General introduction, aim and outline of the thesis

Partially based on: Boers SA, Jansen R, Hays JP. Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) for use in the routine clinical microbiological diagnostic laboratory. Submitted for publication.

GENERAL INTRODUCTION

Over millions of years of co-evolution, microorganisms (including bacteria, archaea, fungi, protists and viruses) have adapted to form microbial communities that occupy virtually every accessible environmental niche, such as in or on living organisms (plant or animal life), soil, oceans, and air. There, these microbial communities can participate in important biological processes, such as biogeochemical processes that sustain life on our planet.[1] Humans also possess such microbial communities, where microorganisms usually live in close harmony with their human host, and with each other, forming symbiotic relationships that have a central role in the development and promotion of human health and disease.[2] The current recognition of the essential importance of these communities means that the microbial composition, structure and function of a wide variety of microbial communities are now being actively investigated by the scientific and medical community, from microbial communities on the International Space Station (ISS) to communities collected from many different human body sites here on earth.[3,4] Importantly however, the rapid increase of research activities within this field has been accompanied by confusion in the vocabulary used to describe different aspects of the microbial communities and environments under investigation. In order to avoid confusion, in this thesis the terms used to describe microbial community analysis are based on those terms defined previously by Marchesi and Ravel: *microbiota, metagenome* and *microbiome*.[5]

The microorganisms present within a defined environment is referred to as the *microbiota*, and the assemblage of their genomes (i.e. genes) as the *metagenome*. The term *microbiome* refers to the entire habitat, including the *microbiota*, *metagenome* and the surrounding environmental conditions (Figure 1).

History of microbiome research

Early investigations into the microbial communities from different environments focused on traditional techniques for isolating and culturing individual microorganisms. Although these culture-based methods were able to determine the viable population within a particular environment using broad-range or selective artificial growth media, obtaining a comprehensive overview of the microbial communities using these culturing methods was proven difficult as many microorganisms require specific growth conditions that cannot be (easily) mimicked within a laboratory environment.[6] However, more recent advances in technologies able to detect the presence of microbial genes (via DNA amplification and sequencing), such as the polymerase chain reaction (PCR),[7] dideoxy termination sequencing (Sanger sequencing),[8] and more recently next-generation sequencing (NGS),[9] means that it is now possible to detect a theoretically unlimited number of microorganisms, present in all kinds of microbial samples, using

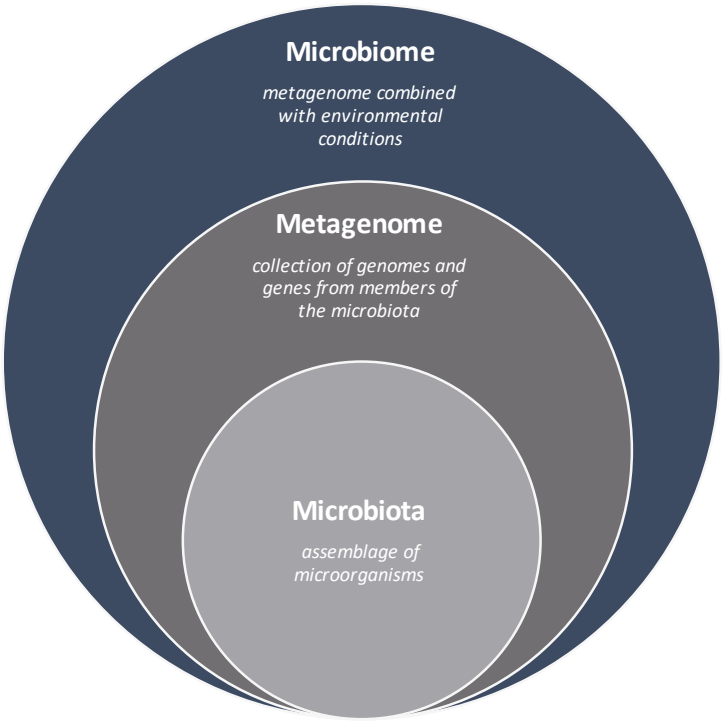

Figure 1. Differentiation of terms used to describe different aspects of research that focus on microbial communities and their environments.

a culture-independent approach. Specifically, Venter and colleagues were the first research group to apply DNA sequencing-based methods on a large scale in order to study microbial dynamics within environmental samples.[10] As a proof of concept, Venter et al. investigated water samples obtained from the Sargasso See, as it was thought that this region of the North Atlantic Ocean contained only a small number of microbial species due to its low nutrient levels. Surprisingly however, their research revealed the presence of at least 1,800 different microbial species, including 148 new bacterial species and over 1.2 million previously unknown genes. This pioneering research illustrated that DNA sequencing-based methods, which are not hampered by the traditional limitations associated with microbial culture, generate more comprehensive characterizations of microbial communities.

The human microbiome and associations with disease

In 2006, Gill and colleagues used the same culture-independent methodology, as described by Venter et al., in order to study the human microbiome.[11] Their study revealed that the microbiome of the human gastrointestinal tract encodes for a larger portion of metabolic pathways – that are important for a healthy human's metabolism – than the

human genome itself. This finding highlighted the crucial importance of the human gut microbiome in health and lay the groundwork for further research to discover new associations between the human microbiota and disease. In the following years, a tremendous amount of (circumstantial) evidence has been collected to suggest a crucial role for the human gut microbiota in health and disease, including for example, in allergic diseases,[12-14] inflammatory bowel diseases,[15,16] and metabolic diseases.[17,18] Additionally, recent discoveries also suggest that the gut microbiota are able to influence psychological disorders, such as anxiety and depressive-like behaviours, via the gut-brain axis.[19] However, the best evidence to indicate the importance of the human gut microbiota in health and disease comes from the clinic, where patients are treated with antibiotics. Antibiotics change the normal composition of the healthy gut microbiota, generating dysbiosis and facilitating the overgrowth of pathobionts such as *Clostridium difficile* bacteria, which are responsible for recurrent diarrhoea.[20] Patients infected with *C. difficile* may be transplanted with a healthy gut microbiota that restores the healthy microbial gut composition, thereby reversing dysbiosis and preventing recurrent episodes of diarrhoea. These so-called faecal microbiota transplantations (FMT) have proven to be more successful for treating recurrent *C. difficile* infections than prescribing yet more antibiotics in order to try to kill or inhibit the overgrowth of *C. difficile*.[21] Interestingly, FMT has also showed promising results for patients diagnosed with Crohn's disease as well.[22,23]

The importance of microbiota detection in routine clinical microbiological diagnostics

The culture-independent microbiota profiling methods used to detect and identify all microbial taxa within a sample should be available not only for research purposes, but also to routine clinical microbiological diagnostics, where the detection and identification of microbial pathogens is the major step in establishing appropriate antimicrobial treatment for infectious diseases. For a long time, routine clinical microbiological diagnostic testing has been performed almost exclusively using culture-based methods that have been highly optimized for the efficient cultivation of known clinically-relevant microorganisms. However, the causative agent of an infection may not always be detected using current 'gold standard' culturing methods and, therefore, culture-independent molecular detection methods are required to identify 'non-culturable' microorganisms. For example, the discovery of the causative pathogens of bacillary angiomatosis (*Bartonella quintana*) and Whipple's disease (*Tropheryma whipplei*) were made possible using Sanger sequencing-based methods, as both aerobic bacteria are very difficult to culture in a laboratory.[24,25] In addition, the use of NGS-based methods has also been shown to improve the detection of obligate anaerobic bacteria in clinical samples.[26,27] Obligate anaerobes are known to cause serious infections, yet their detection may be sub-optimal within routine clinical microbiological diagnostic laboratories as special precautions

are required to help preserve the anaerobic environment during specimen collection, transport and culture.[28] Therefore, culture-independent microbiota profiling methods could play an important role in the identification of the aetiology of anaerobic infections, or any other infections caused by fastidious and/or unexpected microorganisms. A second important point is that obtaining a comprehensive overview of polymicrobial populations within clinical samples means that the whole microbial community per se could be taken into account when making clinical decisions. However, before steps can be taken to implement such testing in the routine clinical microbiological diagnostic laboratory, it is important to understand the current NGS-based methodologies available for characterizing microbial communities, and the potential pitfalls and biases that can influence the results obtained. Armed with this information, the aim and outline of the current thesis will become clearer to the reader.

NGS-based methodologies for characterizing microbial communities

The advent of NGS has enabled researchers to investigate the composition and function of microbial populations in very diverse environments with unprecedented resolution and throughput. Currently, the majority of these investigations apply NGS by focussing on either targeted amplicon sequencing with the 16S ribosomal RNA (rRNA) gene as phylogenetic target (i.e. 16S rRNA gene NGS) or on shotgun metagenomics. A general overview of both methods is shown in Figure 2 and the strengths and weaknesses of each method will be discussed in the following section.

Targeted amplicon sequencing. Amplicon sequencing methods have been widely used as a targeted approach for characterizing microbial communities. Here, DNA is extracted from all cells in a sample and subjected to PCR amplification using a taxonomically informative genetic marker that is common to virtually all microorganisms of interest. The resultant amplicons are sequenced and then characterized using bioinformatics tools in combination with reference databases to determine which microorganism are present in the sample and at what relative abundance. Advances in this technology now mean that the latest amplicon-based NGS protocols enable extensive multiplexing, which allows researchers to process and analyse millions of PCR amplicons derived from hundreds of samples on a single NGS-run.[29]

The 16S rRNA gene is by far the most established genetic marker used for prokaryotic identification and classification ever since Woese and Fox first utilized rRNA sequence characterization to define the three domains of life in 1977.[30] Because the 16S rRNA gene encodes for the RNA component of the small subunit (SSU) of prokaryotic ribosomes, which performs essential functions within the translation process, it is present among all bacteria and archaea and possess a slow rate of evolution that allows researchers to infer microbial phylogenetic relationships. The 16S rRNA gene is approximately 1,500

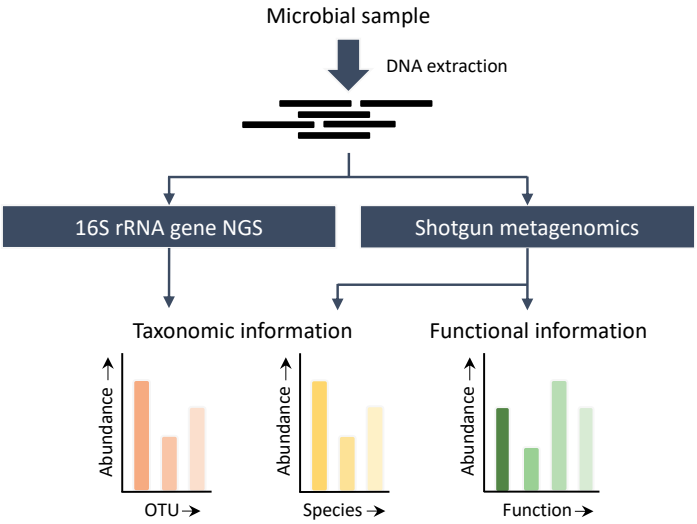

Figure 2. General overview of 16S rRNA gene NGS and shotgun metagenomics methods. Both methods start with the extraction of nucleic acids from a microbial sample. Next, the extracted DNA is either subjected to 16S rRNA gene PCR amplification (16S rRNA gene NGS) or sheared into small DNA fragments (shotgun metagenomics). The resultant 16S rRNA gene amplicons, or sheared DNA fragments, are sequenced using NGS-based techniques. Finally, all sequence data are processed using an extensive array of bioinformatics algorithms that allows the researcher to explore the taxonomic composition and/or the functional capacity of the sample tested.
OTU = operational taxonomic units, a group of very similar sequences.

base pairs (bp) in size and its gene structure is defined by an alteration of nine highly conserved and nine hypervariable regions (V1-V9). The conserved regions can serve as universal primer binding sites for the PCR amplification of gene fragments, whereas the hypervariable regions contain considerable sequence diversity, useful for prokaryotic identification.[31] By comparing these hypervariable regions to 16S rRNA gene sequences of designated type strains that are available on large public databases (e.g. SILVA, RDP, GreenGenes, or NCBI), researchers can obtain accurate taxonomic identifications of prokaryotic taxa.[32-35] However, it is important to note that the sequencing of partial 16S rRNA genes, which is currently the most commonly used microbiota profiling strategy, often lacks the discriminatory power to differentiate prokaryotes at the species taxonomic level and is generally restricted to genus-level classifications.[36] For this reason, there has been a continuous search for alternative marker genes that can improve phylogenetic resolution among prokaryotic species. For example, sequence-based analysis of the *rpoB* gene has previously been demonstrated to improve the discriminative power for characterizing prokaryotic species (when compared to 16S rRNA gene sequencing methods) among several bacterial families and genera, including *Bacillus*,[37] *Enterobacteriaceae*,[38] *Staphylococcus*,[39] and others.[40] The *rpoB* gene encodes the highly

conserved beta subunit of the prokaryotic RNA polymerase and apparently possesses the same key attributes as the 16S rRNA gene.[41] However, 16S rRNA gene sequencing studies profit from the massive amounts of sequence information already available in large publicly accessible reference databases. Hence, although alternative phylogenetic markers such as *rpoB* (and many others) are very promising,[42] these biomarkers still face the challenge of competing with thousands of publications that utilize extensive databases of 16S rRNA gene sequencing information.

The characterization of eukaryotic communities is also an active research area that often employs targeted amplicon sequencing approaches. For this, the 18S rRNA gene, which is the eukaryotic nuclear homologue of the 16S rRNA gene in prokaryotes, have been used as a genetic marker in studies investigating fungi and protists. For example, novel phylogenetic groups of fungal microorganisms have been defined using 18S rRNA gene based sequencing,[43] and a diversity of small eukaryotes were for the first time reported at high ocean depths (250 – 3,000 meters) using the same method.[44] Despite these efforts, a multi-laboratory consortium proposed the nuclear ribosomal internal transcribed spacer (ITS) region as the primary genetic marker for fungi in 2012.[45] The ITS region was preferred over the 18S rRNA gene due to the higher sequence variability in the ITS region and the presence of a more curated and comprehensive reference database. Nevertheless, it is argued that the uneven lengths of ITS fragments may promote preferential PCR amplification of shorter ITS sequences that could lead to a biased quantification of relative abundances of fungal taxa and, therefore, the (additional) use of non-ITS targets in sequencing-based microbiota studies for fungi is desirable.[46]

Finally, the detection and characterization of viruses requires a different detection approach altogether. Unlike for cellular life forms, there is not a single gene or genomic region that is homologous across all viral genomes.[47] For virus detection, microarrays that span the 'middle ground' between NGS-based and PCR-based methodologies have been developed. These microarrays are designed to detect known viruses (including phages), sometimes in combination with the simultaneous detection of prokaryotes and microbial eukaryotes.[48-50] The main advantage of these methods is the ability to simultaneously test for the presence of hundreds of viruses in a single assay and thereby remove the need for an *a priori* knowledge of the presence of a suspected virus. However, the range of detectable viruses is limited by the content of the viral probes that are initially spotted on the detection microarray, which may not represent the full genetic diversity of a viral community derived from a microbial sample.

Shotgun metagenomics. Shotgun metagenomics is an alternative approach to characterize microbial communities that, in contrast to targeted amplicon methods, uses the entire nucleic acid content of a microbial sample and produces relative abundance information for all genes, functions and microorganisms. Here, nucleic acids are again

extracted from the sample, but are sheared into small fragments that are independently sequenced. The first shotgun metagenomics approaches to characterize microbial communities used cloned libraries to facilitate DNA sequencing using automated Sanger sequencing instruments.[10,11] However, advances in NGS technologies mean that the cloning step is no longer necessary and greater yields of sequencing data can be obtained without this cloning bias-sensitive, labour-intensive and costly step.

Since shotgun metagenomics is PCR-independent and, therefore, not biased by primers designed on the basis of expectations of sequence conservation, this method is able to detect microorganisms which may not be detected using targeted amplicon-based NGS methods. For example, Brown and colleagues described a notable subset of bacterial taxa – known as candidate phyla radiation (CPR) bacteria – that could evade detection by 16S rRNA gene NGS methods due to self-splicing introns and proteins encoded within their rRNA genes, both because they occur in regions targeted by PCR primers and because they increase the length of the target sequence.[51] Of note, four members of the *Thiotrichaceae* family are the only other bacteria outside the CPR known to have self-splicing introns within their 16S rRNA genes, illustrating their rarity in bacteria.[52] In addition, there are no broad-range genetic markers for viruses as mentioned before. For that reason, shotgun metagenomics has revolutionized the field of virology with comprehensive applications that includes viral detection and virus discovery in clinical and environmental samples.[53,54] In fact, the genomes of DNA viruses can be recovered through shotgun metagenomics of DNA that was directly extracted from a sample, whereas extracted RNA has to be converted to complementary DNA (cDNA) first in order to detect RNA viruses.[55]

Obtaining genome sequences using shotgun metagenomics improves the researchers' ability to discriminate microorganisms on a species-level, or even strain-level, taxonomically. This is in contrast to 16S rRNA gene NGS methods that offer often limited resolution at lower taxonomic levels due to the high sequence conservation at these taxonomic levels of the amplicons produced.[36] The identification of microbial strains is of particular importance during epidemic outbreaks caused by microorganisms, where rapid and accurate pathogen identification and characterization is essential for the management of individual cases and of an entire outbreak. For example, the genome sequence of the outbreak strain of Shiga-toxigenic *Escherichia coli* (STEC) 0104:H4, which caused over 50 deaths in Germany in 2011, was reconstructed early in the outbreak using a culture-dependent whole-genome sequencing method.[56] As a result, rapid PCR screening tests were quickly developed using the available genome sequence,[57,58] which aided in tracing back the source of the outbreak to fenugreek seeds from Egypt.[59] Importantly, two years later, researchers were able to reconstruct the genome sequence of this outbreak strain using shotgun metagenomics directly on faecal samples that were collected from subjects during the outbreak.[60] This result highlights the potential

of shotgun metagenomics to identify and characterize pathogens directly from (clinical) samples and supports its future prospective use during outbreaks of life threatening infections caused by unknown pathogens.

Finally, shotgun metagenomics provides access to the functional gene composition of microbial communities and thus gives a much broader description of microbial community genetics than single gene phylogenetic surveys. In general, functional annotation involves two steps, namely gene prediction and gene annotation. During the gene prediction step, various bioinformatics algorithms are used to determine which sequences may (partially) encode proteins. Once identified, protein coding sequences are compared to a database of protein families and functionally annotated with the matching family's function.[61] This information can then be used to discover new genes and to formulate functional pathways.[62] Importantly, since shotgun metagenomics generally targets genomic DNA, it cannot distinguish whether the predicted genes are actually expressed under particular conditions. The measurement of gene expression can be achieved by using metatranscriptomics approaches,[63] which are beyond the scope of this chapter.

Experimental pitfalls and biases

Regardless of the types of microorganisms targeted, or the methodology used to characterize them, choices made at every step – from sample handling to data analysis – can have a serious impact on biasing the final results obtained. The effects of bias can lead to the discovery of spurious correlations and to missing true correlations. Therefore, it is recommended that technicians and researchers use synthetic microbial community (SMC) mixes (also known as mock samples), containing multiple fully-characterized microbial species, in order to calibrate their chosen protocols and identify biases introduced by their techniques.[64] In the following section, the focus is primarily directed towards the potential biases created for protocols utilizing 16S rRNA gene NGS methods, which are shown in Figure 3. This is because 16S rRNA gene NGS methods are more rapid, less complicated and cheaper compared to techniques such as shotgun metagenomics and therefore more likely to be implemented in routine (clinical) microbiological diagnostic laboratories within a shorter timeframe.

Sample handling. The choice of the most optimal sampling protocol depends on the sample type to be investigated. However, they all have in common that samples are transported to the laboratory and stored for a certain period of time before these samples are processed. The transport and storage conditions of biological samples are important factors that can impact DNA yield and DNA quality prior to microbiota investigations. Therefore, several studies have evaluated how different storage and transit conditions may affect the stability of the microbial composition. For example, Carroll et

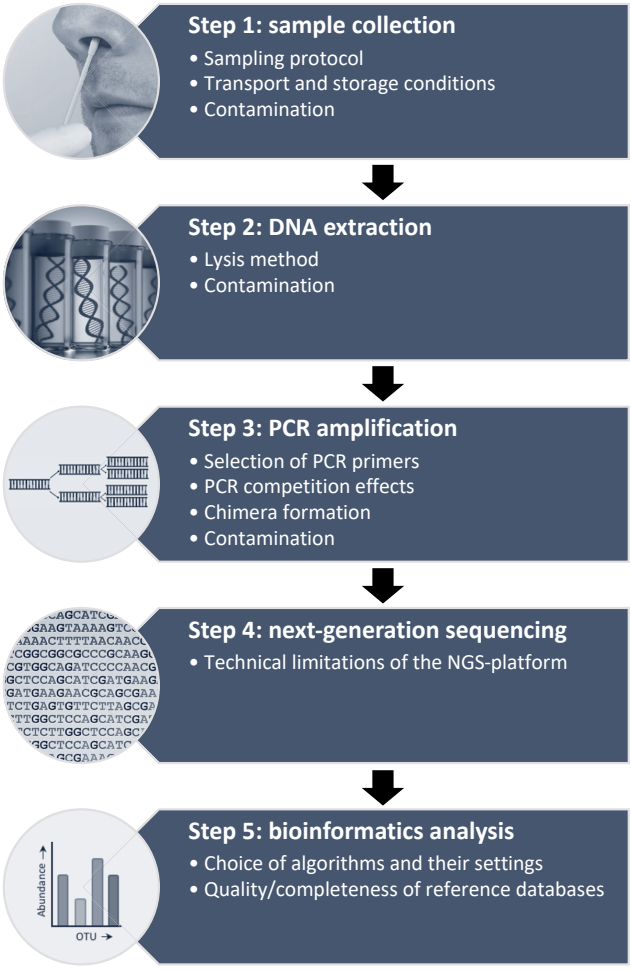

Figure 3. Schematic overview of the workflow for 16S rRNA gene-based analysis of microbial communities, showing the potential biases created for each step of the process.

al. demonstrated microbial stability of faecal samples over a 24-hours period at room temperature and 6 months of long-term storage at -80°C.[65] Others have shown that storage of faecal samples for three days at room temperature did not affect total DNA purity and relative 16S rRNA gene contents,[66] but that DNA became fragmented when samples were inconsistently freeze thawed or when samples had been kept for over 2 weeks at room temperature.[67] Interestingly, a recent study by Shaw et al. illustrated that faecal samples stored for more than 2 years at -80°C are still largely representative of the original microbial community composition.[68] Although these studies show that the effects of storage and transit conditions on microbial diversity and structure are surprisingly small

for faecal samples, the most widely accepted protocols for optimal preservation involves immediate freezing followed by long-term storage at -80°C.[69]

DNA extraction. All DNA-based methods, including 16S rRNA gene NGS methods, rely on the effective lysis of microorganisms to liberate genomic material for downstream analysis. In order to achieve effective lysis, several procedures have been developed, including the chemical or mechanical disruption of cells, lysis using detergents, or a combination of these approaches. However, some cell types may resist common mechanical or chemical lysis methods that may result in important differences in the performance of commercially available DNA extraction kits.[70,71] For example, some methods have been previously shown to yield in a reduced recovery of Gram-positive microorganisms compared to Gram-negative microorganisms (presumably due to differences in the composition of the respective microbial cell envelopes),[72] and an effective cell lysis becomes even more problematic for microorganisms whose cell envelope contains the difficult to lyse component mycolic acid, such as in mycobacteria.[73] Essentially, the choice of the most optimal DNA extraction method is greatly dependent on the sample type and target microbial species to be investigated, but should be employed consistently within a microbiota study.

Contaminating DNA. The validity of microbiota results is threatened by the presence of contaminating DNA derived from the (laboratory) environment and/or the reagents/consumables used during sample processing. For example, PCRs may yield billions of amplicons, which combined with the extreme sensitivity of PCR amplification, means that there is a high risk of amplicon contamination within research and diagnostic laboratories that regularly use PCR. For this reason, many laboratories spatially separate pre- and post-PCR steps in order to limit the risk of amplicon cross-contamination between distinct PCR experiments. Additionally, Glassing et al. showed that commercially available DNA extraction and PCR amplification kits may generate up to 20,000 16S rRNA gene sequences, representing more than 80 prokaryotic genera, even without the addition of any sample.[74] These contamination issues are particularly important for the accurate analysis of the microbial composition of low biomass samples. Salter et al. clearly illustrated how contaminating DNA can affect the microbiota results obtained.[75] These researchers sequenced a pure culture of the bacterium *Salmonella bongori* as well as a series of diluted versions and showed that DNA contamination increased with each dilution and quickly drowned out the original *S. bongori* signal. Therefore, in order to minimize the chance of erroneous conclusions derived from microbiota surveys, it is essential that negative extraction controls (specifically, template-free 'blanks' processed with the same DNA extraction and PCR amplification kits as the actual samples) be

included in 16S rRNA gene NGS protocols in order to allow for the identification of amplicon sequences that originate from DNA contamination.

Selection of 16S rRNA gene PCR primers. Universal 16S rRNA gene PCR primer sets are designed to amplify as many different 16S rRNA gene sequences from as wide a range of prokaryotic species as possible. However, it is well-known that there are no suitable 100% conserved regions of the 16S rRNA gene available for PCR amplification, which can lead to inaccurate microbiota profiles due to inefficient PCR primer binding. In order to ensure the detection of the specific microbial taxa of interest in a particular study, several researchers have reported on the adaptation of universally applicable 16S rRNA gene PCR primer sets via the introduction of degenerate base pairs at the positions of 16S rRNA gene/primer sequence mismatches.[76,77] In addition, the multiple hypervariable regions of each 16S rRNA gene exhibit different degrees of sequence diversity resulting in an ongoing debate about the most efficient hypervariable regions to be used for accurate phylogenetic analysis and taxonomic classification.[78,79] However, the choice for a particular hypervariable region also depends on the technological limitations of the NGS-platforms used. For example, the short length of the 16S rRNA gene V4 region (~250 bp) allows for a full overlap of DNA sequences that are obtained from both ends of the PCR amplicon using Illumina's MiSeq NGS-platform, which is currently the most commonly used NGS-platform. This strategy generates the lowest error rates, which have resulted in more accurate taxonomic classifications, compared to the results obtained from the not completely overlapping V3-V4 and V4-V5 regions.[29] Indeed, the amplification and sequencing of multiple hypervariable regions,[64] or even the generation of (near) full-length 16S rRNA gene sequences using upcoming third generation sequencing platforms,[80,81] give the most complete description of microbiota profiles within a microbial sample.

PCR competition effects. Although often neglected in 16S rRNA gene NGS studies, PCR is a competitive process meaning that the presence of multiple 16S rRNA gene template molecules in a single reaction tube may lead to the preferential PCR amplification of a subset of 16S rRNA gene targets that amplify more efficiently compared to other 16S rRNA gene targets.[82] These differences in template DNA amplification efficiencies may lead to inaccurate microbiota profiling results. There are several mechanisms (relating to the differences in 16S rRNA gene target sequence composition) that could lead to such preferential PCR amplification, including primer binding capacity, sequence length, and GC-content.[82,83] However, compensating for these different amplification efficiencies requires optimized PCR conditions that guarantee equal amplification efficiency for each individual 16S rRNA gene target, which is practically impossible when investigating polymicrobial samples of unknown composition. An extra complication based on our

own experience investigating clinical samples (Chapter 4, this thesis), is that PCR amplification efficiencies of 16S rRNA gene template molecules may be reduced in samples that contain high levels of human DNA and low levels of prokaryotic DNA, probably via the formation of competing non-specific amplicons. Thus, although NGS is a very sensitive detection platform, differences in PCR amplification efficiency of 16S rRNA gene targets within a polymicrobial sample may lead to a biased (and even false) outcome of the original sample composition. Therefore, methodological steps should be taken to try to reduce the effect of PCR amplification efficiency bias.

Chimera formation. 16S rRNA gene PCRs will generate chimeric amplification products (whereby a single DNA amplicon comprises sequences that originate from multiple different 16S rRNA genes), which may be falsely interpreted as a novel microorganism or an existing but absent microorganism, thus inflating the apparent sample richness (i.e. the number of microbial taxa present within a sample). The most commonly described mechanism of chimera formation involves prematurely terminated PCR products that can serve as PCR primers to amplify related template DNA molecules on subsequent PCR cycles.[84] In addition, chimera formation might also occur due to template-switching events during DNA synthesis,[85] or via the incorporation of random DNA fragments, such as shortened PCR primers and degraded amplicons that might be produced by proofreading enzymes during PCR amplification.[86] Importantly, chimeras are frequent artefacts in 16S rRNA gene NGS studies and have been detected at a frequency of up to 30%, although the frequency of chimera production decreases, as expected, when template DNA similarity diminishes.[87] In order to reduce the chance of chimera formation, optimized PCR protocols have been proposed that include the use of a highly processive polymerase and a minimized number of PCR cycles,[88] but no method has been shown to eliminate these artefacts entirely. In addition, numerous computational approaches have been developed over the years to detect and remove chimeric sequences from 16S rRNA gene NGS datasets,[84,89-91] but these different methods often disagree with one another.[84,92] Thus, chimeras continue to be of a major cause of concern to researchers performing 16S rRNA gene NGS research, and even more disturbing, public 16S rRNA gene reference databases are already suspected of containing a significant number of chimeric sequences that further complicate the reliable taxonomic classifications obtained from 16S rRNA gene NGS experiments.[90] Optimized methodologies need to be developed that reduce the generation of chimeric amplification products without relying on bioinformatics-based chimera identification and filtering steps.

Bioinformatics analysis. The analysis of 16S rRNA gene NGS data requires an extensive array of bioinformatics algorithms that are involved in computational intensive steps such as quality filtering, operational taxonomic units (OTU) clustering, and sequence

classification. Currently, there are many different bioinformatics algorithms available for this purpose, which makes it difficult for non-bioinformatics educated scientists to identify the most accurate approaches for 16S rRNA gene NGS analysis. Importantly however, multiple studies have shown that the choice of certain bioinformatics algorithms and their settings can affect the final microbiota results obtained.[93,94] For this reason, popular open-source programs, such as mothur and QIIME, have aided in these issues through rewriting specific bioinformatics algorithms (e.g. mothur) or combining original published bioinformatics algorithms (e.g. QIIME) into single optimized software packages.[95-96] These programs have excellent online tutorials and forums to further support the (inexperienced) user, but their use remains complex as both programs have implemented a collection of command-line tools that represent a large number of bioinformatics algorithms and settings. Therefore, there remains a strong need for 'easy-to-use' bioinformatics pipelines that can be operated by non-bioinformatics educated users, including most employees in routine (clinical) microbiological diagnostic laboratories.

In summary, the experimental pitfalls and biases that are described in this chapter frustrate the standardization of the many 16S rRNA gene NGS protocols currently published. Standardization of methods is arguably best-practice to ensure quality, as well as a necessity to compare results obtained in different laboratories. Although the urgent need for standardized 16S rRNA gene NGS protocols has been recognized in recent years,[97] improvements in reproducibility and accuracy are still required before these methods can make the transition from research tool to diagnostic applications.

AIM AND OUTLINE OF THE THESIS

The overall aim of this thesis was to develop and validate an accurate and standardized 16S rRNA gene NGS platform for use in the routine (clinical) microbiological diagnostic laboratory. For this, several issues relating to the previously described experimental pitfalls and biases of current 16S rRNA gene NGS protocols needed to be overcome. These include inevitable PCR amplification biases, such as chimera formation and PCR competition, and the introduction of contaminating DNA derived from the laboratory environment and reagents used in the experimental set-up. In addition, analysis of 16S rRNA gene NGS data requires a combination of bioinformatics skills and computational resources that is nowadays mostly absent in routine (clinical) microbiological diagnostic laboratories. In this respect, special emphasis has been placed on: i) the development of a novel PCR amplification protocol to reduce chimera formation and PCR competition biases, ii) the development of a protocol to remove DNA contamination from 16S rRNA gene NGS results, and iii) the establishment of an 'easy-to-use' and fully automated bioinformatics pipeline for 16S rRNA gene NGS data analysis (in conjunction with colleagues from the Department of Bioinformatics at the Erasmus MC).

Chapter 1 contains a general introduction and short outline of the thesis. This introduction particularly focusses on the experimental pitfalls and biases associated with current microbiota profiling research that could for example easily result in erroneous conclusions of associations between the human microbiota and disease. In **Chapter 2**, we described a 'Ten-E' protocol that can be used by scientists and clinicians to quickly and critically evaluate claims derived from microbiota-based research. The subsequent six chapters are then divided into two themes relating to the development (**Chapters 3**, **4**, **5**, **6**) and the evaluation (**Chapters 7** and **8**) of the newly developed 16S rRNA gene NGS platform referred to as 'MYcrobiota'.

Current 16S rRNA gene NGS methods involve PCR amplification protocols that simultaneously amplify multiple 16S rRNA gene template molecules in a single reaction tube. Such multi-template PCRs are known to generate chimeric amplicons and can also be affected by PCR competition effects, thereby reducing the accuracy of 16S rRNA gene NGS results. To overcome these limitations, we first developed a novel micelle-based PCR (micPCR) amplification strategy, which is described in **Chapter 3**. In **Chapter 4**, we described the addition of an internal calibrator (IC) to our micPCR protocol, which allows for the standardization of microbiota profiling results, thereby generating quantitative microbiota profiles and facilitating the subtraction of contaminating DNA. In order to develop a comprehensive 16S rRNA gene profiling platform, we provided access to complex command-line bioinformatics tools via the 'Galaxy mothur Toolset (GmT)', which allows non-bioinformatics educated users to build and apply bioinformatics pipelines for 16S rRNA gene NGS data analysis through an 'easy-to-use' Galaxy web interface (**Chap-**

ter 5). **Chapter 6** describes how a dedicated GmT-based bioinformatics pipeline was coupled to our specific micPCR use-case in order to generate an 'end-to-end' microbiota diagnostic analysis service. The resulting platform (MYcrobiota) was evaluated for use in the field of routine clinical microbiological diagnostics by processing a range of clinical samples and comparing the results obtained using MYcrobiota to those obtained using routine culture-based methods.

In **Chapter 7**, the performance of MYcrobiota to detect bacterial DNA in clinical samples was further evaluated. In this respect, we analysed low biomass joint fluids obtained from patients suspected of bacterial septic arthritis and compared the results from MYcrobiota to routine cultures. Additionally, in **Chapter 8**, the universal applicability of MYcrobiota was assessed by employing the methodology as a microbial monitoring tool in the field of drinking water management. Here, the microbial dynamics was investigated within an operational drinking water distribution system using MYcrobiota and conventional techniques (including heterotrophic plate counts, adenosine triphosphate measurements and flow cytometry) as comparator.

Chapter 9 summarizes and discusses the research presented in the thesis, as well as the future perspectives of MYcrobiota and microbiota profiling per se for use in the routine (clinical) microbiological diagnostic laboratory. Finally, the main findings of the thesis are summarized in Dutch in **Chapter 10**.

REFERENCES

1. Pedrós-Alió C. Genomics and marine microbial ecology. *Int Microbiol* 2006; **9**: 191-197.
2. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012; **13**: 260-270.
3. Be, NA, Avila-Herrera A, Allen JE, et al. Whole metagenome profiles of particulates collected from the International Space Station. *Microbiome* 2017; **5**: 81.
4. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486**: 207-214.
5. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome* 2015; **3**: 31.
6. Lagier JC, Hugon P, Khelaifia S, et al. The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clin Microbiol Rev* 2015; **28**: 237-264.
7. Mullis KB. The unusual origin of the polymerase chain reaction. *Sci Am* 1990; **262**: 56-65.
8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; **74**: 5463-5467.
9. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016; **17**: 333-351.
10. Venter JC, Remington K, Heidelberg JF, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004; **304**: 66-74.
11. Gill SR, Pop M, DeBoy RT, et al. Metagenomic analysis of the human distal gut microbiome. *Science* 2006; **312**: 1355-1359.
12. Arrieta MC, Stiemsma LT, Dimitriu PA, et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci Transl Med* 2015; **7**: 307ra152.
13. Abrahamsson TR, Jakobsson HE, Andersson AF, et al. Low gut microbiota diversity in early infancy precedes asthma at school age. *Clin Exp Allergy* 2014; **44**: 842-850.
14. West CE, Rydén P, Lundin D, et al. Gut microbiome and innate immune response patterns in IgE-associated eczema. *Clin Exp Allergy* 2015; **45**: 1419-1429.
15. Frank DN, St Amand AL, Feldman RA, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* 2007; **104**: 13780-13785.
16. Fujimoto T, Imaeda H, Takahashi K, et al. Decreased abundance of *Faecalibacterium prausnitzii* in the gut microbiota of Crohn's disease. *J Gastroenterol Hepatol* 2013; **28**: 613-619.
17. Barlow GM, Yu A, Mathur R. Role of the gut microbiome in obesity and diabetes mellitus. *Nutr Clin Pract* 2015; **30**: 787-797.
18. Komaroff AL. The microbiome and risk for obesity and diabetes. *JAMA* 2017; **317**: 355-356.
19. Foster JA, McVey Neufeld KA. Gut-brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci* 2013; **36**: 305-312.
20. Rupnik M, Wilcox MH, Gerding DN. *Clostridium difficile* infection: new developments in epidemiology and pathogenesis. *Nat Rev Microbiol* 2009; **7**: 526-536.
21. van Nood E, Vrieze A, Nieuwdorp M, et al. Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *N Engl J Med* 2013; **368**: 407-415.

22. Cui B, Feng Q, Wang H, et al. Fecal microbiota transplantation through mid-gut for refractory Crohn's disease: safety, feasibility, and efficacy trial results. *J Gastroenterol Hepatol* 2015; **30**: 51-58.
23. He Z, Li P, Zhu J, et al. Multiple fresh fecal microbiota transplants induces and maintains clinical remission in Crohn's disease complicated with inflammatory mass. *Sci Rep* 2017; **7**: 4753.
24. Relman DA, Loutit JS, Schmidt TM, et al. The agent of bacillary angiomatosis. An approach to the identification of uncultured pathogens. *N Engl J Med* 1990; **323**: 1573-1580.
25. Wilson KH, Blitchington R, Frothingham R, et al. Phylogeny of the Whipple's-disease-associated bacterium. *Lancet* 1991; **338**: 474-475.
26. Cummings LA, Kurosawa K, Hoogestraat DR, et al. Clinical next generation sequencing outperforms standard microbiological culture for characterizing polymicrobial samples. *Clin Chem* 2016; **62**: 1465-1473.
27. Rhoads DD, Cox SB, Rees EJ, et al. Clinical identification of bacteria in human chronic wound infections: culturing vs. 16S ribosomal DNA sequencing. *BMC Infect Dis* 2012; **12**: 321.
28. Brook I. Clinical review: bacteremia caused by anaerobic bacteria in children. *Crit Care* 2002; **6**: 205-211.
29. Kozich JJ, Westcott SL, Baxter NT, et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 2013; **79**: 5112-5120.
30. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 1977; **74**: 5088-5090.
31. Van de Peer Y, Chapelle S, De Wachter R. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res* 1996; **24**: 3381-3391.
32. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007; **35**: 7188-7196.
33. Cole JR, Chai B, Farris RJ, et al. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* 2005; **33**: D294-6.
34. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006; **72**: 5069-5072.
35. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res* 2012; **40**: D136-43.
36. Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 2007; **10**: 504-509.
37. Blackwood KS, Turenne CY, Harmsen D, et al. Reassessment of sequence-based targets for identification of *bacillus* species. *J Clin Microbiol* 2004; **42**: 1626-1630.
38. Mollet C, Drancourt M, Raoult D. *rpoB* sequence analysis as a novel basis for bacterial identification. *Mol Microbiol* 1997; **26**: 1005-1011.
39. Drancourt M, Raoult D. *rpoB* gene sequence-based identification of *Staphylococcus* species. *J Clin Microbiol* 2002; **40**: 1333-1338.
40. Adekambi T, Drancourt M, Raoult D. The *rpoB* gene as a tool for clinical microbiologists. *Trends Microbiol* 2009; **17**: 37-45.

41. Dahllof I, Baillie H, Kjelleberg S. *rpoB*-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl Environ Microbiol* 2000; **66**: 3376-3380.
42. Lan Y, Rosen G, Hershberg R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* 2016; **4**: 18.
43. Jones MD, Forn I, Gadelha C, et al. Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* 2011; **474**: 200-203.
44. López-Garcia P, Rodriguez-Valera F, Pedrós-Alió C, et al. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 2001; **409**: 603-607.
45. Schoch CL, Seifert K, Huhndorf S, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proc Natl Acad Sci USA* 2012; **109**: 6241-6246.
46. De Filippis F, Laiola M, Blaiotta G, et al. Different amplicon targets for sequencing-based studies of fungal diversity. *Appl Environ Microbiol* 2017; **83**: e00905-17.
47. Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol* 2005; **3**: 504-510.
48. Gardner SN, Jaing CJ, McLoughlin KS, et al. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* 2010; **11**: 668.
49. Wang D, Coscoy L, Zylberberg M, et al. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 2002; **99**: 15687-156892.
50. Palacios G, Quan PL, Jabado OJ, et al. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis* 2007; **13**: 73-81.
51. Brown CT, Hug LA, Thomas BC, et al. Unusual biology across a group comprising more than 15% of domain *Bacteria*. *Nature* 2015; **523**: 208-211.
52. Salman V, Amann R, Shub DA, et al. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc Natl Acad Sci USA* 2012; **109**: 4203-4208.
53. Capobianchi MR, Giombini E, Rozera G. Next-generation sequencing technology in clinical virology. *Clin Microbiol Infect* 2013; **19**: 15-22.
54. Smits SL, Osterhaus AD. Virus discovery: one step beyond. *Curr Opin Virol* 2013; **3**: e1-e6.
55. Batty EM, Wong THN, Trebes A, et al. A modified RNA-Seq approach for whole genome sequencing of RNA viruses from faecal and blood samples. *PLoS One* 2013; **8**: e66129.
56. Mellmann A, Harmsen D, Cummings CA, et al. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 2011; **6**: e22751.
57. Bielaszewska M, Mellman A, Zhang W, et al. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect Dis* 2011; **11**: 671-676.
58. Qin J, Cui Y, Zhao X, et al. Identification of the Shiga toxin-producing *Escherichia coli* O104:H4 strain responsible for a food poisoning outbreak in Germany by PCR. *J Clin Microbiol* 2011; **49**: 3439-3440.
59. King LA, Nogareda F, Weill FX, et al. Outbreak of Shiga toxin-producing *Escherichia coli* O104:H4 associated with organic fenugreek sprouts, France, June 2011. *Clin Infect Dis* 2012; **54**: 1588-1594.

60. Loman NJ, Constantinidou C, Christner M, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 2013; **309**: 1502-1510.
61. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci* 2014; **5**: 209.
62. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010; **464**: 59-65.
63. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. *Bioinform Biol Insights* 2016; **10**: 19-25.
64. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PloS one* 2012; **7**: e39315.
65. Carroll IM, Ringel-Kulka T, Siddle JP, et al. Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. *PLoS One* 2012; **7**: e46953.
66. Dominianni C, Wu J, Hayes RB, et al. Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiol* 2014; **14**: 103.
67. Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol* 2012; **12**: 158.
68. Shaw AG, Sim K, Powell E, et al. Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room? *Microbiome* 2016; **4**: 40.
69. Goodrich JK, Di Rienzi SC, Poole AC, et al. Conducting a microbiome study. *Cell* 2014; **158**: 250-262.
70. Kennedy NA, Walker AW, Berry SH, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* 2014; **9**: e88982.
71. Wu GD, Lewis JD, Hoffmann C, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol* 2010; **10**: 206.
72. Hendolin PH, Paulin L, Ylikoski J. Clinically applicable multiplex PCR for four middle ear pathogens. *J Clin Microbiol* 2000; **38**: 125-32.
73. Vandeventer PE, Weigel KM, Salazar J, et al. Mechanical disruption of lysis-resistant bacterial cells by use of a miniature, low-power, disposable device. *J Clin Microbiol* 2011; **49**: 2533-2539.
74. Glassing A, Dowd SE, Galandiuk S, et al. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog* 2016; **8**: 24.
75. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12**: 87.
76. Sim K, Cox MJ, Wopereis H, et al. Improved detection of bifidobacteria with optimised 16S rRNA-gene based pyrosequencing. *PLoS One* 2012; **7**: e32543.
77. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 2016; **18**: 1403-1414.

78. Chakravorty S, Helb D, Burday M, et al. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* 2007; **69**: 330-339.

79. Yang B, Wang Y, Qian PY. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 2016; **17**: 135.

80. Benitez-Paez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer. *Gigascience* 2016; **5**: 4.

81. Schloss PD, Jenior M, Koumpouras CC, et al. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* 2016; **4**: e1869.

82. Kalle E, Kubista M, Rensing C. Multi-template polymerase chain reaction. *Biomol Detect Quantif* 2014; **2**: 11-29.

83. Frank JA, Reich CI, Sharma S, et al. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol* 2008; **74**: 2461-2470.

84. Haas BJ, Gevers D, Earl AM, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 2011; **21**: 494-504.

85. Odelberg SJ, Weiss RB, Hata A, et al. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res* 1995; **23**: 2049-2057.

86. Zylstra P, Rothenfluh H, Weiller GF, et al. PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunol Cell Biol* 1998; **76**: 395-405.

87. Wang GC, Wang Y. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* 1996; **142**: 1107-1114.

88. Gohl DM, Vangay P, Garbe J, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* 2016; **34**: 942-949.

89. Edgar RC, Haas BJ, Clemente JC, et al. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; **27**: 2194-2200.

90. Ashelford KE, Chuzhanova NA, Fry JC, et al. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol* 2005; **71**: 7724-7736.

91. Wright ES, Yilmaz LS, Noguera DR. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl Environ Microbiol* 2012; **78**: 717-725.

92. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 2011; **6**: e27310.

93. Kopylova E, Navas-Molina JA, Mercier C, et al. Open-source sequence clustering methods improve the state of the art. *mSystems* 2016; **1**; e00003-15.

94. Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 2015; **3**: e1487.

95. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537-7541.

96. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; **7**: 335-336.

97. Hiergeist A, Reischl U, Priority Program Intestinal Microbiota Consortium/quality assessment participants, et al. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int J Med Microbiol* 2016; **306**: 334-342.