussion

4.1. Full-Length SSU rRNA Gene Sequencing Improves Species-Level Classification but Shows Primer Issues

A variety of different factors can bias sequencing approaches targeting the 16S rRNA gene. The most well-known and studied are the sample collection, stabilization and transport to the laboratory, DNA extraction method, DNA concentration, primers targeting different V-regions, PCR condition and settings, laboratory practice, analysis pipelines, and reference databases [24,26,36–39]. In our results, it became obvious that, e.g., primer pairs 27F and 338R underrepresent *Akkermansia*, whereas this is not the case for 338F and 785R. Because the LoopSeq kit uses the forward 27F primer, low amounts of *Akkermansia* and *Bifidobacterium* were expected, similar to V1–V2, which also used the 27F-region for priming. However, the failure of the detection of *Collinsella* and *Eggerthella* in the LoopSeq results could not be explained by using this particular forward primer. In contrast, it was shown in a previous study that the use of primer pair 1115F and 1492R led to an underrepresentation of those two genera [26], which might suggest that the reverse primer 1492R is responsible for this inferior result concerning the latter two genera. This highlights that studies must

carefully test whether the primers used are suitable for the expected results. Furthermore, improved full-length primer pairs or strategies such as the rRNA full-length approach of Karst et al. [5] are needed. The method of Karst et al. is primer-independent because it starts from the actual rRNA molecule, which is reverse transcribed in cDNA, and not from rRNA genes.

Overall, we suggest that the LoopSeq approach could be improved by enhancing the 16S and 18S targeting primers. For instance, the current forward primer 1 is the commonly used 27F-CM. It has already been shown [40] that this primer poses three mismatches when *Bifidobacteria* should be amplified, and thus shows a decreased binding towards the 16S rRNA genes of this genus. Accordingly, we saw a dramatic underrepresentation of *Bifidobacteria* in the ZIEL2 mock (Figure 2E). Perhaps the primer mix should include further or other 27F-based primers (e.g., 27F-YM) that are improved in bacterial targeting.

We want to emphasize a well-thought-out study design and the need for sufficiently complex mock communities, as we have shown that the Zymo mock is too simple to illustrate possible biasing effects (Figure 2). This becomes even more important when low biomass samples are analyzed. It was previously shown that milk samples are of low bacterial biomass, and are therefore prone to be contaminated by bacteria from the environment [27,38,41]. Thus, controls and mock communities of sufficient complexity should always be sequenced in order to secure the reliability of a study. Eisenhofer et al. [42] published a well-thought-out list of several methods which can be used to minimize the influence of contaminant DNA on low bacterial biomass samples that should be taken into account. To name a few, the use of controls, suitable protective clothing including gloves, masks and clean suits, decontamination and cleaning steps, and protection steps during the sample processing like the use of unique barcodes should be considered [42].

We showed that by using a full-length SSU rRNA sequencing method, species-level identification clearly is increased (Figure 2). Nonetheless, to further improve species-level classification, we see a need for higher resolution environment-specific databases, such as those described by Dueholm et al. [43] and Escapa et al. [44], allowing precise taxonomic comparison and classification. Escapa et al. could, for example, show that the training of the Ribosomal Database Project (RDP) classifier using a habitat-specific training set improved the taxonomical assignment for short- as well as long-read sequences at the species-level. Such bioinformatical developments, besides methodological improvements (e.g., using full-length strategies), will increase the overall amount of reads which can be classified down to the species level. Improvements in species-level classification have also been shown by Jeong et al. [45] for the use of the LoopSeq approach on human fecal samples. Furthermore, for those samples, the taxonomic resolution was improved when compared to short amplicon V3–V4 protocols while analyzing *alpha*-diversity, relative abundance frequency and identification accuracy.

4.2. Using Full-Length Sequencing Approaches for Microbial Monitoring

Microbial monitoring using sequencing-based approaches can facilitate and allow for the detection of possible contaminants in food samples. Thus, we assessed, in a proof-of-principle, whether we could detect putative mastitis pathogens in our sample set. Importantly, in our bulk tank milk samples, we found 17 out of 25 listed putative mastitis-causing bacteria in the full-length data [20]. For instance, *S. uberis* was found in the highest amounts among the analyzed species. However, most of the reads were contributed by three samples (880, 504 reads; 978, 422 reads; and 983, 422 reads), which made up 80% of all of the combined *S. uberis* reads found in all ten milk samples. Thus, we believe that a future study including samples of known (sub-)clinical mastitis cases is of further interest. Concerning *S. uberis*, for example, it is yet unclear which amount might be tolerated. Possibly, a threshold in either relative or absolute abundance, or the relative abundances between different mastitis-causing bacteria (i.e., a dysbiotic state) must be defined for these organisms. Mastitis-causing pathogens are often found to be opportunistic and, therefore, the mere presence of those species currently does not allow

us to draw conclusions about a possible state of inflammation or disease. Nevertheless, full-length SSU rRNA gene sequencing allows relative quantifications, which could be used to determine dysbiotic states.

As our study design was intended to be a proof-of-principle concerning a species-level detection of potentially pathogenic bacteria in milk, we assessed whether we could find such species in our dataset. However, these bacteria could also be simple contaminants, as previously reported [12,14,46–49]. Nevertheless, in many of those previously performed studies, short-amplicon sequencing strategies were applied targeting either V1-V2, V3-V4 or V4 alone. These studies are, therefore, limited in their taxonomic resolution. In contrast, targeting the full-length SSU rRNA gene helps to identify bacteria, archaea, and eukaryotes at improved taxonomic levels. For instance, Catozzi et al. [17] published a full-length 16S rRNA sequencing strategy for the milk microbiota of water buffalos. In this study, the authors demonstrated that full-length strategies are suitable for species-level detection. Nevertheless, their strategy had some drawbacks, such as, for instance, a higher error rate of the reads obtained by using a MinION sequencer, comparability issues due to the use of different reference databases, and difficulties in processing the raw reads of this sequencer. In contrast, the MiSeq sequencer of Illumina used in our study has much lower error rates in sequencing and is the most widely available short-read sequencer. Furthermore, the LoopGenomics pipeline conducts the pre-processing into full-length reads, which is easy to use. Next, the reads are identified using the SILVA database as a reference for both short and full-length sequences, which currently is one of the best databases available for 16S rRNA sequences [26]. In addition, as previously stated, not only bacteria but also eukaryotic and archaeal microorganisms are discovered due to the primer mix included. This might be of importance for further studies that want to assess the impact of yeast-associated mastitis, such as those performed by previous studies, which found that even though yeast-associated mastitis is rare, it could be of importance for some clinical cases of intramammary infections [34,50]. Besides this, the species-level identification of milk-associated microorganisms is of great interest for animal husbandry and dairies in general.

One further example in this mentioned respect is the identification and differentiation of *Pseudomonas* spp., which are often associated with milk spoilage [51]. In our study, using the LoopSeq approach, we could identify *P. brenneri*, *P. canadensis*, *P. fluorescens*, *P. gessardii*, *P. helleri*, *P. lundensis*, *P. mucidolens*, *P. pseudoalcaligenes*, *P. putida* and *P. rhizosphaerae*, some of which are known to be proteolytic *Pseudomonas* spp. that occur frequently in retail milk [52]. In contrast, short amplicon data were not useful for the identification of *Pseudomonas* at the species-level, neither in V1–V2 nor in V3–V4 data, with only a questionable *Pseudomonas lurida*, because this species is not found in the full-length data. Most probably, this is a misclassification corresponding to *P. fluorescens* identified in the LoopSeq data, as *P. lurida* and *P. fluorescens* differ by only one nucleotide in the V3–V4 part of their 16S rRNA genes.

5. Conclusions

Using the 16S/18S LoopSeq kit suitable for Illumina sequencing, we could not only identify bacteria at the species level but also the archaeal and eukaryotic microorganisms present in raw milk samples. The number of eukaryotic and archaeal reads varied from sample to sample, accounting for up to over 50% of all of the reads. Obviously, the bovine milk microbiome is highly diverse and different from sample to sample [7,12,53], which is reinforced by our study. The advantage of full-length SSU rRNA gene sequencing over short amplicon sequencing approaches is an improved species-level classification, as well as the simultaneous analysis of not only the bacteria present in the sample but also the identification of archaeal and eukaryotic species. Moreover, the LoopSeq kit as a commercial product allows for easy standardization across labs, and the downstream pipelines allow simple and convenient analysis with little bioinformatic knowledge required from the user.

Supplementary Materials: The Supplementary Materials can be found at https://www.mdpi.com/article/10.3390/microorganisms9061251/s1. Table S1: PCR condition for the first step short amplicon

16S PCR. Table S2: PCR condition for the step short amplicon 16S PCR. Table S3: 16S rRNA gene short amplicon primer sequences. Table S4: PCR condition for LoopSeq Enrichment PCR. Table S5: List of the complete taxonomy of microorganisms detected within the milk samples processed using the full-length SSU rRNA gene sequencing approach (Excel table).

Author Contributions: Conceptualization, I.A.-S. and K.N.; methodology, I.A.-S., A.S. and K.H.; software, I.A.-S. and K.N.; validation, I.A.-S. and K.N.; formal analysis, I.A.-S.; investigation, I.A.-S. and K.N.; resources, K.N. and M.W.; data curation, I.A.-S. and K.N.; writing—original draft preparation, I.A.-S.; writing—review and editing, A.S., K.H., M.W. and K.N.; visualization, I.A.-S.; supervision, K.N. and M.W.; project administration, K.N.; funding acquisition, M.W. and K.N. All authors have read and agreed to the published version of the manuscript.

Funding: I.A.-S. was funded by the ZIEL—Institute for Food & Health with a grant for a doctorate position, and was partially funded by a grant from the Research Foundation of Dairy Science at the Technical University of Munich (VFMF), both given to K.N. The project was supported by funds from the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program (281A105616) given to M.W.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw sequencing data (for the short amplicon sequencing) and the de-multiplexed fastq files of the contigs (for LoopSeq) are available at the Sequence Read Archive within the BioProject PRJNA719984.

Acknowledgments: We want to thank Angela Sachsenhauser, Lukas Mix and Caroline Ziegler for their excellent technical assistance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Janda, J.M.; Abbott, S.L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* **2007**, *45*, 2761–2764. [CrossRef]
2. Patel, J.B. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol. Diagn.* **2001**, *6*, 313–321. [CrossRef]
3. Reitmeier, S.; Kiessling, S.; Neuhaus, K.; Haller, D. Comparing Circadian Rhythmicity in the Human Gut Microbiome. *STAR Protoc.* **2020**, 100148. [CrossRef] [PubMed]
4. Burke, C.M.; Darling, A.E. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ* **2016**, *4*, e2492. [CrossRef]
5. Karst, S.M.; Dueholm, M.S.; McIlroy, S.J.; Kirkegaard, R.H.; Nielsen, P.H.; Albertsen, M. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **2018**, *36*, 190–195. [CrossRef] [PubMed]
6. Chandler, M. Prokaryotic DNA Transposons: Classes and Mechanism. *eLS* **2017**, 1–16. [CrossRef]
7. Porcellato, D.; Meisal, R.; Bombelli, A.; Narvhus, J.A. A core microbiota dominates a rich microbial diversity in the bovine udder and may indicate presence of dysbiosis. *Sci. Rep.* **2020**, *10*, 21608. [CrossRef] [PubMed]
8. Taponen, S.; McGuinness, D.; Hiitiö, H.; Simojoki, H.; Zadoks, R.; Pyörälä, S. Bovine milk microbiome: A more complex issue than expected. *Vet. Res.* **2019**, *50*, 44. [CrossRef]
9. Metzger, S.A.; Hernandez, L.L.; Skarlupka, J.H.; Walker, T.M.; Suen, G.; Ruegg, P.L. A Cohort Study of the Milk Microbiota of Healthy and Inflamed Bovine Mammary Glands From Dryoff Through 150 Days in Milk. *Front. Vet. Sci.* **2018**, *5*, 247. [CrossRef] [PubMed]
10. Cremonesi, P.; Ceccarani, C.; Curone, G.; Severgnini, M.; Pollera, C.; Bronzo, V.; Riva, F.; Addis, M.F.; Filipe, J.; Amadori, M.; et al. Milk microbiome diversity and bacterial group prevalence in a comparison between healthy Holstein Friesian and Rendena cows. *PLoS ONE* **2018**, *13*, e0205054. [CrossRef]
11. Metzger, S.A.; Hernandez, L.L.; Skarlupka, J.H.; Suen, G.; Walker, T.M.; Ruegg, P.L. Influence of sampling technique and bedding type on the milk microbiota: Results of a pilot study. *J. Dairy Sci.* **2018**, *101*, 6346–6356. [CrossRef] [PubMed]
12. Pang, M.; Xie, X.; Bao, H.; Sun, L.; He, T.; Zhao, H.; Zhou, Y.; Zhang, L.; Zhang, H.; Wei, R.; et al. Insights Into the Bovine Milk Microbiota in Dairy Farms With Different Incidence Rates of Subclinical Mastitis. *Front. Microbiol.* **2018**, *9*, 2379. [CrossRef]
13. Doyle, C.J.; Gleeson, D.; O'Toole, P.W.; Cotter, P.D. High-throughput metataxonomic characterization of the raw milk microbiota identifies changes reflecting lactation stage and storage conditions. *Int. J. Food Microbiol.* **2017**, *255*, 1–6. [CrossRef] [PubMed]

14. Oultram, J.W.H.; Ganda, E.K.; Boulding, S.C.; Bicalho, R.C.; Oikonomou, G. A Metataxonomic Approach Could Be Considered for Cattle Clinical Mastitis Diagnostics. *Front. Vet. Sci.* **2017**, *4*, 36. [CrossRef]
15. Sokolov, S.; Fursova, K.; Shulcheva, I.; Nikanova, D.; Artyemieva, O.; Kolodina, E.; Sorokin, A.; Dzhelyadin, T.; Shchannikova, M.; Shepelyakovskaya, A.; et al. Comparative Analysis of Milk Microbiomes and Their Association with Bovine Mastitis in Two Farms in Central Russia. *Animals* **2021**, *11*, 1401. [CrossRef]
16. Li, N.; Wang, Y.; You, C.; Ren, J.; Chen, W.; Zheng, H.; Liu, Z. Variation in Raw Milk Microbiota Throughout 12 Months and the Impact of Weather Conditions. *Sci. Rep.* **2018**, *8*, 2371. [CrossRef]
17. Catozzi, C.; Ceciliani, F.; Lecchi, C.; Talenti, A.; Vecchio, D.; De Carlo, E.; Grassi, C.; Sánchez, A.; Francino, O.; Cuscó, A. Short communication: Milk microbiota profiling on water buffalo with full-length 16S rRNA using nanopore sequencing. *J. Dairy Sci.* **2020**, *103*, 2693–2700. [CrossRef]
18. Contreras, G.A.; Rodríguez, J.M. Mastitis: Comparative Etiology and Epidemiology. *J. Mammary Gland Biol. Neoplasia* **2011**, *16*, 339–356. [CrossRef] [PubMed]
19. Dufour, S.; Labrie, J.; Jacques, M. The Mastitis Pathogens Culture Collection. *Microbiol. Resour. Announc.* **2019**, *8*, e00133-19. [CrossRef]
20. Cobirka, M.; Tancin, V.; Slama, P. Epidemiology and Classification of Mastitis. *Animals* **2020**, *10*, 2212. [CrossRef] [PubMed]
21. Bolte, J.; Zhang, Y.; Wente, N.; Krömker, V. In Vitro Susceptibility of Mastitis Pathogens Isolated from Clinical Mastitis Cases on Northern German Dairy Farms. *Vet. Sci.* **2020**, *7*, 10. [CrossRef]
22. Dalanezi, F.M.; Joaquim, S.F.; Guimarães, F.F.; Guerra, S.T.; Lopes, B.C.; Schmidt, E.M.S.; Cerri, R.L.A.; Langoni, H. Influence of pathogens causing clinical mastitis on reproductive variables of dairy cows. *J. Dairy Sci.* **2020**, *103*, 3648–3655. [CrossRef] [PubMed]
23. Traversari, J.; van den Borne, B.H.P.; Dolder, C.; Thomann, A.; Perreten, V.; Bodmer, M. Non-aureus Staphylococci Species in the Teat Canal and Milk in Four Commercial Swiss Dairy Herds. *Front. Vet. Sci.* **2019**, *6*, 186. [CrossRef]
24. Siebert, A.; Hofmann, K.; Staib, L.; Doll, E.V.; Scherer, S.; Wenning, M. Amplicon-sequencing of raw milk microbiota: Impact of DNA extraction and library-PCR. *Appl. Microbiol. Biotechnol.* **2021**. [CrossRef] [PubMed]
25. Godon, J.J.; Zumstein, E.; Dabert, P.; Habouzit, F.; Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol.* **1997**, *63*, 2802–2813. [CrossRef] [PubMed]
26. Abellan-Schneyder, I.; Matchado, M.S.; Reitmeier, S.; Sommer, A.; Sewald, Z.; Baumbach, J.; List, M.; Neuhaus, K. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *Msphere* **2021**, *6*, e01202-20. [CrossRef]
27. Salter, S.J.; Cox, M.J.; Turek, E.M.; Calus, S.T.; Cookson, W.O.; Moffatt, M.F.; Turner, P.; Parkhill, J.; Loman, N.J.; Walker, A.W. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **2014**, *12*, 87. [CrossRef]
28. Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glöckner, F.O. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **2012**, *41*, e1. [CrossRef]
29. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 3. [CrossRef]
30. Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]
31. Clarke, K.R. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **1993**, *18*, 117–143. [CrossRef]
32. Cendron, F.; Niero, G.; Carlino, G.; Penasa, M.; Cassandro, M. Characterizing the fecal bacteria and archaea community of heifers and lactating cows through 16S rRNA next-generation sequencing. *J. Appl. Genet.* **2020**, *61*, 593–605. [CrossRef]
33. Zhu, Z.; Kristensen, L.; Difford, G.F.; Poulsen, M.; Noel, S.J.; Abu Al-Soud, W.; Sørensen, S.J.; Lassen, J.; Løvendahl, P.; Højberg, O. Changes in rumen bacterial and archaeal communities over the transition period in primiparous Holstein dairy cows. *J. Dairy Sci.* **2018**, *101*, 9847–9862. [CrossRef] [PubMed]
34. Dworecka-Kaszak, B.; Krutkiewicz, A.; Szopa, D.; Kleczkowski, M.; Biegańska, M. High prevalence of *Candida* yeast in milk samples from cows suffering from mastitis in poland. *Sci. World J.* **2012**, *2012*, 196347. [CrossRef]
35. Spanamberg, A.; Augusto Wünder, E.; Isabel Brayer Pereira, D.; Argenta, J.; Maria Cavallini Sanches, E.; Valente, P.; Ferreiro, L. Etiología de la mastitis bovina producida por levaduras en el sur de Brasil. *Rev. Iberoam. De Micol.* **2008**, *25*, 154–156. [CrossRef]
36. Usman, T.; Yu, Y.; Liu, C.; Fan, Z.; Wang, Y. Comparison of methods for high quantity and quality genomic DNA extraction from raw cow milk. *Genet. Mol. Res.* **2014**, *13*, 3319–3328. [CrossRef] [PubMed]
37. Kennang Ouamba, A.J.; LaPointe, G.; Dufour, S.; Roy, D. Optimization of Preservation Methods Allows Deeper Insights into Changes of Raw Milk Microbiota. *Microorganisms* **2020**, *8*, 368. [CrossRef]
38. Dahlberg, J.; Sun, L.; Persson Waller, K.; Östensson, K.; McGuire, M.; Agenäs, S.; Dicksved, J. Microbiota data from low biomass milk samples is markedly affected by laboratory and reagent contamination. *PLoS ONE* **2019**, *14*, e0218257. [CrossRef]
39. Xue, Z.; Kable, M.E.; Marco, M.L. Impact of DNA Sequencing and Analysis Methods on 16S rRNA Gene Bacterial Community Analysis of Dairy Products. *Msphere* **2018**, *3*. [CrossRef]
40. Walker, A.W.; Martin, J.C.; Scott, P.; Parkhill, J.; Flint, H.J.; Scott, K.P. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **2015**, *3*, 26. [CrossRef]
41. McHugh, A.J.; Yap, M.; Crispie, F.; Feehily, C.; Hill, C.; Cotter, P.D. Microbiome-based environmental monitoring of a dairy processing facility highlights the challenges associated with low microbial-load samples. *NPJ Sci. Food* **2021**, *5*, 4. [CrossRef]

42. Eisenhofer, R.; Minich, J.J.; Marotz, C.; Cooper, A.; Knight, R.; Weyrich, L.S. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol.* **2019**, *27*, 105–117. [CrossRef] [PubMed]
43. Dueholm, M.S.; Andersen, K.S.; McIlroy, S.J.; Kristensen, J.M.; Yashiro, E.; Karst, S.M.; Albertsen, M.; Nielsen, P.H. Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). *MBio* **2020**, *11*, e01557-20. [CrossRef]
44. Escapa, I.F.; Huang, Y.; Chen, T.; Lin, M.; Kokaras, A.; Dewhirst, F.E.; Lemon, K.P. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* **2020**, *8*, 65. [CrossRef]
45. Jeong, J.; Yun, K.; Mun, S.; Chung, W.-H.; Choi, S.-Y.; Nam, Y.-d.; Lim, M.Y.; Hong, C.P.; Park, C.; Ahn, Y.; et al. The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. *Sci. Rep.* **2021**, *11*, 1727. [CrossRef] [PubMed]
46. Oikonomou, G.; Bicalho, M.L.; Meira, E.; Rossi, R.E.; Foditsch, C.; Machado, V.S.; Teixeira, A.G.V.; Santisteban, C.; Schukken, Y.H.; Bicalho, R.C. Microbiota of Cow's Milk; Distinguishing Healthy, Sub-Clinically and Clinically Diseased Quarters. *PLoS ONE* **2014**, *9*, e85904. [CrossRef]
47. Kuehn, J.S.; Gorden, P.J.; Munro, D.; Rong, R.; Dong, Q.; Plummer, P.J.; Wang, C.; Phillips, G.J. Bacterial community profiling of milk samples as a means to understand culture-negative bovine clinical mastitis. *PLoS ONE* **2013**, *8*, e61959. [CrossRef] [PubMed]
48. Wang, Y.; Nan, X.; Zhao, Y.; Jiang, L.; Wang, M.; Wang, H.; Zhang, F.; Xue, F.; Hua, D.; Liu, J.; et al. Rumen microbiome structure and metabolites activity in dairy cows with clinical and subclinical mastitis. *J. Anim. Sci. Biotechnol.* **2021**, *12*, 36. [CrossRef] [PubMed]
49. Hamel, J.; Zhang, Y.; Wente, N.; Krömker, V. Non-*S. aureus* staphylococci (NAS) in milk samples: Infection or contamination? *Vet. Microbiol.* **2020**, *242*, 108594. [CrossRef]
50. Zaragoza, C.S.; Olivares, R.A.; Watty, A.E.; Moctezuma Ade, L.; Tanaca, L.V. Yeasts isolation from bovine mammary glands under different mastitis status in the Mexican High Plateu. *Rev. Iberoam Micol.* **2011**, *28*, 79–82. [CrossRef]
51. Meng, L.; Zhang, Y.; Liu, H.; Zhao, S.; Wang, J.; Zheng, N. Characterization of *Pseudomonas* spp. and Associated Proteolytic Properties in Raw Milk Stored at Low Temperatures. *Front. Microbiol.* **2017**, *8*, 2158. [CrossRef] [PubMed]
52. Maier, C.; Hofmann, K.; Huptas, C.; Scherer, S.; Wenning, M.; Lücking, G. Simultaneous quantification of the most common and proteolytic *Pseudomonas* species in raw milk by multiplex qPCR. *Appl. Microbiol. Biotechnol.* **2021**, *105*, 1693–1708. [CrossRef] [PubMed]
53. Addis, M.F.; Tanca, A.; Uzzau, S.; Oikonomou, G.; Bicalho, R.C.; Moroni, P. The bovine milk microbiota: Insights and perspectives from -omics studies. *Mol. Biosyst.* **2016**, *12*, 2359–2372. [CrossRef] [PubMed]

Supplementary Materials

Table S1. PCR condition for 1st step short amplicon 16S PCR.

PCR step		Temperature		Time	Step
Initial Denaturation		98 °C		2 min	1 cycle
Denaturation		98 °C		10 s	
Annealing	V1V2:	57 °C	milk:	30 s	20 cycles
	V3–V4:	55 °C	mock:	30 s	15 cycles
Extension		72 °C	milk:	90 s	
		72 °C	mock:	40 s	
Final Extension		72 °C		2 min	1 cycle
Storage		4 °C		hold	

Table S2. PCR condition for 2nd step short amplicon 16S PCR.

PCR step	Temperature	Time	Step
Initial Denaturation	98 °C	40 s	1 cycle
Denaturation	98 °C	20 s	
Annealing	55 °C	40 s	10 cycles
Extension	72 °C	40 s	
Final Extension	72 °C	2 min	1 cycle
Storage	4° C	hold	

Table S3. 16S rRNA gene short amplicon primer sequences.

Targeted region	Forward primer (5′–3′)	Reverse primer (5′–3′)	Reference
V1–V2	AGA GTT TGA TYM TGG CTC AG	GCT GCC TCC CGT AGG AGT	Salter, et al. [40]
V3–V4	CCT ACG GGN GGC WGC AG	GAC TAC HVG GGT ATC TAA TCC	Klindworth, et al. [52]

Table S4. PCR condition for LoopSeq Enrichment PCR.

PCR step	Temperature	Time	Step	Ramp speed
Initial Denaturation	95 °C	3 min	1 cycle	2 °C/s
Denaturation	98 °C	15 s		
Annealing	52 °C	20 s	30 cycles	2 °C/s
Extension	72 °C	2 min		
Storage	4 °C	hold		2 °C/s

Table S5. List of complete taxonomy of microorganisms detected within the milk samples processed using the full-length SSU rRNA gene sequencing approach (Excel table).

3. General Discussion and Conclusion

Even though the terms “microbiota” and “microbiome” are often used interchangeably in the last couple of years, differences exist. While microbiota describes the living organisms of a defined ecosystem, the microbiome includes, besides the microbiota, also the molecules produced by the microorganisms (Berg *et al.*, 2020, Marchesi and Ravel, 2015). The interchangeably use of both terms reflects a general standardization problem.

While conducting my thesis, I was wondering if the interchangeable and sometime carless use of both terms indicates problems exceeding terminology. Indeed, while 16S rRNA sequencing seems to be a standard technology for about 20 years, it seems that many, even fundamental questions about this technology, have not been answered profoundly and have neither be standardized. In my thesis, different sequencing approaches (including short, long, and synthetic long sequence approaches) were tested, created, evaluated, and compared. After carefully analyzing my data, I see further potential and need to standardize not only terminology but also several other factors in the field of microbiota and microbiome research.

Already in 2011, the European Commission decided to fund an international project called the “International Human Microbiome Standards” (IHMS), which aimed to develop protocols that will allow and facilitate data comparability between different scientific studies performed in the field of human microbiome research. The consortium published within the first four years several standard operating procedures (SOPs), aiming to guarantee best practices for sample collection, identification, and extraction, as well as for sequencing and data analysis (Guarner *et al.*, 2015). Nonetheless, those SOPs must followed strictly, which is difficult and will probably not be feasible (Editorials, 2016). In order to evaluate comparability of different labs, the Microbiome Quality Control (MBQC) project sent different sample types (human stool samples, samples from chemostats, and mock communities) to 15 different laboratories and compared the results produced. Most differences in the taxonomical profiles were due to differences in sample type and origin, but also due to DNA extraction method, sample handling, and bioinformatical tools used (Sinha *et al.*, 2017). Thus, for performing standardized experiments, reference reagents were provided by the National Institute for Biological Standards and Control (NIBSC) with the support of the World Health Organization (WHO). The standards are intended for gut microbiome analyses which are performed using NGS-based techniques (Amos *et al.*, 2020). Further, both shallow metagenomic sequencing and 16S rRNA gene sequencing were used to assess and investigate thresholds that should be met when the distributed standards are used in benchmarking existing analysis pipelines (Amos *et al.*, 2020). In any

case, the most important aspects in standardization are summarized in Fig 3.1 and will be further discussed in detail.

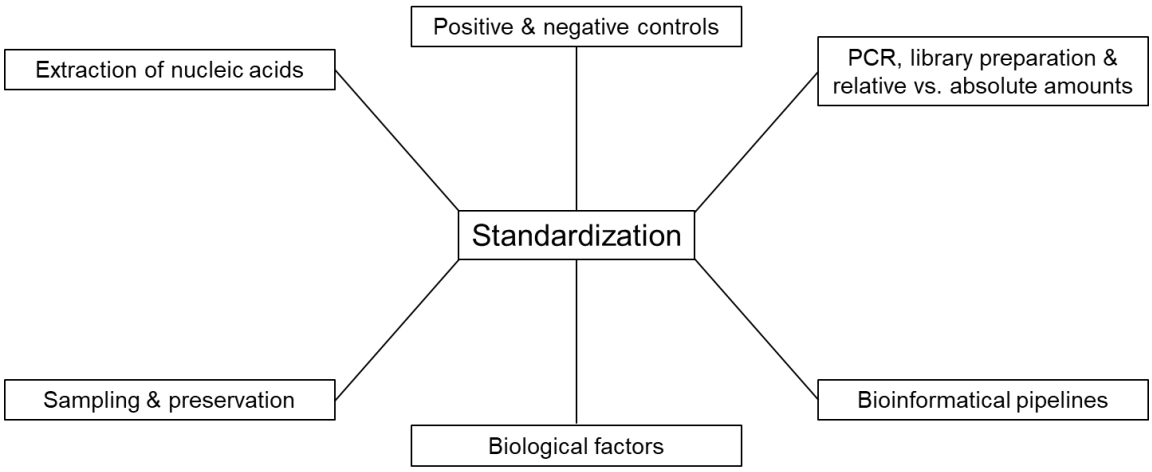

Figure 3.1: Factors impacting standardization in the field of microbiome research.

Sampling and Preservation

It was shown that different sampling and preservation approaches influence the resulting taxonomical profiles generated after sequencing (e.g., Bjerre *et al.*, 2019, Camacho-Sanchez *et al.*, 2013, Gorzelak *et al.*, 2015, Hsieh *et al.*, 2016, McDonald *et al.*, 2018, Sinha *et al.*, 2016, Wu *et al.*, 2010a). In this study, we confirmed that the use of different preservation solutions influences downstream analysis. I showed that the commercially available RL and the self-made SStab showed better stabilizing potentials than the commercial DNA stool stabilizer for stabilizing RNA. For the DNA extraction efficiency and quality, on the other hand, we do not see an explicit difference. When samples stored in either DS or SStab were extracted and sequenced using short amplicon 16S rRNA gene sequencing, differences in β-diversity were small and, thus, results were overall comparable. Importantly, if other stabilizing agents shall be used in future, it must be checked beforehand if taxonomical profiles shift, or if the integrity of the nucleic acids is impaired, or differences in the α- and β-diversity are detected.

Nonetheless, it should be noted that self-made stabilizing agents bear the potential of being of insufficient quality. It is more difficult to guarantee a reasonable and stringent quality for several batches produced and to strictly quality control raw materials and each step producing such liquids. Thus, we suggest producing self-made SStab in sufficiently large quantities needed for the complete experiment and enough back-up. Further, it might be reasonable to purchase a batch of ingredients from the same vendors in a high quality. Otherwise, commercial products are recommended with that hope that comparability is given due to the companies' quality control. Next to this, different storage temperatures could affect the stabilizing agent itself (decay of chemicals or precipitation). In any case, samples in the preservation buffer should be stored as cold as possible, since certain

bacterial populations have been shown to grow in such liquids (Nel Van Zyl *et al.*, 2020). Obviously, the integrity of the nucleic acids is preserved longer at colder temperatures. Besides, it should be evaluated if stabilizing agents could be omitted (i.e., collection close to the lab) as Menke *et al.* (2017) showed that sampling without any buffer and immediate (flash) freezing gave best results.

Extraction of nucleic acids

DNA extraction from samples stored in the different tested buffers resulted in similar quality and quantity of the extracted gDNA product (section 2.1). For the gDNA extraction, different extractions protocols were not evaluated as the used protocol was found to perform best in a previous study (Reitmeier, 2021). For the extraction of RNAs, we found that the Zymo Quick RNA isolation kit for stool samples performed better in terms of stability and efficiency compared to the phenol-chloroform-based approach adapted from Zoetendal *et al.* (2006).

Generally, it is very important to assess and regularly re-evaluate the used nucleic acid extraction protocols as previous reports showed that this aspect might be the most biasing factor in microbiota/microbiome research besides real biological variation (Bartolomaeus *et al.*, 2020, Costea *et al.*, 2017, Sinha *et al.*, 2017). Greathouse *et al.* (2019) mentioned that automation of the extraction process could improve comparability between laboratories and quality control samples such as mock communities as positive controls and buffer or water samples as negative controls should always be used to further identify and address possible issues. Finally, for low biomass samples the risk of contamination is more relevant during the extraction process as contaminating DNA could displace the initials sample DNA (Salter *et al.*, 2014).Thus special or updated protocols (e.g., Bjerre *et al.*, 2019, Douglas *et al.*, 2020, Siebert *et al.*, 2021) should be used for such samples. Villette *et al.* (2021) for expample showed that an extended and repeated mechanical lysing step improved the representation of bacterial composition from low biomass samples. Other approaches included the use of denoising approaches to remove spurious OTUs and potential contaminants (Claassen-Weitz *et al.*, 2020) or automation as described in the KatharoSeq protocol (Minich *et al.*, 2018).

Positive and negative controls

In this work, I showed that it is of great importance to use sufficiently complex mock communities. Biasing effects that could influence taxonomical profiles and downstream analyses were not observed when simple mock communities only were used (see section 2.2, 2.5, and 2.6). The importance of negative controls becomes most obvious when very low gDNA input or low biomass (i.e., low bacterial counts) samples are used (section 2.3).

It is crucial to assess whether profiles and results reported are reflecting the real situation or if contaminants influence the results such that they might even show similar or indistinguishable profiles compared to the negative controls. Here, bioinformatical tools that allow the removal of contaminant sequences, as suggested by Davis *et al.* (2018) or McKnight *et al.* (2019), might be useful, but have not been tested here.

Interestingly, Hornung *et al.* (2019) checked in a small survey of 265 publications how many of those have reported using either or both negative and positive controls. They found that only roughly 30% and 10% of the studies use negative and positive controls, respectively, which is alarming. It should be mandatory to have strict rules on how to perform and describe such experiments and include sufficiently complex mock communities. Reviewers in the scientific community must impose mandatory reports for the use of controls without which publication would not be possible (Hornung *et al.*, 2019, Kim *et al.*, 2017, Schierwagen *et al.*, 2020).

PCR and library preparation

Several aspects have to be considered for any library preparation. First, it must be decided whether full-length SSU rRNA gene sequencing or short amplicon sequencing shall be used. The advantage of full-length over short amplicon sequencing is an improved species-level classification on the cost of longer preparation times, a different primer bias (see section 2.5), and overall higher costs, mainly due to an increased library-preparation costs and sequencing depth required. Nonetheless, full-length sequencing, if pathogens or spoilage-associated microorganisms should be identified at the species level, is highly beneficial (see section 2.6). Moreover, it could be shown previously that full-length 16S rRNA gene sequencing might not only allow taxonomic resolution of the species but even for strain-level (Johnson *et al.*, 2019). For our MinION approach, we could not reach sufficiently valid and reliable results compared to the other methodologies tested. Thus, we believe that further investigation in this direction is needed and will be worthwhile. It should be tested whether applying or adapting the approaches presented by Karst *et al.* (2021) or Matsuo *et al.* (2021) could improve the taxonomical classification and, therefore, the reliability of full-length sequencing. Further, it should be tested whether the sequencing accuracy could be improved by adapting the $1D^2$ technology of ONT, which allows to sequence both strands of the targeted DNA one after the other (Santos *et al.*, 2020).

Secondly, the use of different primer pairs or even primer mixes is of great importance and must be considered carefully. For instance, the commercially available LoopSeq kit (LoopGenomics), which performed in our comparison best for the full-length approaches (see section 2.5), uses a primer-mixture allowing not only the assessment of bacteria present within a sample but also targets the 16S rRNA of archaea and the 18S rRNA of

eukaryotic microorganisms. This might be of interest for environmental studies, as a lot of the often used short amplicon primers are only optimized for bacteria, and therefore primer pairs must be tested beforehand to guarantee that archaea are targeted as well (Bahram *et al.*, 2018, Fischer *et al.*, 2016).

For short amplicon sequencing approaches, we showed that the selection of different primer pairs amplifying distinct V-regions is crucial (see section 2.2.). Several primer pairs were shown to miss certain genera, e.g., primer pair 27F and 338R (V1-V2) *Akkermansia* or 515F and 806R (V4) *Salmonella*. Several studies have already reported similar findings (e.g., Chen *et al.*, 2019, Klindworth *et al.*, 2012, Thijs *et al.*, 2017, Tremblay *et al.*, 2015) and are the reason why studies using different primer sets cannot be compared easily. In our test set, composed of different mock communities and human fecal samples, we found that, overall, primers used for amplifying V3-V4 performed best and are therefore recommended.

Concerning the specific case of low-biomass samples, library preparation procedures should be adapted as well. We showed here that by the implementation of a ddPCR step, sequencing and taxonomical profiles could be reliably produced and achieved even with gDNA input amounts <1 ng (see section 2.3). Several aspects of improving sequencing of low-biomass samples were performed previously (e.g., Claassen-Weitz *et al.*, 2020, Minich *et al.*, 2018, Saladié *et al.*, 2020), but no similar technical approach has yet been described. Thus, we think that this method will allow further investigation in fields that have to deal with low biomass issues and will facilitates gaining reliable 16S profiles.

Relative and absolute amounts

Measuring the relative amount of bacteria in a sample is informative but has drawbacks. The expansion of one group of bacteria evitably decreases the other group, even though their numbers might not have changed. Further, low bacterial counts might be informative concerning some illnesses, but when all samples are normalized to the same amount (i.e., 100%), this information is not available anymore.

Several methods measuring bacterial numbers in a sample have been proposed, of which I tested flow cytometric counting and spike-in sequencing. The latter approach was described by Tourlousse *et al.* (2017), and I used their spike-in sequences proposed in short amplicon-based synthetic spike-in sequencing. The resulting spike-in normalized bacterial abundances were compared to bacterial counts using flow cytometry for a test set consisting of IBD patients' fecal samples and control samples from healthy volunteers. For both the spike-in normalized abundances determined through spike-in sequencing and the cell counting by flow cytometry, a higher number of bacterial cells were determined for healthy controls than for the IBD patients (see section 2.4). As the spike-in sequencing

strategy was shown to not affect the normalized abundances and was highly correlated to sequencing results performed without spike-in, we propose to use spike-in strategies regularly or even as the standard short amplicon sequencing approach. This is also reinforced by recent publications, which state that experimental quantitative data is needed and should be collected habitually (Barlow *et al.*, 2020, Jian *et al.*, 2020, Lloréns-Rico *et al.*, 2021, Zemb *et al.*, 2020).

Bioinformatical pipelines

A great variety of different software tools for the analysis of microbiome data exist. Those different software options use distinct reference databases, clustering approaches, filtering options, and cut-offs. In our analysis (see section 2.2), we found that RDP and Silva were the most accurate reference databases. GreenGenes is still a frequently used reference database but was shown to perform poorly in our analysis. This is most likely because of missing updates, as the last one was performed in 2013 and the database is outdated. Concerning the clustering approach, no major deviation in the resulting taxonomical profiles were detected when either OTUs, zOTUs, or ASVs were produced. Nonetheless, approaches that include denoising steps (zOTU and ASVs) performed slightly better than standard OTUs and are therefore recommended.

Comparability of datasets that were analyzed using different software and databases is difficult (see section 2.2. and 2.5) as nomenclature might be different, and the estimation of relative abundances are affected. This was also shown in other publications (Marizzoni *et al.*, 2020, Prodan *et al.*, 2020, Straub *et al.*, 2020) and reinforces the need for standardization and, most importantly, proper documentation of pipelines, reference databases, and settings used.

Biological factors

Ideally, only biological factors (health status, age, gender, etc.) and environmental factors (location, diet, etc.) should cause different results in microbiome studies. The effect of biomass is an additional vital factor to consider. The overall small number of microorganisms in environments that are of low biomass challenges correct community composition. Moreover, sample contamination is more relevant for such samples and must therefore be controlled and assessed.

Besides the risk of samples being contaminated, e.g., through laboratory or kit contaminations, other technical and methodological bias exists. Due to the vast number of variables, both in biological and technical terms, it is still challenging to clearly describe factors influencing a healthy microbiome to become dysbiotic. McBurney et al. (2019) stated that validated biomarkers and precisely defined measures are needed to properly

study and investigate microbiome-host interactions and to contribute to defining a healthy human microbiome. Shanahan *et al.* (2021) stated that 85% of compositional variance detected in population-based studies can still not be explained. The authors further concluded that strain-level information is required on the one hand, and thus, technical aspects and strategies must be improved. On the other hand, intestinal microorganisms besides bacteria, such as viruses, fungi, and archaea, must be further studied in detail (Shanahan *et al.*, 2021).

Conclusion

In this work, I presented several aspects improving SSU rRNA gene sequencing strategies, from using low amounts of gDNA, choosing V-regions, over proper bioinformatics to validation using complex mock communities and sufficient controls. However, it is mandatory for ongoing and future research to have appropriate setups well thought through and a detailed reporting of the methods used, including those in bioinformatics. The most critical elements to consider for future studies have been assembled in Fig 3.2.

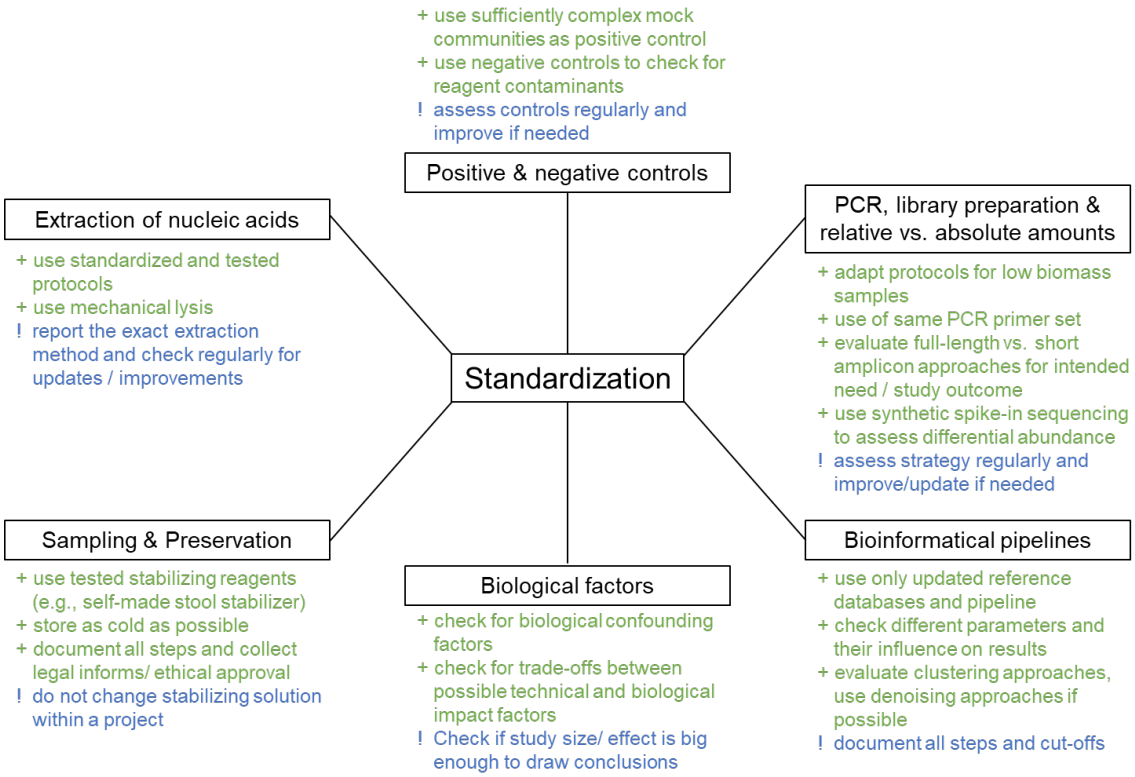

Figure 3.2: Important factors for standardization in SSU rRNA gene sequencing approaches. In green, the most important aspects to consider are listed. In blue, we highlighted what researchers have to do at least when performing microbiota/microbiome-based studies.

4. References

Abebe R, Hatiya H, Abera M, Megersa B, Asmare K. Bovine mastitis: prevalence, risk factors and isolation of Staphylococcus aureus in dairy herds at Hawassa milk shed, South Ethiopia. BMC Veterinary Research 2016, 12(1):270.

Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. mSphere 2021a, 6(1):e01202-01220.

Abellan-Schneyder I, Siebert A, Hofmann K, Wenning M, Neuhaus K. Full-Length SSU rRNA Gene Sequencing Allows Species-Level Detection of Bacteria, Archaea, and Yeasts Present in Milk. Microorganisms 2021b, 9(6):1251.

Addis MF, Tanca A, Uzzau S, Oikonomou G, Bicalho RC, Moroni P. The bovine milk microbiota: insights and perspectives from -omics studies. Molecular BioSystems 2016, 12(8):2359-2372.

Ahmadian A, Ehn M, Hober S. Pyrosequencing: History, biochemistry and future. Clinica Chimica Acta 2006, 363(1):83-94.

Ahn JH, Kim BY, Song J, Weon HY. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. J Microbiol 2012, 50(6):1071-1074.

Allaband C, McDonald D, Vázquez-Baeza Y, Minich JJ, Tripathi A, Brenner DA, Loomba R, Smarr L, Sandborn WJ, Schnabl B *et al.* Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. Clin Gastroenterol Hepatol 2019, 17(2):218-230.

Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience 2018, 7(5).

Ameur A, Kloosterman WP, Hestand MS. Single-Molecule Sequencing: Towards Clinical Applications. Trends in Biotechnology 2019, 37(1):72-85.

Amos GCA, Logan A, Anwar S, Fritzsche M, Mate R, Bleazard T, Rijpkema S. Developing standards for the microbiome field. Microbiome 2020, 8(1):98.

Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010.

Ansorge WJ, Katsila T, Patrinos GP: Chapter 8 - Perspectives for Future DNA Sequencing Techniques and Applications. In: Molecular Diagnostics (Third Edition). Edited by Patrinos GP: Academic Press; 2017: 141-153.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic acids research 2018, 46(5):2159-2168.

Atreya R, Neurath MF, Siegmund B. Personalizing Treatment in IBD: Hype or Reality in 2020? Can We Predict Response to Anti-TNF? Frontiers in Medicine 2020, 7(517).

Aziz RK. A hundred-year-old insight into the gut microbiome! Gut pathogens 2009, 1(1):21-21.

Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. Environmental microbiology reports 2018.

Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 2003, 55(3):541-555.

Barlow JT, Bogatyrev SR, Ismagilov RF. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. Nature communications 2020, 11(1):2590.

Bartolomaeus TUP, Birkner T, Bartolomaeus H, Löber U, Avery EG, Mähler A, Weber D, Kochlik B, Balogh A, Wilck N *et al.* Quantifying technical confounders in microbiome studies. Cardiovascular Research 2020, 117(3):863-875.

Bashiardes S, Zilberman-Schapira G, Elinav E. Use of Metatranscriptomics in Microbiome Research. Bioinform Biol Insights 2016, 10:19-25.

Baunwall SMD, Lee MM, Eriksen MK, Mullish BH, Marchesi JR, Dahlerup JF, Hvas CL. Faecal microbiota transplantation for recurrent Clostridioides difficile infection: An updated systematic review and meta-analysis. EClinicalMedicine 2020, 29.

Bellali S, Lagier JC, Raoult D, Bou Khalil J. Among Live and Dead Bacteria, the Optimization of Sample Collection and Processing Remains Essential in Recovering Gut Microbiota Components. Frontiers in microbiology 2019, 10:1606.

Bender JM, Li F, Adisetiyo H, Lee D, Zabih S, Hung L, Wilkinson TA, Pannaraj PS, She RC, Bard JD *et al.* Quantification of variation and the impact of biomass in targeted 16S rRNA gene sequencing studies. Microbiome 2018, 6(1):155.

Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience 2016, 5(1):4.

Benitez-Paez A, Sanz Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION portable nanopore sequencer. Gigascience 2017, 6(7):1-12.

Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, Chen X, Cocolin L, Eversole K, Corral GH *et al.* Microbiome definition re-visited: old concepts and new challenges. Microbiome 2020, 8(1):103.

Berg JM, Tymoczko JL, Stryer L: Stryer Biochemie, 7. Auflage edn. Berlin, Heidelberg: Springer Spektrum; 2013.

Bernstein CN. Treatment of IBD: Where We Are and Where We Are Going. Official journal of the American College of Gastroenterology | ACG 2015, 110(1).

Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Applied and environmental microbiology 2011, 77(21):7846-7849.

Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Briefings in Bioinformatics 2019, 22(1):178-193.

Bjerre RD, Hugerth LW, Boulund F, Seifert M, Johansen JD, Engstrand L. Effects of sampling strategy and DNA extraction on human skin microbiome investigations. Scientific reports 2019, 9(1):17287.

Blackburn J, Wong T, Madala BS, Barker C, Hardwick SA, Reis ALM, Deveson IW, Mercer TR. Use of synthetic DNA spike-in controls (sequins) for human genome sequencing. Nature protocols 2019, 14(7):2119-2151.

Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. The ISME journal 2013, 7(11):2061-2068.

Blevins SM, Bronze MS. Robert Koch and the 'golden age' of bacteriology. International Journal of Infectious Diseases 2010, 14(9):e744-e751.

Bolte J, Zhang Y, Wente N, Krömker V. In Vitro Susceptibility of Mastitis Pathogens Isolated from Clinical Mastitis Cases on Northern German Dairy Farms. Vet Sci 2020, 7(1).

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology 2019, 37(8):852-857.

Brandt J, Albertsen M. Investigation of Detection Limits and the Influence of DNA Extraction and Primer Choice on the Observed Microbial Communities in Drinking Water Samples Using 16S rRNA Gene Amplicon Sequencing. Frontiers in microbiology 2018, 9:2140.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X *et al.* The potential and challenges of nanopore sequencing. Nature biotechnology 2008, 26(10):1146-1153.

Breitenwieser F, Doll EV, Clavel T, Scherer S, Wenning M. Complementary Use of Cultivation and High-Throughput Amplicon Sequencing Reveals High Biodiversity Within Raw Milk Microbiota. Frontiers in microbiology 2020, 11(1557).

Brennecke J, Falkenberg U, Wente N, Krömker V. Are Severe Mastitis Cases in Dairy Cows Associated with Bacteremia? Animals 2021, 11(2):410.

Brown MR, Hands CL, Coello-Garcia T, Sani BS, Ott AIG, Smith SJ, Davenport RJ. A flow cytometry method for bacterial quantification and biomass estimates in activated sludge. J Microbiol Methods 2019, 160:73-83.

Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Scientific data 2019, 6(1):190007.

Burke CM, Darling AE. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. PeerJ 2016, 4:e2492.

Callahan BJ, Grinevich D, Thakur S, Balamotis MA, Yehezkel TB. Ultra-accurate microbial amplicon sequencing with synthetic long reads. Microbiome 2021, 9(1):130.

Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal 2017, 11(12):2639-2643.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods 2016, 13(7):581-583.

Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic acids research 2019, 47(18):e103.

Calus ST, Ijaz UZ, Pinto AJ. NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. GigaScience 2018, 7(12).

Camacho-Sanchez M, Burraco P, Gomez-Mestre I, Leonard JA. Preservation of RNA and DNA from mammal samples under field conditions. Molecular ecology resources 2013, 13(4):663-673.

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

Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N *et al.* Moving pictures of the human microbiome. Genome biology 2011, 12(5):R50.

Cardona S, Eck A, Cassellas M, Gallart M, Alastrue C, Dore J, Azpiroz F, Roca J, Guarner F, Manichanh C. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC microbiology 2012, 12:158.

Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods 2007, 69(2):330-339.

Chen C-C, Wu W-K, Chang C-M, Panyod S, Lu T-P, Liou J-M, Fang Y-J, Chuang EY, Wu M-S. Comparison of DNA stabilizers and storage conditions on preserving fecal microbiota profiles. Journal of the Formosan Medical Association 2020, 119(12):1791-1798.

Chen Z, Hui PC, Hui M, Yeoh YK, Wong PY, Chan MCW, Wong MCS, Ng SC, Chan FKL, Chan PKS. Impact of Preservation Method and 16S rRNA Hypervariable Region on Gut Microbiota Profiling. mSystems 2019, 4(1):e00271-00218.

Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nature reviews Genetics 2012, 13(4):260-270.

Chongming W, Tianda C, Wenyi X, Tingting Z, Yuwei P, Yanan Y, Fang Z, Hao G, Qingshi W, Li W *et al.* The Maintenance of Microbial Community in Human Fecal Samples by a Self-made Cost-effective Preservation Buffer. Scientific reports 2021.

Choo JM, Leong LE, Rogers GB. Sample storage conditions significantly influence faecal microbiome profiles. Scientific reports 2015, 5:16350.

Chung N, Van Goethem MW, Preston MA, Lhota F, Cerna L, Garcia-Pichel F, Fernandes V, Giraldo-Silva A, Kim HS, Hurowitz E *et al.* Accurate Microbiome Sequencing with Synthetic Long Read Sequencing. bioRxiv 2020:2020.2010.2002.324038.

Ciuffreda L, Rodríguez-Pérez H, Flores C. Nanopore sequencing and its application to the study of microbial communities. Computational and Structural Biotechnology Journal 2021, 19:1497-1511.

Claassen-Weitz S, Gardner-Lubbe S, Mwaikono KS, du Toit E, Zar HJ, Nicol MP. Optimizing 16S rRNA gene profile analysis from low biomass nasopharyngeal and induced sputum specimens. BMC microbiology 2020, 20(1):113.

Claassen S, du Toit E, Kaba M, Moodley C, Zar HJ, Nicol MP. A comparison of the efficiency of five different commercial DNA extraction kits for extraction of DNA from faecal samples. Journal of microbiological methods 2013, 94(2):103-110.

Clarridge JE, 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 2004, 17(4):840-862.

Clavel T, Reitmeier S, Hitch TCA, Treichel N, Fikas N, Hausmann B, Ramer-Tait A, Neuhaus K, Berry D, Haller D *et al.* Handling of Spurious Sequences Affects the Outcome of High-Throughput 16S rRNA Gene Amplicon Profiling. Research Square 2020.

Clooney AG, Fouhy F, Sleator RD, A OD, Stanton C, Cotter PD, Claesson MJ. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. PloS one 2016, 11(2):e0148028.

Cobirka M, Tancin V, Slama P. Epidemiology and Classification of Mastitis. Animals 2020, 10(12):2212.

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-D642.

Coleman OI, Haller D. Bacterial Signaling at the Intestinal Epithelial Interface in Inflammation and Cancer. Frontiers in Immunology 2018, 8(1927).

Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung F-E *et al.* Towards standards for human fecal sample processing in metagenomic studies. Nature biotechnology 2017, 35(11):1069-1076.

Couto N, Schuele L, Raangs EC, Machado MP, Mendes CI, Jesus TF, Chlebowicz M, Rosema S, Ramirez M, Carriço JA *et al.* Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Scientific reports 2018, 8(1):13767.

Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon [version 2; peer review: 2 approved, 3 approved with reservations]. F1000Research 2019, 7(1755).

D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC genomics 2016, 17:55.

Dahlberg J, Sun L, Persson Waller K, Östensson K, McGuire M, Agenäs S, Dicksved J. Microbiota data from low biomass milk samples is markedly affected by laboratory and reagent contamination. PloS one 2019, 14(6):e0218257-e0218257.

Dahlberg J, Williams JE, McGuire MA, Peterson HK, Östensson K, Agenäs S, Dicksved J, Waller KP. Microbiota of bovine milk, teat skin, and teat canal: Similarity and variation due to sampling technique and milk fraction. Journal of Dairy Science 2020, 103(8):7322-7330.

Dalanezi FM, Joaquim SF, Guimarães FF, Guerra ST, Lopes BC, Schmidt EMS, Cerri RLA, Langoni H. Influence of pathogens causing clinical mastitis on reproductive variables of dairy cows. J Dairy Sci 2020, 103(4):3648-3655.

Davies J. Where have All the Antibiotics Gone? Can J Infect Dis Med Microbiol 2006, 17(5):287-290.

Davis A, Kohler C, Alsallaq R, Hayden R, Maron G, Margolis E. Improved yield and accuracy for DNA extraction in microbiome studies with variation in microbial biomass. BioTechniques 2019, 66(6):285-289.

Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018, 6(1):226.

De Almeida CV, de Camargo MR, Russo E, Amedei A. Role of diet and gut microbiota on colorectal cancer immunomodulation. World J Gastroenterol 2019, 25(2):151-162.

De Filippis F, Parente E, Zotta T, Ercolini D. A comparison of bioinformatic approaches for 16S rRNA gene profiling of food bacterial microbiota. International Journal of Food Microbiology 2018, 265:9-17.

Demeke T, Dobnik D. Critical assessment of digital PCR for the detection and quantification of genetically modified organisms. Analytical and bioanalytical chemistry 2018, 410(17):4039-4050.

Derakhshani H, Fehr KB, Sepehri S, Francoz D, De Buck J, Barkema HW, Plaizier JC, Khafipour E. Microbiota of the bovine udder: Contributing factors and potential implications for udder health and mastitis susceptibility. Journal of Dairy Science 2018, 101(12):10605-10625.

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.

Deutscher AT, Burke CM, Darling AE, Riegler M, Reynolds OL, Chapman TA. Near full-length 16S rRNA gene next-generation sequencing revealed Asaia as a common midgut bacterium of wild and domesticated Queensland fruit fly larvae. Microbiome 2018, 6(1):85.

Di L, Fu Y, Sun Y, Li J, Liu L, Yao J, Wang G, Wu Y, Lao K, Lee RW *et al.* RNA sequencing by direct tagmentation of RNA/DNA hybrids. Proceedings of the National Academy of Sciences 2020, 117(6):2886-2893.

Dietrich A, Matchado MS, Zwiebel M, Ölke B, Lauber M, Baumbach J, Haller D, Reitmeier S, List M. Namco: A microbiome explorer unpublished 2021.

Dominianni C, Wu J, Hayes RB, Ahn J. Comparison of methods for fecal microbiome biospecimen collection. BMC microbiology 2014, 14:103.

Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. Nature Reviews Microbiology 2016, 14(1):20-32.

Douglas CA, Ivey KL, Papanicolas LE, Best KP, Muhlhausler BS, Rogers GB. DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. Scientific reports 2020, 10(1):123.

Doyle CJ, Gleeson D, O'Toole PW, Cotter PD. Impacts of Seasonal Housing and Teat Preparation on Raw Milk Microbiota: a High-Throughput Sequencing Study. Applied and environmental microbiology 2017, 83(2):e02694-02616.

Drengenes C, Eagan TML, Haaland I, Wiker HG, Nielsen R. Exploring protocol bias in airway microbiome studies: one versus two PCR steps and 16S rRNA gene region V3 V4 versus V4. BMC genomics 2021, 22(1):3.

Dreo T, Pirc M, Ramsak Z, Pavsic J, Milavec M, Zel J, Gruden K. Optimising droplet digital PCR analysis approaches for detection and quantification of bacteria: a case study of fire blight and potato brown rot. Analytical and bioanalytical chemistry 2014, 406(26):6513-6528.

Du B, Meng L, Liu H, Zheng N, Zhang Y, Guo X, Zhao S, Li F, Wang J. Impacts of Milking and Housing Environment on Milk Microbiota. Animals 2020, 10(12):2339.

Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, Albertsen M, Nielsen PH. Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). mBio 2020, 11(5):e01557-01520.

Dufour S, Labrie J, Jacques M. The Mastitis Pathogens Culture Collection. Microbiol Resour Announc 2019, 8(15):e00133-00119.

Dulanto Chiang A, Dekker JP. From the Pipeline to the Bedside: Advances and Challenges in Clinical Metagenomics. The Journal of Infectious Diseases 2019, 221(Supplement_3):S331-S340.

Duquenoy A, Bellais S, Gasc C, Schwintner C, Dore J, Thomas V. Assessment of Gram- and Viability-Staining Methods for Quantifying Bacterial Community Dynamics Using Flow Cytometry. Frontiers in microbiology 2020, 11(1469).

Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, Palmer JN, Workman AD, Blasetti M, Sen B *et al.* Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. Microbiome 2018, 6(1):190.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. Diversity of the Human Intestinal Microbial Flora. Science (New York, NY) 2005, 308(5728):1635-1638.

Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv 2016:081257.

Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 2013, 10(10):996-998.

Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 2018, 34(14):2371-2375.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011, 27(16):2194-2200.

Editorials NM. Raising standards in microbiome research. Nature microbiology 2016, 1(7):16112.

Egerton FN. A History of the Ecological Sciences, Part 19: Leeuwenhoek's Microscopic Natural History. The Bulletin of the Ecological Society of America 2006, 87(1):47-58.

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 (New York, NY) 2009, 323(5910):133-138.

Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends in Microbiology 2019, 27(2):105-117.

Emerson JB, Adams RI, Román CMB, Brooks B, Coil DA, Dahlhausen K, Ganz HH, Hartmann EM, Hsu T, Justice NB *et al.* Schrödinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. Microbiome 2017, 5(1):86.

Escapa IF, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, Lemon KP. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 2020, 8(1):65.

Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL *et al.* The long-term stability of the human gut microbiota. Science (New York, NY) 2013, 341(6141):1237439-1237439.

Farré-Maduell E, Casals-Pascual C. The origins of gut microbiome research in Europe: From Escherich to Nissle. Human Microbiome Journal 2019, 14:100065.

Fidler G, Tolnai E, Stagel A, Remenyik J, Stundl L, Gal F, Biro S, Paholcsek M. Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling. Scientific reports 2020, 10(1):3419.

Fiedorová K, Radvanský M, Němcová E, Grombiříková H, Bosák J, Černochová M, Lexa M, Šmajs D, Freiberger T. The Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community Recovery. Frontiers in microbiology 2019, 10(821).

Fischer MA, Güllert S, Neulinger SC, Streit WR, Schmitz RA. Evaluation of 16S rRNA Gene Primer Pairs for Monitoring Microbial Community Structures Showed High Reproducibility within and Low Comparability between Datasets Generated with Multiple Archaeal and Bacterial Primer Pairs. Frontiers in microbiology 2016, 7:1297.

Flores R, Shi J, Yu G, Ma B, Ravel J, Goedert JJ, Sinha R. Collection media and delayed freezing effects on microbial composition of human stool. Microbiome 2015, 3:33.

Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC microbiology 2016, 16(1):123.

Fouhy F, Deane J, Rea MC, O'Sullivan O, Ross RP, O'Callaghan G, Plant BJ, Stanton C. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. PloS one 2015a, 10(3):e0119355.

Fouhy F, Deane J, Rea MC, O'Sullivan Ó, Ross RP, O'Callaghan G, Plant BJ, Stanton C. The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. PloS one 2015b, 10(3):e0119355.

Fox GE. Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol 2010, 2(9):a003483-a003483.

Francoz D, Wellemans V, Dupré JP, Roy JP, Labelle F, Lacasse P, Dufour S. Invited review: A systematic review and qualitative analysis of treatments other than conventional antimicrobials for clinical mastitis in dairy cows. Journal of Dairy Science 2017, 100(10):7751-7770.

Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D *et al.* Relating the metatranscriptome and metagenome of the human gut. Proceedings of the National Academy of Sciences of the United States of America 2014, 111(22):E2329-E2338.

Fremin BJ, Sberro H, Bhatt AS. MetaRibo-Seq measures translation in microbiomes. Nature communications 2020, 11(1):3268.

Fricker AM, Podlesny D, Fricke WF. What is new and relevant for sequencing-based microbiome research? A mini-review. Journal of Advanced Research 2019, 19:105-112.

Friedmann H. Escherich and Escherichia. EcoSal Plus 2014.

Frossard A, Hammes F, Gessner MO. Flow Cytometric Assessment of Bacterial Abundance in Soils, Sediments and Sludge. Frontiers in microbiology 2016, 7:903.

Gest H. The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, Fellows of The Royal Society. Notes and Records of the Royal Society of London 2004, 58(2):187-201.

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-naive microbiome in new-onset Crohn's disease. Cell host & microbe 2014, 15(3):382-392.

Ghurye JS, Cepeda-Espinoza V, Pop M. Metagenomic Assembly: Overview, Challenges and Applications. Yale J Biol Med 2016, 89(3):353-362.

Giehren F: Deciphering metatranslatomes - the chain link to understand functionality of gut microbiota. 2021.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A *et al.* Genome sequencing identifies major causes of severe intellectual disability. Nature 2014, 511(7509):344-347.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science (New York, NY) 2006, 312(5778):1355-1359.

Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathogens 2016, 8(1):24.

Gobert G, Cotillard A, Fourmestraux C, Pruvost L, Miguet J, Boyer M. Droplet digital PCR improves absolute quantification of viable lactic acid bacteria in faecal samples. J Microbiol Methods 2018, 148:64-73.

Godon JJ, Zumstein E, Dabert P, Habouzit F, Moletta R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. Applied and environmental microbiology 1997, 63(7):2802-2813.

Gołębiewski M, Tretyn A. Generating amplicon reads for microbial community assessment with next-generation sequencing. Journal of applied microbiology 2020, 128(2):330-354.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 2016, 17(6):333-351.

Gordon S. Elie Metchnikoff, the Man and the Myth. Journal of Innate Immunity 2016, 8(3):223-227.

Gorokhova E. Effects of preservation and storage of microcrustaceans in RNAlater on RNA and DNA degradation. Limnology and Oceanography: Methods 2005, 3(2):143-148.

Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. PloS one 2015, 10(8):e0134802.

Greathouse KL, Sinha R, Vogtmann E. DNA extraction for human microbiome studies: the issue of standardization. Genome biology 2019, 20(1):212.

Gremion F, Chatzinotas A, Harms H. Comparative 16S rDNA and 16S rRNA sequence analysis indicates that Actinobacteria might be a dominant part of the metabolically active bacteria in heavy metal-contaminated bulk and rhizosphere soil. Environmental Microbiology 2003, 5(10):896-907.

Gryp T, Glorieux G, Joossens M, Vaneechoutte M. Comparison of five assays for DNA extraction from bacterial cells in human faecal samples. Journal of applied microbiology 2020, 129(2):378-388.

IHMS_SOP01_V1: Standard Operating Procedure for Sample Identification

Guo Y, Li SH, Kuang YS, He JR, Lu JH, Luo BJ, Jiang FJ, Liu YZ, Papasian CJ, Xia HM *et al.* Effect of short-term room temperature storage on the microbial community in infant fecal samples. Scientific reports 2016, 6:26648.

Hacker J, Blum-Oehler G. In appreciation of Theodor Escherich. Nature Reviews Microbiology 2007, 5(12):902-902.

Halasa T, Kirkeby C. Differential Somatic Cell Count: Value for Udder Health Management. Frontiers in Veterinary Science 2020, 7(1153).

Han M, Hao L, Lin Y, Li F, Wang J, Yang H, Xiao L, Kristiansen K, Jia H, Li J. A novel affordable reagent for room temperature storage and transport of fecal samples for metagenomic analyses. Microbiome 2018, 6(1):43-43.

Hardwick SA, Chen WY, Wong T, Kanakamedala BS, Deveson IW, Ongley SE, Santini NS, Marcellin E, Smith MA, Nielsen LK *et al.* Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nature communications 2018, 9(1):3096.

Harrison JG, John Calder W, Shuman B, Alex Buerkle C. The quest for absolute abundance: The use of internal standards for DNA-based community ecology. Molecular ecology resources 2021, 21(1):30-43.

Hart ML, Meyer A, Johnson PJ, Ericsson AC. Comparative Evaluation of DNA Extraction Methods from Feces of Multiple Host Species for Downstream Next-Generation Sequencing. PloS one 2015, 10(11):e0143334.

Hatch A, Horne J, Toma R, Twibell BL, Somerville KM, Pelle B, Canfield KP, Genkin M, Banavar G, Perlina A *et al.* A Robust Metatranscriptomic Technology for Population-Scale Studies of Diet, Gut Microbiome, and Human Health. International Journal of Genomics 2019, 2019:1718741.

Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics 2016, 107(1):1-8.

Heikema AP, Horst-Kreft D, Boers SA, Jansen R, Hiltemann SD, de Koning W, Kraaij R, de Ridder MAJ, van Houten CB, Bont LJ *et al.* Comparison of Illumina versus Nanopore 16S rRNA Gene Sequencing of the Human Nasal Microbiota. Genes (Basel) 2020, 11(9).

Hickl O, Heintz-Buschart A, Trautwein-Schult A, Hercog R, Bork P, Wilmes P, Becher D. Sample Preservation and Storage Significantly Impact Taxonomic and Functional Profiles in Metaproteomics Studies of the Human Gut Microbiome. Microorganisms 2019, 7(9):367.

Hiitiö H, Simojoki H, Kalmus P, Holopainen J, Pyörälä S, Taponen S. The effect of sampling technique on PCR-based bacteriological results of bovine milk samples. Journal of Dairy Science 2016, 99(8):6532-6541.

Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems 2018, 3(6):e00069-00018.

Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 2011, 83(22):8604-8610.

Hong LZ, Hong S, Wong HT, Aw PPK, Cheng Y, Wilm A, de Sessions PF, Lim SG, Nagarajan N, Hibberd ML *et al.* BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. Genome biology 2014, 15(11):517.

Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. FEMS microbiology ecology 2019, 95(5):fiz045.

Hsieh Y-H, Peterson CM, Raggio A, Keenan MJ, Martin RJ, Ravussin E, Marco ML. Impact of Different Fecal Processing Methods on Assessments of Bacterial Diversity in the Human Intestine. Frontiers in microbiology 2016, 7:1643-1643.

Hücker SM, Ardern Z, Goldberg T, Schafferhans A, Bernhofer M, Vestergaard G, Nelson CW, Schloter M, Rost B, Scherer S *et al.* Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome. PloS one 2017, 12(9):e0184119.

Hungate RE. The anaerobic mesophilic cellulolytic bacteria. Bacteriol Rev 1950, 14(1):1-49.

Hungate RE: Chapter IV A Roll Tube Method for Cultivation of Strict Anaerobes. In: Methods in Microbiology. Edited by Norris JR, Ribbons DW, vol. 3: Academic Press; 1969: 117-132.

Hungate RE. Studies on Cellulose Fermentation: I. The Culture and Physiology of an Anaerobic Cellulose-digesting Bacterium. J Bacteriol 1944, 48(5):499-513.

Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology 2016, 17(1):239-239.

Janda JM, Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. Journal of Clinical Microbiology 2007, 45(9):2761-2764.

Jeong J, Yun K, Mun S, Chung W-H, Choi S-Y, Nam Y-d, Lim MY, Hong CP, Park C, Ahn Y *et al.* The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. Scientific reports 2021, 11(1):1727.

Jian C, Luukkonen P, Yki-Järvinen H, Salonen A, Korpela K. Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. PloS one 2020, 15(1):e0227285.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome research 2011, 21(9):1543-1551.

Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature communications 2019, 10(1):5029.

Jumpstart Consortium Human Microbiome Project Data Generation Working G. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. PloS one 2012, 7(6):e39315.

Kable ME, Srisengfa Y, Laird M, Zaragoza J, McLeod J, Heidenreich J, Marco ML. The Core and Seasonal Microbiota of Raw Bovine Milk in Tanker Trucks and the Impact of Transfer to a Milk Processing Facility. mBio 2016, 7(4):e00836-00816.

Kamke J, Taylor MW, Schmitt S. Activity profiles for marine sponge-associated bacteria obtained by 16S rRNA vs 16S rRNA gene comparisons. The ISME journal 2010, 4(4):498-508.

Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nature biotechnology 2018, 36(2):190-195.

Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nature methods 2021, 18(2):165-169.

Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, Braun J, McWeeney S. Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. mSystems 2019, 4(4):e00290-00219.

Kendall AI. Some Observations on the Study of the Intestinal Bacteria. Journal of Biological Chemistry 1909, 6(6):499-507.

Kennang Ouamba AJ, LaPointe G, Dufour S, Roy D. Optimization of Preservation Methods Allows Deeper Insights into Changes of Raw Milk Microbiota. Microorganisms 2020, 8(3):368.

Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. Nat Rev Gastroenterol Hepatol 2016, 13(9):508-516.

Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J *et al.* Optimizing methods and dodging pitfalls in microbiome research. Microbiome 2017, 5(1):52.

Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. J Microbiol Methods 2011, 84(1):81-87.

Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic acids research 2001, 29(1):181-184.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic acids research 2012, 41(1):e1-e1.

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D *et al.* Best practices for analysing microbiomes. Nature Reviews Microbiology 2018, 16(7):410-422.

Koliada A, Syzenko G, Moseiko V, Budovska L, Puchkov K, Perederiy V, Gavalko Y, Dorofeyev A, Romanenko M, Tkach S *et al.* Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC microbiology 2017, 17(1):120.

Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. Development, Growth & Differentiation 2019, 61(5):316-326.

Kuehn JS, Gorden PJ, Munro D, Rong R, Dong Q, Plummer PJ, Wang C, Phillips GJ. Bacterial community profiling of milk samples as a means to understand culture-negative bovine clinical mastitis. PloS one 2013, 8(4):e61959-e61959.

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

Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, Clavel T. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific reports 2016, 6:33721.

Lagkouvardos I, Klaring K, Heinzmann SS, Platz S, Scholz B, Engel KH, Schmitt-Kopplin P, Haller D, Rohn S, Skurk T *et al.* Gut metabolites and bacterial community networks during a pilot intervention study with flaxseeds in healthy adult men. Molecular nutrition & food research 2015, 59(8):1614-1628.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences 1985, 82(20):6955.

Lane N. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. Philos Trans R Soc Lond B Biol Sci 2015, 370(1666):20140344.

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

Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 2018, 35(3):421-432.

Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS microbiology letters 2010, 307(1):80-86.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and Quantification 2015, 3:1-8.

Leahy SC, Higgins DG, Fitzgerald GF, van Sinderen D. Getting better with bifidobacteria. Journal of applied microbiology 2005, 98(6):1303-1315.

Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, Martin S, Peel N, Acford-Palmer H, Hoyles L *et al.* Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. Nature microbiology 2019.

Leidy J: A flora and fauna within living animals. Washington: Smithsonian Institution; 1853.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. Nature 2006, 444(7122):1022-1023.

Li F, Hitch TCA, Chen Y, Creevey CJ, Guan LL. Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle. Microbiome 2019, 7(1):6.

Li N, Wang Y, You C, Ren J, Chen W, Zheng H, Liu Z. Variation in Raw Milk Microbiota Throughout 12 Months and the Impact of Weather Conditions. Scientific reports 2018, 8(1):2371.

Li R, Tun HM, Jahan M, Zhang Z, Kumar A, Dilantha Fernando WG, Farenhorst A, Khafipour E. Comparison of DNA-, PMA-, and RNA-based 16S rRNA Illumina sequencing for detection of live bacteria in water. Scientific reports 2017, 7(1):5752-5752.

Lim MY, Hong S, Kim B-M, Ahn Y, Kim H-J, Nam Y-D. Changes in microbiome and metabolomic profiles of fecal samples stored with stabilizing solution at room temperature: a pilot study. Scientific reports 2020, 10(1):1789.

Lim MY, Song E-J, Kim SH, Lee J, Nam Y-D. Comparison of DNA extraction methods for human gut microbial community profiling. Systematic and Applied Microbiology 2018, 41(2):151-157.

Lima SF, Bicalho MLdS, Bicalho RC. Evaluation of milk sample fractions for characterization of milk microbiota from healthy and clinical mastitis cows. PloS one 2018, 13(3):e0193671.

Liu J, Meng Z, Liu X, Zhang X-H. Microbial assembly, interaction, functioning, activity and diversification: a review derived from community compositional data. Marine Life Science & Technology 2019a, 1(1):112-128.

Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nature communications 2019b, 10(1):2449.

Liu Y-X, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein & Cell 2021, 12(5):315-330.

Lloréns-Rico V, Vieira-Silva S, Gonçalves PJ, Falony G, Raes J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. Nature communications 2021, 12(1):3562.

Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome medicine 2016, 8(1):51.

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 2017, 550(7674):61-66.

Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences 2016, 113(21):5970-5975.

Londei P: Archaeal Ribosomes. In: eLS John Wiley & Sons, Ltd (Ed). 2010.

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.

Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS microbiology reviews 1994, 15(2-3):155-173.

Ma J, Sheng L, Hong Y, Xi C, Gu Y, Zheng N, Li M, Chen L, Wu G, Li Y *et al.* Variations of Gut Microbiome Profile Under Different Storage Conditions and Preservation Periods: A Multi-Dimensional Evaluation. Frontiers in microbiology 2020, 11(972).

Madigan MT, Martinko JM, Bender KS, Buckley DH, Stahl DA: Brock Biology of Microorganisms, Global Edition : UEL. United Kingdom, UNITED KINGDOM: Pearson Education Limited; 2014.

Mahnic A, Breskvar M, Dzeroski S, Skok P, Pintar S, Rupnik M. Distinct Types of Gut Microbiota Dysbiosis in Hospitalized Gastroenterological Patients Are Disease Non-related and Characterized With the Predominance of Either Enterobacteriaceae or Enterococcus. Frontiers in microbiology 2020, 11(120).

Mallott EK, Malhi RS, Amato KR. Assessing the comparability of different DNA extraction and amplification methods in gut microbial community profiling. Access Microbiol 2019, 1(7):e000060.

Manzari C, Oranger A, Fosso B, Piancone E, Pesole G, D'Erchia AM. Accurate quantification of bacterial abundance in metagenomic DNAs accounting for variable DNA integrity levels. Microbial Genomics 2020, 6(10).

Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome 2015, 3(1):31.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437(7057):376-380.

Marino SM. A computational strategy for rapid on-site 16S metabarcoding with Oxford Nanopore sequencing. bioRxiv 2020:2020.2008.2025.267591.

Marizzoni M, Gurry T, Provasi S, Greub G, Lopizzo N, Ribaldi F, Festari C, Mazzelli M, Mombelli E, Salvatore M *et al.* Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples. Frontiers in microbiology 2020, 11(1262).

Markusková B, Minarovičová J, Véghová A, Drahovská H, Kaclíková E. Impact of DNA extraction methods on 16S rRNA-based profiling of bacterial communities in cheese. Journal of Microbiological Methods 2021, 184:106210.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 2011, 17(1):3.

Martínez JL. Effect of antibiotics on bacterial populations: a multi-hierachical selection process. F1000Research 2017, 6:51-51.

Martins SAM, Martins VC, Cardoso FA, Germano J, Rodrigues M, Duarte C, Bexiga R, Cardoso S, Freitas PP. Biosensors for On-Farm Diagnosis of Mastitis. Front Bioeng Biotechnol 2019, 7:186-186.

Martinson JNV, Walk ST. Escherichia coli Residency in the Gut of Healthy Human Adults. EcoSal Plus 2020, 9(1):10.1128/ecosalplus.ESP-0003-2020.

Matsuo Y, Komiya S, Yasumizu Y, Yasuoka Y, Mizushima K, Takagi T, Kryukov K, Fukuda A, Morimoto Y, Naito Y *et al.* Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. BMC microbiology 2021, 21(1):35.

McBurney MI, Davis C, Fraser CM, Schneeman BO, Huttenhower C, Verbeke K, Walter J, Latulippe ME. Establishing What Constitutes a Healthy Human Gut Microbiome: State of the Science, Regulatory Considerations, and Future Directions. The Journal of Nutrition 2019, 149(11):1882-1895.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier A-S. Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PloS one 2014, 9(9):e106689.

McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems 2018, 3(3):e00031-00018.

McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. Environmental DNA 2019, 1(1):14-25.

McOrist AL, Jackson M, Bird AR. A comparison of five methods for extraction of bacterial DNA from human faecal samples. Journal of Microbiological Methods 2002, 50(2):131-139.

Menke S, Gillingham MA, Wilhelm K, Sommer S. Home-Made Cost Effective Preservation Buffer Is a Better Alternative to Commercial Preservation Methods for Microbiome Research. Frontiers in microbiology 2017, 8:102.

Meola M, Rifa E, Shani N, Delbès C, Berthoud H, Chassard C. DAIRYdb: a manually curated reference database for improved taxonomy annotation of 16S rRNA gene sequences from dairy products. BMC genomics 2019, 20(1):560.

Metwaly A: Functional Characterization of Human Gut Microbiota in Inflammatory Bowel Disease Patients Using Gnotobiotic Humanized Mice. Freising: Technical University of Munich; 2020.

Metzger SA, Hernandez LL, Skarlupka JH, Suen G, Walker TM, Ruegg PL. Influence of sampling technique and bedding type on the milk microbiota: Results of a pilot study. J Dairy Sci 2018, 101(7):6346-6356.

Michail S, Durbin M, Turner D, Griffiths AM, Mack DR, Hyams J, Leleiko N, Kenche H, Stolfi A, Wine E. Alterations in the gut microbiome of children with severe ulcerative colitis. Inflamm Bowel Dis 2012, 18(10):1799-1808.

Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J *et al.* KatharoSeq Enables High-Throughput Microbiome Analysis from Low-Biomass Samples. mSystems 2018, 3(3).

Mosca A, Leclerc M, Hugot JP. Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem? Frontiers in microbiology 2016, 7(455).

Moxon ER. Applications of molecular microbiology to vaccinology. The Lancet 1997, 350(9086):1240-1244.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 1986, 51 Pt 1:263-273.

Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in enzymology 1987, 155:335-350.

Multinu F, Harrington SC, Chen J, Jeraldo PR, Johnson S, Chia N, Walther-Antonio MR. Systematic Bias Introduced by Genomic DNA Template Dilution in 16S rRNA Gene-Targeted Microbiota Profiling in Human Stool Homogenates. mSphere 2018, 3(2):e00560-00517.

Myer PR, McDaneld TG, Kuehn LA, Dedonder KD, Apley MD, Capik SF, Lubbers BV, Harhay GP, Harhay DM, Keele JW *et al.* Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing. PloS one 2020, 15(7):e0235498.

Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ 2018, 6:e5364.

Nel Van Zyl K, Whitelaw AC, Newton-Foot M. The effect of storage conditions on microbial communities in stool. PloS one 2020, 15(1):e0227486.

Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. PloS one 2014, 9(4):e94249.

Neuhaus K, Landstorfer R, Simon S, Schober S, Wright PR, Smith C, Backofen R, Wecko R, Keim DA, Scherer S. Differentiation of ncRNAs from small mRNAs in Escherichia coli O157:H7 EDL933 (EHEC) by combined RNAseq and RIBOseq - ryhB encodes the regulatory RNA RyhB and a peptide, RyhP. BMC genomics 2017, 18(1):216.

NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA *et al.* The NIH Human Microbiome Project. Genome research 2009, 19(12):2317-2323.

Nissle. Ueber die Grundlagen einer neuen ursächlichen Bekämpfung der pathologischen Darmflora1). Dtsch Med Wochenschr 1916, 42(39):1181-1184.

Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, Brodie EL, Malamud D, Poles MA, Pei Z. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol 2010, 16(33):4135-4144.

Nygaard AB, Tunsjø HS, Meisal R, Charnock C. A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. Scientific reports 2020, 10(1):3209.

Oikonomou G, Addis MF, Chassard C, Nader-Macias MEF, Grant I, Delbès C, Bogni CI, Le Loir Y, Even S. Milk Microbiota: What Are We Exactly Talking About? Frontiers in microbiology 2020, 11(60).

Oikonomou G, Bicalho ML, Meira E, Rossi RE, Foditsch C, Machado VS, Teixeira AGV, Santisteban C, Schukken YH, Bicalho RC. Microbiota of Cow's Milk; Distinguishing Healthy, Sub-Clinically and Clinically Diseased Quarters. PloS one 2014, 9(1):e85904.

Oikonomou G, Machado VS, Santisteban C, Schukken YH, Bicalho RC. Microbial diversity of bovine mastitic milk as described by pyrosequencing of metagenomic 16s rDNA. PloS one 2012, 7(10):e47671.

Ou F, McGoverin C, Swift S, Vanholsbeeck F. Absolute bacterial cell enumeration using flow cytometry. Journal of applied microbiology 2017, 123(2):464-477.

Pace NR, Sapp J, Goldenfeld N. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proceedings of the National Academy of Sciences 2012, 109(4):1011-1018.

Pacocha N, Scheler O, Nowak MM, Derzsi L, Cichy J, Garstecki P. Direct droplet digital PCR (dddPCR) for species specific, accurate and precise quantification of bacteria in mixed samples. Analytical Methods 2019, 11(44):5730-5735.

Panaccione R. Mechanisms of inflammatory bowel disease. Gastroenterol Hepatol (N Y) 2013, 9(8):529-532.

Park JC, Im S-H. Of men in mice: the development and application of a humanized gnotobiotic mouse model for microbiome therapeutics. Experimental & Molecular Medicine 2020, 52(9):1383-1396.

Park S-C, Won S. Evaluation of 16S rRNA Databases for Taxonomic Assignments Using Mock Community. Genomics Inform 2018, 16(4):e24-e24.

Parmar P, Lopez-Villalobos N, Tobin JT, Murphy E, McDonagh A, Crowley SV, Kelly AL, Shalloo L. The Effect of Compositional Changes Due to Seasonal Variation on Milk Density and the Determination of Season-Based Density Conversion Factors for Use in the Dairy Industry. Foods 2020, 9(8):1004.

Parracho HM, Bingham MO, Gibson GR, McCartney AL. Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children. J Med Microbiol 2005, 54(Pt 10):987-991.

Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, Martinez X, Varela E, Sarrabayrouse G, Machiels K *et al.* A microbial signature for Crohn's disease. Gut 2017, 66(5):813.

Penington JS, Penno MAS, Ngui KM, Ajami NJ, Roth-Schulze AJ, Wilcox SA, Bandala-Sanchez E, Wentworth JM, Barry SC, Brown CY *et al.* Influence of fecal collection conditions and 16S rRNA gene sequencing at two centers on human gut microbiota analysis. Scientific reports 2018, 8(1):4386.

Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microbial genomics 2020, 6(8):mgen000409.

Petersen C, Round JL. Defining dysbiosis and its influence on host immunity and disease. Cell Microbiol 2014, 16(7):1024-1033.

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific reports 2018, 8(1):10950.

Pinna NK, Dutta A, Monzoorul Haque M, Mande SS. Can Targeting Non-Contiguous V-Regions With Paired-End Sequencing Improve 16S rRNA-Based Taxonomic Resolution of Microbiomes?: An In Silico Evaluation. Frontiers in genetics 2019, 10(653).

Podolsky SH. Metchnikoff and the microbiome. The Lancet 2012, 380(9856):1810-1811.

Pollock J, Glendinning L, Wisedchanwet T, Watson M. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Applied and environmental microbiology 2018, 84(7).

Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, Polozov RV, Nechipurenko YD, Grokhovsky SL. Non-random DNA fragmentation in next-generation sequencing. Scientific reports 2014, 4:4532-4532.

Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PloS one 2020, 15(1):e0227434.

Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 2012, 28(14):1823-1829.

Pruvost M, Grange T, Geigl E-M. Minimizing DNA contamination by using UNG-coupled quantitative real-time PCR on degraded DNA samples: application to ancient DNA studies. BioTechniques 2005, 38(4):569-575.

Quan L, Dong R, Yang W, Chen L, Lang J, Liu J, Song Y, Ma S, Yang J, Wang W *et al.* Simultaneous detection and comprehensive analysis of HPV and microbiome status of a cervical liquid-based cytology sample using Nanopore MinION sequencing. Scientific reports 2019, 9(1):19337-19337.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 2013, 41(D1):D590-D596.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature biotechnology 2017, 35(9):833-844.

Rainard P. Mammary microbiota of dairy ruminants: fact or fiction? Veterinary research 2017, 48(1):25-25.

Rausch P, Rühlemann M, Hermes BM, Doms S, Dagan T, Dierking K, Domin H, Fraune S, von Frieling J, Hentschel U *et al.* Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 2019, 7(1):133.

Reitmeier S: Arrhythmic Gut Microbiome Signatures for Type 2 Diabetes Risk Profiling. Freising: Technical University of Munich; 2021.

Reitmeier S, Hitch TCA, Treichel N, Fikas N, Hausmann B, Ramer-Tait AE, Neuhaus K, Berry D, Haller D, Lagkouvardos I *et al.* Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. ISME Communications 2021.

Reitmeier S, Kiessling S, Neuhaus K, Haller D. Comparing Circadian Rhythmicity in the Human Gut Microbiome. STAR Protocols 2020:100148.

Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 2015, 13(5):278-289.

Robinson CJ, Bohannan BJ, Young VB. From structure to function: the ecology of host-associated microbial communities. Microbiology and molecular biology reviews : MMBR 2010, 74(3):453-476.

Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. Bioinformatics 2020.

Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science (New York, NY) 1998, 281(5375):363, 365.

Rosselli R, Romoli O, Vitulo N, Vezzi A, Campanaro S, de Pascale F, Schiavon R, Tiarca M, Poletto F, Concheri G *et al.* Direct 16S rRNA-seq from bacterial communities: a PCR-

independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon. Scientific reports 2016, 6:32165.
Rubin DC, Shaker A, Levin MS. Chronic intestinal inflammation: inflammatory bowel disease and colitis-associated colon cancer. Frontiers in immunology 2012, 3:107-107.
Saladié M, Caparrós-Martín JA, Agudelo-Romero P, Wark PAB, Stick SM, O'Gara F. Microbiomic Analysis on Low Abundant Respiratory Biomass Samples; Improved Recovery of Microbial DNA From Bronchoalveolar Lavage Fluid. Frontiers in microbiology 2020, 11(2477).
Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology 2014, 12(1):87.
Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 1977, 74(12):5463-5467.
Santiago-Rodriguez TM, Garoutte A, Adams E, Nasser W, Ross MC, La Reau A, Henseler Z, Ward T, Knights D, Petrosino JF *et al.* Metagenomic Information Recovery from Human Stool Samples Is Influenced by Sequencing Depth and Profiling Method. Genes (Basel) 2020, 11(11):1380.
Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, Guarner F, Manichanh C. Processing faecal samples: a step forward for standards in microbial community analysis. BMC microbiology 2014, 14(1):112.
Santos A, van Aerle R, Barrientos L, Martinez-Urtaza J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. Comput Struct Biotechnol J 2020, 18:296-305.
Scaldaferri F, Gerardi V, Mangiola F, Lopetuso LR, Pizzoferrato M, Petito V, Papa A, Stojanovic J, Poscia A, Cammarota G *et al.* Role and mechanisms of action of Escherichia coli Nissle 1917 in the maintenance of remission in ulcerative colitis patients: An update. World J Gastroenterol 2016, 22(24):5505-5511.
Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Human Molecular Genetics 2010, 19(R2):R227-R240.
Schaedler RW, Dubs R, Costello R. Association of Germfree Mice with Bacteria Isolated from Normal Mice. J Exp Med 1965, 122(1):77-82.
Schierwagen R, Alvarez-Silva C, Servant F, Trebicka J, Lelouvier B, Arumugam M. Trust is good, control is better: technical considerations in blood microbiome analysis. Gut 2020, 69(7):1362.
Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW, Ananthakrishnan AN, Andrews E, Barron G, Lake K *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nature microbiology 2018, 3(3):337-346.
Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PloS one 2011, 6(12):e27310-e27310.
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.
Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature methods 2016, 13(5):435-438.
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.
Shakya M, Lo C-C, Chain PSG. Advances and Challenges in Metatranscriptomic Analysis. Frontiers in genetics 2019, 10(904).
Shanahan F, Ghosh TS, O'Toole PW. The Healthy Microbiome—What Is the Definition of a Healthy Gut Microbiome? Gastroenterology 2021, 160(2):483-494.

Library preparation protocol to sequence full length 16S rRNA gene in Nanopore MinION sequencer

Sheahan T, Hakstol R, Kailasam S, Glaister GD, Hudson AJ, Wieden H-J. Rapid metagenomics analysis of EMS vehicles for monitoring pathogen load using nanopore DNA sequencing. PloS one 2019, 14(7):e0219961-e0219961.

Siebert A, Hofmann K, Staib L, Doll EV, Scherer S, Wenning M. Amplicon-sequencing of raw milk microbiota: impact of DNA extraction and library-PCR. Applied Microbiology and Biotechnology 2021.

Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, Corby P, Guo Y, Ruff RR, Alekseyenko AV *et al.* The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community. Genes (Basel) 2020, 11(8):878.

Siezen RJ, Wilson G. Probiotics genomics. Microb Biotechnol 2010, 3(1):1-9.

Sinha R, Abnet CC, White O, Knight R, Huttenhower C. The microbiome quality control project: baseline study design and future directions. Genome biology 2015, 16(1):276.

Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Abnet CC *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nature biotechnology 2017, 35(11):1077-1086.

Sinha R, Chen J, Amir A, Vogtmann E, Shi J, Inman KS, Flores R, Sampson J, Knight R, Chia N. Collecting Fecal Samples for Microbiome Analyses in Epidemiology Studies. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2016, 25(2):407-416.

Smith KA. Louis pasteur, the father of immunology? Frontiers in immunology 2012, 3:68-68.

Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems 2016, 1(3):e00021-00016.

Sonnenborn U. Escherichia coli strain Nissle 1917—from bench to bedside and back: history of a special Escherichia coli strain with probiotic properties. FEMS microbiology letters 2016, 363(19).

Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, Gessner A, Spang R. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 2016, 4(1):28.

Stapleton JA, Kim J, Hamilton JP, Wu M, Irber LC, Maddamsetti R, Briney B, Newton L, Burton DR, Brown CT *et al.* Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing. PloS one 2016, 11(1):e0147229.

Stinson LF, Keelan JA, Payne MS. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. Lett Appl Microbiol 2019, 68(1):2-8.

Straub D, Blackwell N, Langarica-Fuentes A, Peltzer A, Nahnsen S, Kleindienst S. Interpretations of Environmental Microbial Community Studies Are Biased by the Selected 16S rRNA (Gene) Amplicon Sequencing Pipeline. Frontiers in microbiology 2020, 11(2652).

Sze MA, Schloss PD. The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. mSphere 2019, 4(3):e00163-00119.

Tal M, Verbrugghe A, Gomez DE, Chau C, Weese JS. The effect of storage at ambient temperature on the feline fecal microbiota. BMC Vet Res 2017, 13(1):256.

Taponen S, McGuinness D, Hiitiö H, Simojoki H, Zadoks R, Pyörälä S. Bovine milk microbiome: a more complex issue than expected. Veterinary Research 2019, 50(1):44.

Tedjo DI, Jonkers DMAE, Savelkoul PH, Masclee AA, van Best N, Pierik MJ, Penders J. The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. PloS one 2015, 10(5):e0126685-e0126685.

Tembe WD, Pond SJK, Legendre C, Chuang H-Y, Liang WS, Kim NE, Montel V, Wong S, McDaniel TK, Craig DW *et al.* Open-access synthetic spike-in mRNA-seq data for cancer gene fusions. BMC genomics 2014, 15(1):824.

Teng F, Darveekaran Nair SS, Zhu P, Li S, Huang S, Li X, Xu J, Yang F. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. Scientific reports 2018, 8(1):16321.

Theron J, Cloete TE. Molecular Techniques for Determining Microbial Diversity and Community Structure in Natural Environments. Critical Reviews in Microbiology 2000, 26(1):37-57.

Thijs S, Op De Beeck M, Beckers B, Truyens S, Stevens V, Van Hamme JD, Weyens N, Vangronsveld J. Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. Frontiers in microbiology 2017, 8:494-494.

Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2012, 2(1):3-3.

Thursby E, Juge N. Introduction to the human gut microbiota. Biochem J 2017, 474(11):1823-1836.

Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic acids research 2017, 45(4):e23.

Trautwein C. In aller Munde und in jedem Darm: unser intestinales Mikrobiom. Der Klinikarzt 2020, 49(12):528-529.

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. Primer and platform effects on 16S rRNA tag sequencing. Frontiers in microbiology 2015, 6:771-771.

Tritt A, Eisen JA, Facciotti MT, Darling AE. An Integrated Pipeline for de Novo Assembly of Microbial Genomes. PloS one 2012, 7(9):e42304.

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.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature 2007, 449(7164):804-810.

Turner S, Pryer KM, Miao VP, Palmer JD. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol 1999, 46(4):327-338.

Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutrition reviews 2012, 70 Suppl 1:S38-44.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends in Genetics 2018, 34(9):666-681.

Vandeputte D, Kathagen G, D'Hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y *et al.* Quantitative microbiome profiling links gut community variation to microbial load. Nature 2017, 551(7681):507-511.

Velásquez-Mejía EP, de la Cuesta-Zuluaga J, Escobar JS. Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces. Appl Microbiol Biotechnol 2018, 102(1):403-411.

Venkataraman A, Parlov M, Hu P, Schnell D, Wei X, Tiesman JP. Spike-in genomic DNA for validating performance of metagenomics workflows. BioTechniques 2018, 65(6):315-321.

Vetrovsky T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PloS one 2013, 8(2):e57923.

Videnska P, Smerkova K, Zwinsova B, Popovici V, Micenkova L, Sedlar K, Budinska E. Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform. Scientific reports 2019, 9(1):13837.

Villette R, Autaa G, Hind S, Holm JB, Moreno-Sabater A, Larsen M. Refinement of 16S rRNA gene analysis for low biomass biospecimens. Scientific reports 2021, 11(1):10741.

Voelkerding KV, Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry 2009, 55(4):641-658.

Voigt AY, Costea PI, Kultima JR, Li SS, Zeller G, Sunagawa S, Bork P. Temporal and technical variability of human gut metagenomes. Genome biology 2015, 16(1):73.

Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proceedings of the National Academy of Sciences of the United States of America 2018, 115(39):9726-9731.

Vos M, Quince C, Pijl AS, de Hollander M, Kowalchuk GA. A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity. PloS one 2012, 7(2):e30600.

Wagner Mackenzie B, Waite DW, Taylor MW. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Frontiers in microbiology 2015, 6(130).

Walden A. Dr. Louis Pasteur. Primary Care Update for OB/GYNS 2003, 10(2):68-70.

Wang Y, Qian P-Y. Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies. PloS one 2009, 4(10):e7401.

Wassenaar TM. Insights from 100 Years of Research with Probiotic E. Coli. Eur J Microbiol Immunol (Bp) 2016, 6(3):147-161.

Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, Licht TR. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome 2014, 2(1):19.

Wessner DR. The Origins of Viruses. Nature Education 2010, 3(9):37.

Whiteside SA, Razvi H, Dave S, Reid G, Burton JP. The microbiome of the urinary tract—a role beyond infection. Nature Reviews Urology 2015, 12(2):81-90.

Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences 1998, 95(12):6578-6583.

Wilkinson MG. Flow cytometry as a potential method of measuring bacterial viability in probiotic products: A review. Trends in Food Science & Technology 2018, 78:1-10.

Wilson DN, Doudna Cate JH. The structure and function of the eukaryotic ribosome. Cold Spring Harb Perspect Biol 2012, 4(5):a011536.

Winand R, Bogaerts B, Hoffman S, Lefevre L, Delvoye M, Braekel JV, Fu Q, Roosens NH, Keersmaecker SC, Vanneste K. Targeting The 16S rRNA Gene For Bacterial Identification In Complexed Mixed Samples: Comparative Evaluation of Second (Illumina) and Third (Oxford Nanopore Technologies) Generation Sequencing Technologies. Int J Mol Sci 2019, 21(1).

Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences 1977, 74(11):5088.

Woese CR, Stackebrandt E, Macke TJ, Fox GE. A Phylogenetic Definition of the Major Eubacterial Taxa. Systematic and Applied Microbiology 1985, 6(2):143-151.

Wouters Y, Dalloyaux D, Christenhusz A, Roelofs HMJ, Wertheim HF, Bleeker-Rovers CP, te Morsche RH, Wanten GJA. Droplet digital polymerase chain reaction for rapid broad-spectrum detection of bloodstream infections. Microb Biotechnol 2020, 13(3):657-668.

Wu GD, Lewis JD, Hoffmann C, Chen Y-Y, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L *et al.* Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC microbiology 2010a, 10(1):206.

Wu J-Y, Jiang X-T, Jiang Y-X, Lu S-Y, Zou F, Zhou H-W. Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. BMC microbiology 2010b, 10:255-255.

Xu L, Seki M. Recent advances in the detection of base modifications using the Nanopore sequencer. Journal of Human Genetics 2020, 65(1):25-33.

Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 2016, 17(1):135.

Yang F, Sun J, Luo H, Ren H, Zhou H, Lin Y, Han M, Chen B, Liao H, Brix S *et al.* Assessment of fecal DNA extraction protocols for metagenomic studies. GigaScience 2020, 9(7).

Yoon S-H, Ha S-M, Kwon S, Lim J, Kim Y, Seo H, Chun J. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 2017, 67(5):1613-1617.

Yu K, Xing J, Zhang J, Zhao R, Zhang Y, Zhao L. Effect of multiple cycles of freeze-thawing on the RNA quality of lung cancer tissues. Cell and tissue banking 2017, 18(3):433-440.

Zemb O, Achard CS, Hamelin J, De Almeida M-L, Gabinaud B, Cauquil L, Verschuren LMG, Godon J-J. Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard. MicrobiologyOpen 2020, 9(3):e977.

Zhang S, Cao X, Huang H. Sampling Strategies for Three-Dimensional Spatial Community Structures in IBD Microbiota Research. Frontiers in Cellular and Infection Microbiology 2017, 7(51).

Zhulin IB. Classic Spotlight: 16S rRNA Redefines Microbiology. J Bacteriol 2016, 198(20):2764.

Ziegler I, Lindström S, Källgren M, Strålin K, Mölling P. 16S rDNA droplet digital PCR for monitoring bacterial DNAemia in bloodstream infections. PloS one 2019, 14(11):e0224656.

Zoetendal EG, Booijink CC, Klaassens ES, Heilig HG, 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.

5. Supplement

Abbreviations

ASV	Amplicon Sequence Variant
bp	base pair
CD	Crohn's Disease
cDNA	complementary DNA
CSS	Circular Consensus Sequencing
ddPCR	digital droplet PCR
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
dNTPs	deoxynucleotide Triphosphates
DTT	dithiothreitol
F/B	*Firmicutes*/*Bacteroides* ratio
Fw	forward
gDNA	genomic DNA
HMP	Human Microbiome Project
IBD	Inflammatory Bowel Disease
IMNGS	Integrated Microbial Next Generation Sequencing
LSU rRNA	Large Subunit ribosomal Ribonucleic Acid
MDS	Multidimensional Scaling
MOPS	3-(N-morpholino)propanesulfonic acid
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NIH	National Institutes of Health
NMDS	Non-metric Multidimensional Scaling
ONT	Oxford Nanopore Technologies
OTU	Operational Taxonomic Unit
PacBio	Pacific Biosciences
PAGE	Polyacrylamide Gel Electrophoresis
PCR	Polymerase Chain Reaction
RDP	Ribosomal Database Project
RNA	Ribonucleic Acid
rRNA	ribosomal Ribonucleic Acid
RT	Room Temperature (approximately 22°C)
rv	reverse
SMRT	Single Molecule Real-Time
SSU rRNA	Small Subunit ribosomal Ribonucleic Acid
UC	Ulcerative Colitis
UMI	Unique Molecular Identifier
v	version
V-regions	variable Regions of the 16S rRNA gene
ZMWS	Zero-Mode Waveguides
zOTU	zero-radius OTU

Publications and Presentations

Peer-reviewed Publications

Abellan-Schneyder I, Siebert A, Hofmann K, Wenning M, Neuhaus K. Full-length SSU rRNA gene sequencing allows species-level detection of bacteria, archaea, and yeasts present in milk. Microorganisms 2021, 9(6):1251.

Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K (2021). Primer, pipelines, parameters: Issues in 16S rRNA gene sequencing. mSphere 2021, 6(1):e01202-20.

Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Scherer S, Neuhaus K. The novel anaerobiosis-responsive overlapping gene *ano* is overlapping antisense to the annotated gene ECs2385 of *Escherichia coli* O157:H7 Sakai. Front Microbiol. 2018, 9:931.

Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Wecko R, Simon S, Scherer S, Neuhaus K. A novel short L-arginine responsive protein-coding gene (*laoB*) antiparallel overlapping to a CadC-like transcriptional regulator in *Escherichia coli* O157:H7 Sakai originated by overprinting. BMC Evol. Biol. 2018, 18(1):21.

Conference Presentation

10/2020	FEMS Online Conference on Microbiology, Oral Presentation: “Full-length versus short amplicon 16S rRNA sequencing: benefits, drawbacks and risks”.
10/2020	Cold Spring Harbor Laboratory Conference Microbiome, Poster Presentation: “Comparability of 16S rRNA sequencing data depends on multiple parameters and study-cross validations are mandatory”.
07/2019	12th Seeon Conference – Microbiota, Probiotics & Host, Poster Presentation: “Full-length 16S rRNA sequencing of rRNA genes or the actual rRNA allows a deeper insight into strains present in stool samples”.
03/2019	GQ 2019 Event – qPCR, dPCR & NGS 2019, Poster Presentation: “Full-length 16S rRNA Sequencing with the Illumina Barcode Structure Allows a Deeper Insight into Strains Present in Stool Samples”.
04/2018	VAAM, Poster Presentation: “Full-length 16S rRNA Sequencing – New Perspectives in Microbiome Research”.

Acknowledgment

Ich danke meinem Doktorvater Herrn Priv.-Doz. Dr. Klaus Neuhaus ganz herzlich für die Möglichkeit zur Promotion. Lieben Dank für deine kontinuierliche Unterstützung, dein offenes Ohr und dein Vertrauen in mich und meine Arbeit. Danke auch, dass du so eine tolle Arbeitsgruppe zusammengestellt hast, in der jeder seinen Platz hat und sich wohlfühlt.

Weiterhin Danke ich Herrn apl. Prof. Dr. Michael Pfaffl für die Übernahme des Prüfungsvorsitzes und Frau Prof. Dr. Lindsay Hall, dass sie meine Zweitprüferin ist.

Lieben Dank an Dr. Ilias Lagkouvardos, Dr. Sandra Reitmeier und Zeno Sewald für die Unterstützung und Hilfe bezüglich jeglicher bioinformatischen Fragestellungen. Danke auch für die vielen wissenschaftlichen und nichtwissenschaftlichen Ratschläge.

Meinen Kooperationspartnern: Monica Steffi Matchado, Dr. Markus List, Annemarie Siebert, Katharina Hofmann, Mohammed Ahmed und Dr. Amira Metwaly sei auch hier nochmal herzlich gedankt.

Ein großer Dank und ein Lob für die wunderbare Arbeit geht an Caroline Ziegler, Angela Sachsenhauser, Lukas Mix, Christine Fritsch, Romy Wecko, Sandra Nagl und Petra Hartberger. Danke für die Hilfe im Labor, die nette Stimmung und das gute Miteinander.

Im Laufe meiner Promotion durfte ich viele großartige Studierende kennenlernen und auf ihrem Weg begleiten. Ein großer Dank geht an meine Forschungspraktikanten: Max Ruscheweyh, Florian Rothfischer, Ipek Eroglu, Andrea Isabel Proaño Vasco, Alina Sommer, Maximilian Grimm und Li Tran. Weiterhin durfte ich vier Abschlussarbeiten betreuen. Liebe Annika Naumann und lieber Maximilian Grimm, danke für eure Arbeit im Rahmen einer Bachelorarbeit. Jeweils sechs Monate durfte ich Andrea Schusser und Michael Singer im Rahmen ihrer Masterarbeiten betreuen und mit ihnen zusammenarbeiten. Lieber Michi, liebe Andi, von Herzen einen lieben Dank für eure super Arbeit und das wunderbare Miteinander. Liebe Andrea, danke, dass du nach deinem Forschungspraktikum noch lange als HiWi für mich und die Core Facility gearbeitet hast. Ich freue mich, dass du in Freiburg nun deine eigenen Studierenden betreuen wirst, ich bin stolz auf dich. Lieber Flo, danke, dass ich dich vom Beginn deines Studiums bis zum Beginn und hoffentlich auch der Fertigstellung deiner eigenen Doktorarbeit begleiten durfte, bleib wie du bist!

Allen ehemaligen und aktuellen Doktorand/innen, sowohl des Lehrstuhls für Mikrobielle Ökologie (Prof. Siegfried Scherer) als auch der Core Facility Microbiome danke ich für die gute Stimmung, unsere gemeinsame Aktivitäten und die gute Zeit die wir zusammen erleben durften. Hervorzuheben sind hierbei ein paar Namen.

Liebe Sarah, danke, dass ich bei dir mein Praktikum und meine Masterarbeit absolvieren durfte. Weiterhin bin ich dir sehr dankbar für deine vielen Ratschläge und vor allem dafür, dass du mich dazu motiviert hast, mich auf die Stelle bei Klaus zu bewerben. Ich hätte mir damals und auch heute für die Masterarbeit keine bessere Betreuerin vorstellen können.

Liebe Alina, danke für unsere Freundschaft und die Ablenkung vom Laboralltag. Liebe Micha, danke, dass du zum Ende der Promotion mein Wochenend-Buddy warst.

Dear Mohsen, I am very happy that you joined the lab. Stay the way you are but, you must watch Harry Potter! Liebe Anika und liebe Franzi, viel Erfolg für eure anstehenden Doktorarbeiten. Ihr werdet das klasse machen. Danke auch für die viele Unterstützung in den letzten Zügen meiner Arbeit.

Von Herzen Danke ich Franziska Giehren und Sandra Reitmeier, ihr seid und werdet einfach immer die besten Kolleginnen (und vielleicht auch Freundinnen, oder was meinst du Franzi?) bleiben. Danke für die vielen gemeinsamen Stunden, das gemeinsame Leiden, Freuen und die große Unterstützung.

Meinen Studienfreundinnen Jacqui, Katha und Tami sei hier auch ganz herzlich gedankt. Ebenso, danke Sebi, dass du mich stets aufzumuntern wusstest und immer ein offenes Ohr hattest. Danke auch an das wunderbare GoBiochem-Team allen voran Makis und Flo. Meinen Freundinnen Johanna und Constanze danke ich von ganzem Herzen für unsere wunderbare Freundschaft und die beste Ablenkung die man sich wünschen kann. Allen Zumba-Mädels besonders Vroni und Kim sei für die besten Montage gedankt.

Meinen Eltern Bettina und José sowie meinen Großeltern und Schwiegereltern in spe danke ich von Herzen für ihre Unterstützung. Liebe/r Franci, Luis, Oscar, Anne und Mateo, danke, dass ihr immer für mich da seid und mich stets zum Lachen bringen könnt. Ich gehe jetzt erstmal einen Kakao trinken.

Lieber Michi, danke, dass du mich nicht nur seit unserem gemeinsamen Bachelor auf meinem Weg stets unterstützt und bestärkt hast, sondern egal wie und wo immer für mich da warst. Du bist der Beste.