MICROBIOTA ANALYSIS

From reseach tool to diagnostic applications

STEFAN A. BOERS

ISBN: 978-94-6361-161-9
Cover design: Erwin Timmermans, Optima Grafische Communicatie
Layout and printing: Optima Grafische Communicatie, Rotterdam, the Netherlands (www.ogc.nl)

Publication of this thesis was financially supported by the Regional Laboratory of Public Health Kennemerland, Haarlem.

Microbiota Analysis
From research tool to diagnostic applications

Microbiota analyse
van onderzoek naar diagnostische toepassingen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

woensdag 14 november 2018 om 11:30 uur

door

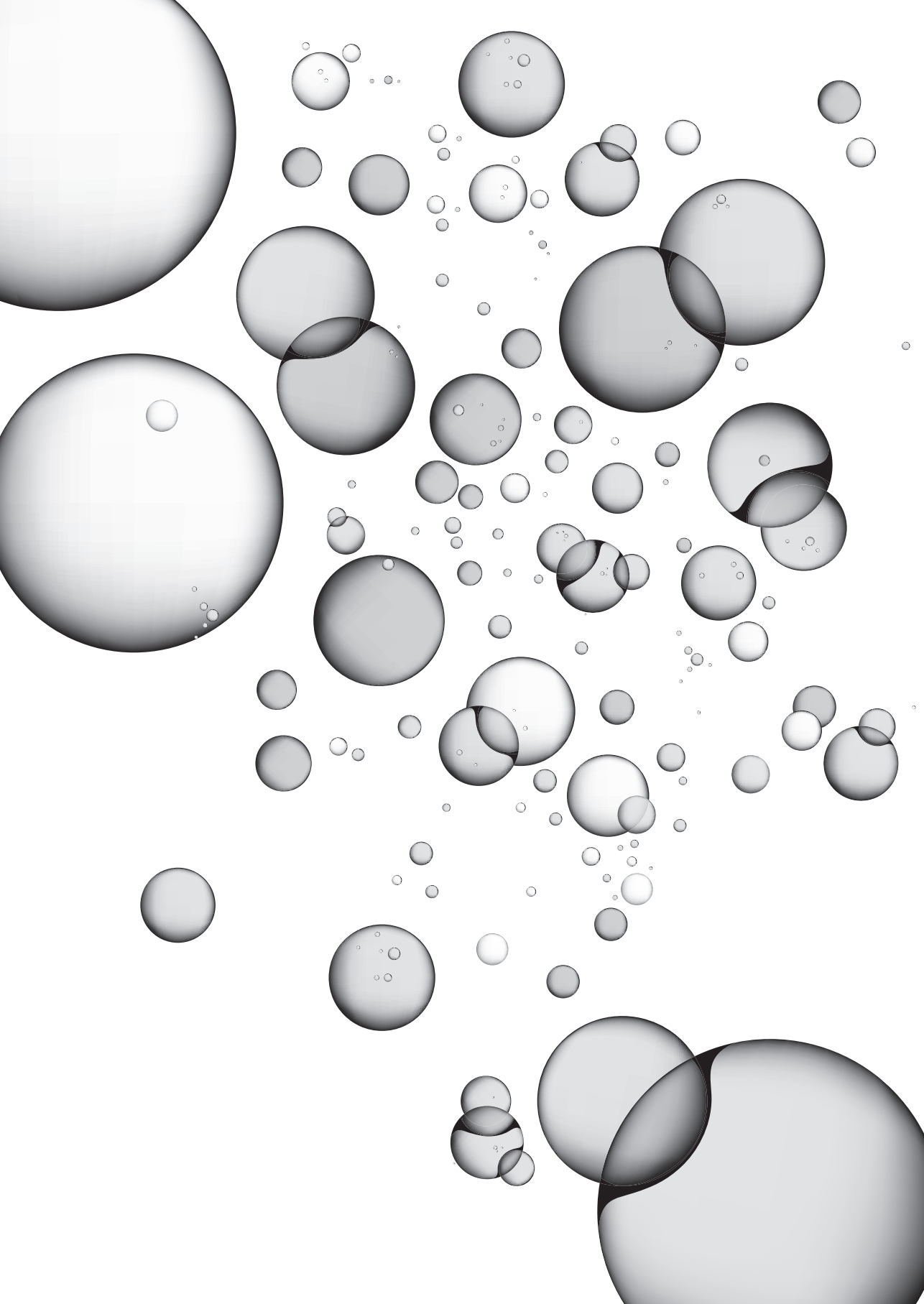
Stefan Alexander Boers
geboren te Oldenzaal

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor	Prof.dr. J.W. Mouton
Overige leden:	Prof.dr. C.A.B. Boucher Prof.dr. P.H.M. Savelkoul Dr. W.J.G. Melchers
Copromotoren:	Dr. J.P. Hays Dr. R. Jansen

TABLE OF CONTENTS



Chapter 1

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

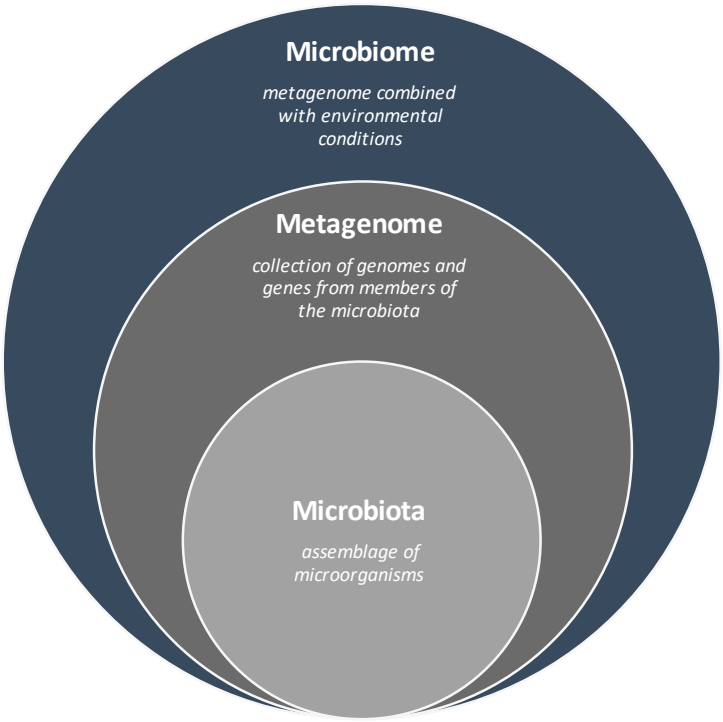


Figure 1. Differentiation of terms used to describe different aspects of research that focus on microbial communities and their environments.

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

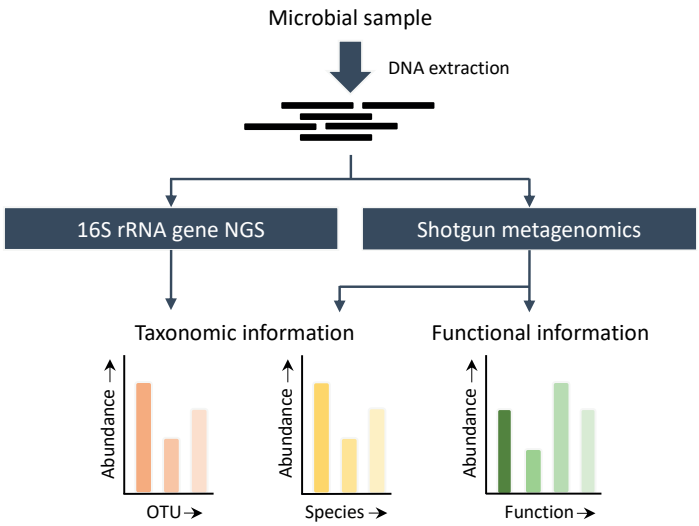


Figure 2. General overview of 16S rRNA gene NGS and shotgun metagenomics methods. Both methods start with the extraction of nucleic acids from a microbial sample. Next, the extracted DNA is either subjected to 16S rRNA gene PCR amplification (16S rRNA gene NGS) or sheared into small DNA fragments (shotgun 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

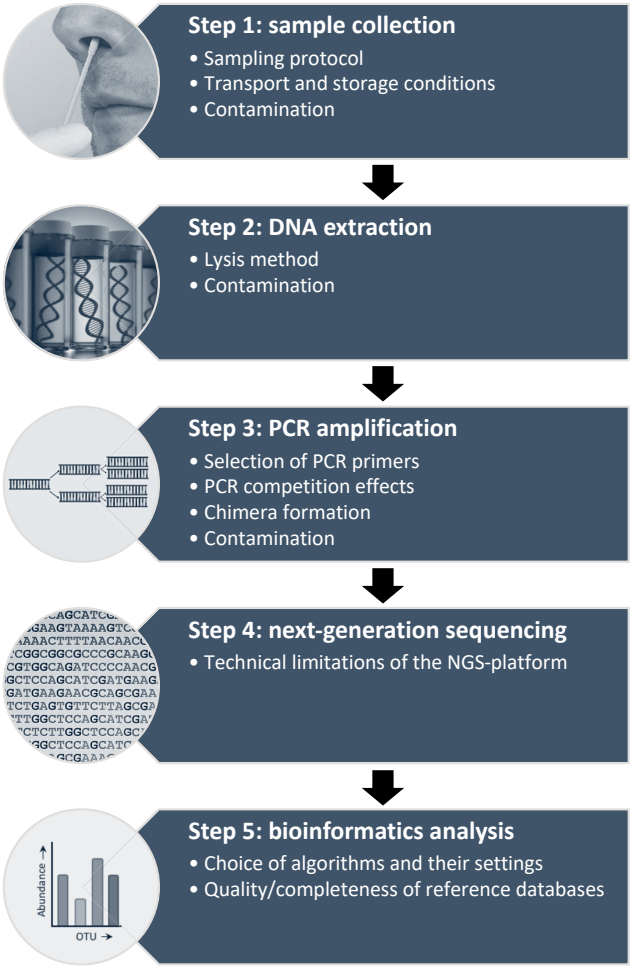


Figure 3. Schematic overview of the workflow for 16S rRNA gene-based analysis of microbial communities, showing the potential biases created for each step of the process.

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.

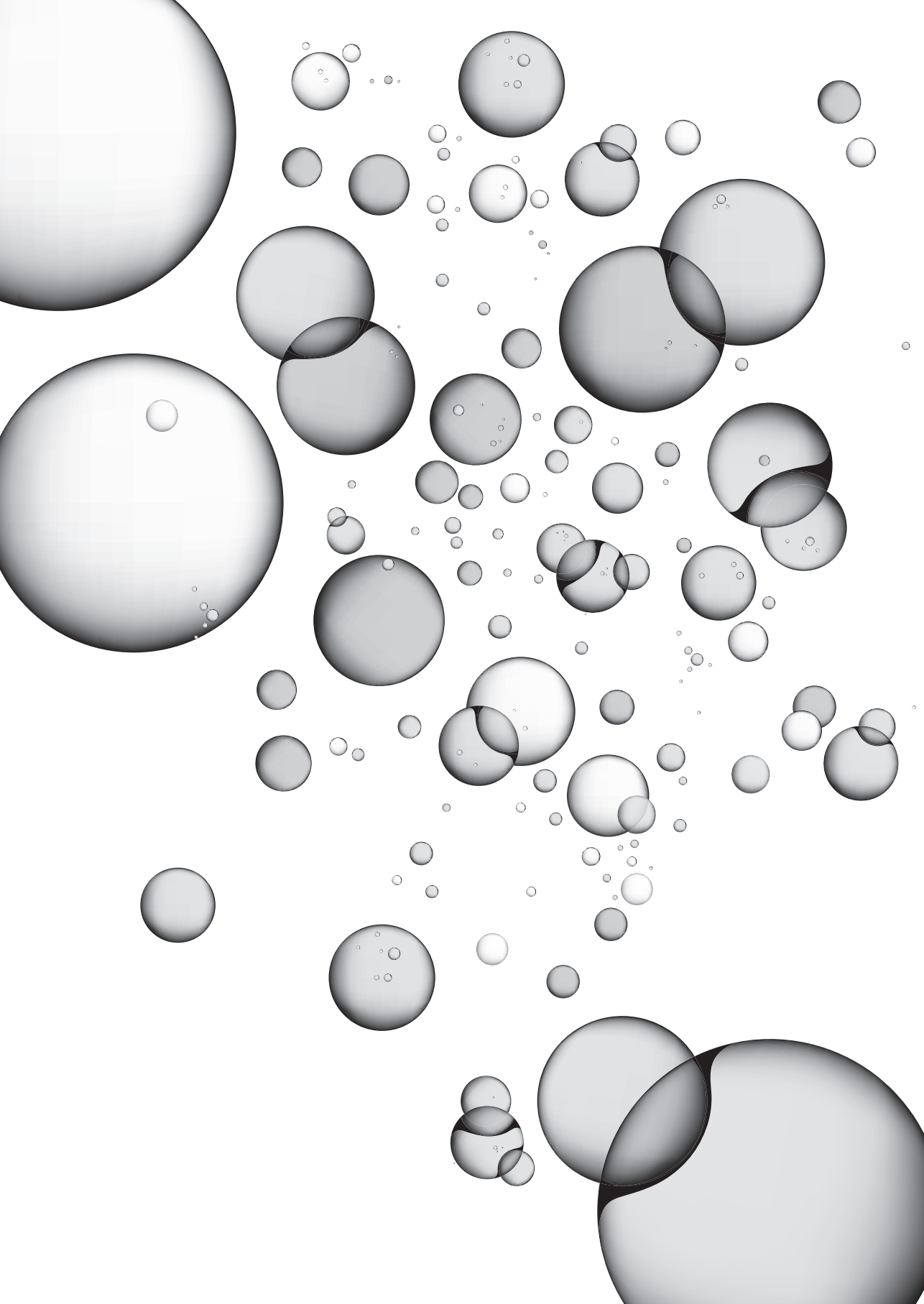
93. Kopylova E, Navas-Molina JA, Mercier C, et al. Open-source sequence clustering methods improve the state of the art. *mSystems* 2016; **1**; e00003-15.

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.



Chapter 2

Suddenly everyone is a microbiota specialist!

Stefan A. Boers
Ruud Jansen
John P. Hays

Clin Microbiol Infect 2016; 22: 581-582.

Recently there has been an explosion in the number of publications linking the human microbiota to various diseases. These microbiota profiles are obtained by either PCR amplification and sequencing of regions of the 16S ribosomal RNA (rRNA) gene of bacteria, or by performing shotgun metagenomics directly on sampled environments. As a simple guide to the critical analysis of microbiota-based publications, the authors present here the 'Ten-E' method. The majority of the described 'Es' can be readily applied to both 16S rRNA gene amplicon sequencing, as well as to shotgun metagenomics-based microbiota profiling studies. As a further note, the authors recommend the adoption of consistent and defined terms within the field of microbiome/microbiota research, as previously published.[1] The ten Es are presented in chronological order of a typical microbiota profiling project, starting with the E of Extraction.

Extraction (E1) – Different DNA extraction methods can seriously impact the final microbiota profiling results. As shown by Kennedy et al., there are significant differences in microbial composition when comparing microbiota profiles obtained from the same specimen using different DNA extraction kits.[2] Therefore caution is necessary when comparing microbiota studies that have used different DNA extraction methodologies.

Environment (E2) – Negative extraction controls should be included and analysed in the experimental protocol for low biomass specimens such as nose swabs, blood or other normally sterile sites. These controls are required to accurately assess the influence of contaminating DNA molecules that may be present in the experimental set-up. These contaminating DNA molecules may already be present in laboratory reagents or commonly used DNA extraction kits. Additionally, contaminating DNA molecules from the laboratory environment may be present on the surface of consumables used during PCR and/or metagenomic microbiota profiling experiments.[3]

Efficiency (E3) – During PCR amplification certain 16S rRNA gene sequences may be amplified more efficiently than others, biasing the resultant microbiota profiles. Amplification efficiency differences are prominent when applying standard PCR protocols but can be overcome by using clonal amplification by micelle PCR. In a micelle PCR, the template DNA molecules are separated into a large number of physically distinct PCR compartments, preventing amplification bias and increasing the accuracy of microbiota profiling methods.[4] Scientists should be aware of the potential for amplification bias during PCRs.

Exaggeration (E4) – Standard 16S rRNA gene PCRs will generate chimeric amplification products, whereby a single DNA amplicon comprises sequences that originate from multiple 16S rRNA genes. Importantly, the inclusion of chimeric sequences that were

not recognized by computational filtering software, leads to incorrect taxonomic identifications and an overestimated microbiota richness in the final microbiota profiling results. These chimeric sequences may be incorrectly identified as new bacterial species. Essentially, the prevention of chimeric sequences will prevent the microbiologist unwittingly becoming a 'bacterial creationist'. One method that can be used to reduce chimera formation is clonal amplification via the use of micelle PCR.[4]

Evaluation (E5) – The evaluation of sequence data by different clustering algorithms may lead to different microbiota results and this fact should be appreciated by scientists.[5] In addition, accurate taxonomic identification of 16S rRNA gene microbiota data depends on the quality and completeness of the reference databases used to identify and classify the sequence data produced, e.g. SILVA, RDP, GreenGenes and NCBI. Since most reference databases contain some unidentified and poorly annotated sequences, and are also inevitably incomplete, manual evaluation of the main sequencing results is to be encouraged. This to ensure that the taxonomic identification of 'key' bacterial genera and species within the microbiota profile are correct.

Elongation (E6) – In general, only short regions of bacterial 16S rRNA genes tend to be sequenced, meaning that these sequences may not have the discriminative power to identify bacteria to the species level. Though some bacterial genera may show sufficient inter-species 16S rRNA gene sequence diversity to allow their accurate identification (e.g. *Akkermansia muciniphila*), other genera may not have sufficient inter-species variation to allow their accurate speciation.[6] Additionally, the naming of species may vary over time.[7] In general, restricting sequence identification to the genus level (when using short 16S rRNA gene sequences), is recommended.

Equality (E7) – 16S rRNA gene sequencing does not generate accurate information regarding the quantification of bacterial species. Different bacterial species carry different numbers of 16S rRNA genes and copy numbers for all bacteria are not known. For example, the *Mycobacterium tuberculosis* genome carries one 16S rRNA gene copy, whereas the *Clostridium beijerinckii* genome carries up to 14 copies of the gene. Therefore, it is recommended that microbiota profiles are expressed as ratios or percentages of '16S rRNA gene copies' rather than ratios of 'species' (which would suggest that bacterial cell or genome copy numbers are being expressed). To provide an accurate number of bacterial genome copies, the use of methods such as calibrated quantitative PCR or digital PCR have to be employed.

Evidence (E8) – Microbiota profiles are generated using bioinformatics approaches and speculations about the clinical importance of the bacterial species usually ignore

Koch's postulates and/or the updated version of Koch's postulates for molecular diagnostics.[8,9] For example, a correlation between an operational taxonomic unit-associated disease, and its corresponding organism, should not be made without first fulfilling Koch's postulates. Currently, many potential disease-associated organisms discovered by microbiota analysis cannot be cultured (although this situation is slowly changing).[10] More effort should be spent on isolating these currently 'non-culturable' organisms before they can be truly associated with a particular disease or condition. Moreover, DNA-based studies do not allow for accurate differentiation between viable, non-viable or dead bacterial cells. This could be important for example, in specimens that have previously been treated with bacteriostatic antibiotics or in environmental samples where 'relic DNA' from dead cells can persist from weeks to years.[11] Therefore, scientists and stakeholders should remain sceptical regarding the scientific claims associated with a microbiota-based article.

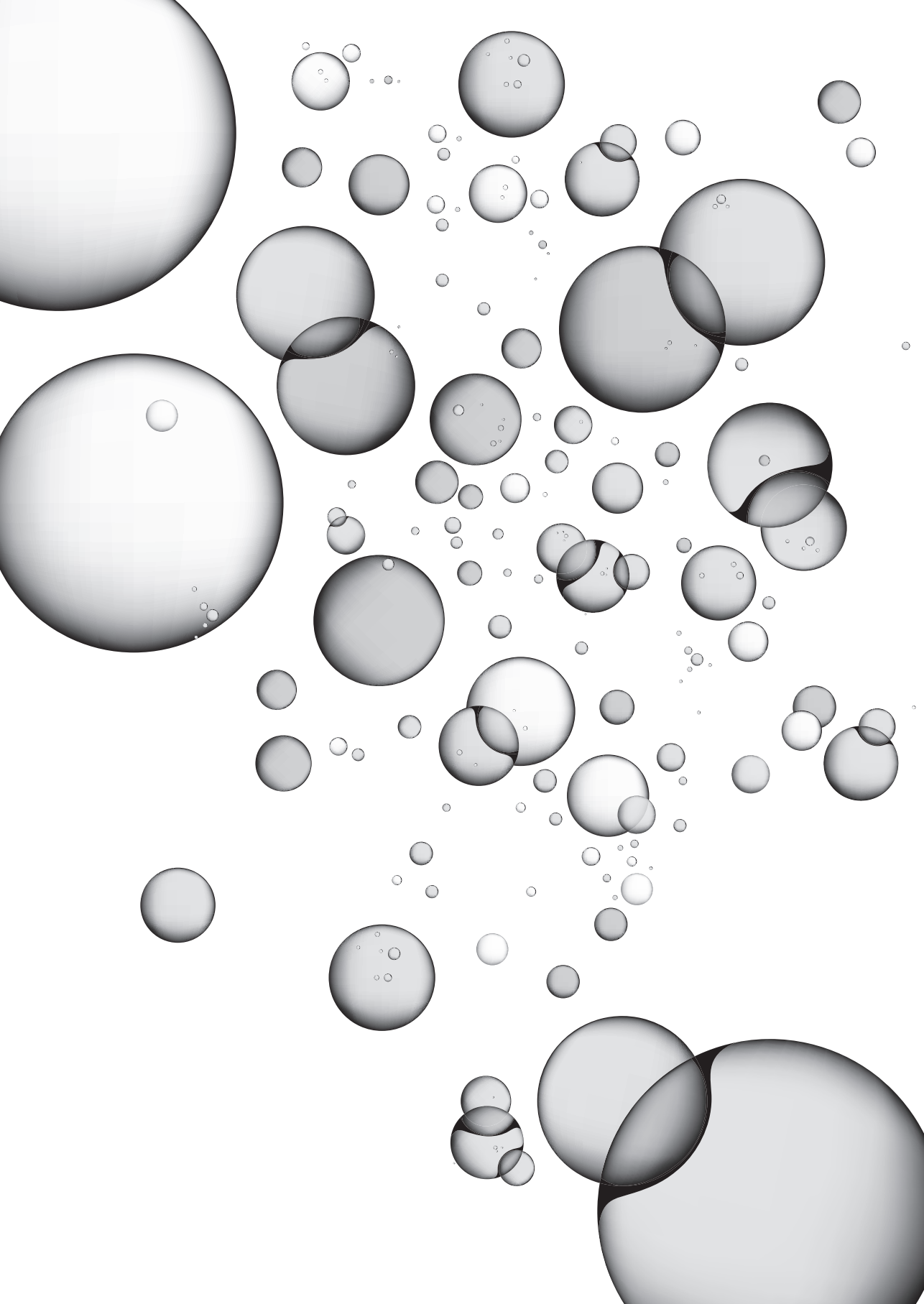
Enrolment (E9) – Microbiota results are often obtained using small cohort-sized studies. However, the microbiota of many ecosystems and environments may be very complex and highly variable, even among similar samples. Many small-scale studies lack the statistical power to test microbiota-based hypotheses to a valid statistical conclusion. This lack of statistical evidence has resulted in a lack of agreement about the microbial composition of many studies published within the scientific literature.[12] Therefore, a larger sample of cohorts and/or meta-cohort analyses should be enrolled when generating conclusions regarding the 'typical' composition of a clinical or environmental sample.

Expectations (E10) – Be aware of possible conflicts of interest between sponsors of microbiota research and the researchers themselves in this highly competitive scientific field. Most journals specifically ask authors to state possible conflicts of interest in their manuscripts. However, readers should still be alert to potential funding biases that may skew published microbiota profiling results.

Finally, the authors hope that the 'Ten-E' protocol published here will aid microbiologists, clinicians, environmentalists, food technologists, journalists and even the general public, to be more critical of the scientific literature when it comes to the reporting of the results of microbiota profiling studies.

REFERENCES

1. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome* 2015; **3**: 31.
2. 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.
3. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12**: 87.
4. Boers SA, Hays JP, Jansen R. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Sci Rep* 2015; **5**: 14181.
5. 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.
6. 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**: 330e9.
7. Collins MD, Lawson PA, Willems A, et al. The phylogeny of the genus Clostridium: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol* 1994; **44**: 812e26.
8. Fredricks DN, Relman DA. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev* 1996; **9**: 18e33.
9. Lipkin WI. Microbe hunting in the 21st century. *Proc Natl Acad Sci USA* 2009; **106**: 6e7.
10. 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**: 237e64.
11. Carini P, Marsden PJ, Leff JW, et al. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol* 2016; **2**: 16242.
12. Gevers D, Kugathasan S, Denson LA, et al. The treatment-naïve microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014; **15**: 382e92.



Chapter 3

Micelle PCR reduces chimera formation in 16S rRNA gene profiling of complex microbial DNA mixtures

Stefan A. Boers
John P. Hays
Ruud Jansen

Sci Rep 2015; 5: 14181.

ABSTRACT

16S rRNA gene profiling has revolutionized the field of microbial ecology. Many researchers in various fields have embraced this technology to investigate bacterial compositions of samples derived from many different ecosystems. However, it is important to acknowledge the current limitations and drawbacks of 16S rRNA gene profiling. Although sample handling, DNA extraction methods and the choice of universal 16S rRNA gene PCR primers are well known factors that could seriously affect the final results of microbiota profiling studies, inevitable amplification artefacts, such as chimera formation and PCR competition, are seldom appreciated. Here we report on a novel micelle-based amplification strategy, which overcomes these limitations via the clonal amplification of targeted DNA molecules. Our results show that micelle PCR drastically reduces chimera formation by a factor of 38 (1.5% vs. 56.9%) compared with traditional PCR, resulting in improved microbial diversity estimates. In addition, compartmentalization during micelle PCR prevents PCR competition due to unequal amplification rates of different 16S rRNA gene template molecules, generating robust and accurate 16S rRNA gene microbiota profiles required for comparative studies (e.g. longitudinal surveys).

INTRODUCTION

Microbiota profiling methods are greatly enhancing our insights into the microbial diversity and taxonomy of many different types of environments and ecosystems, including the relationship between microbiota and host in health and disease.[1] The development of next-generation sequencing (NGS) technologies has highlighted the difficulties of assessing the microbiota using conventional culture methods, as PCR-based NGS of bacterial 16S rRNA genes yields a large diversity of 16S rRNA gene sequences that may be associated with a complex assortment of bacterial taxonomies – from phylum to genus level.[2] Although sequence-based approaches are incredibly powerful, it is important that scientists and bioinformaticians understand and acknowledge the current limitations and drawbacks of NGS technologies and appreciate that the choices made, from study design to DNA extraction and from DNA amplification to data analysis, can have serious impact on the microbiota profiles obtained.[3] For example, Kennedy et al. previously reported significant differences in DNA yield and bacterial DNA composition when comparing DNA extracted from the same faecal sample with different extraction kits.[4] In addition, the use of universal 16S rRNA gene PCR primers has led to inconsistencies in the literature regarding the abundance of the bacteria within similar ecosystems.[5] Essentially, the choice of the most optimal cell lysis procedures, and the most sensitive/specific universal 16S rRNA gene primer pair to be used, are greatly dependent on the sample type and target species to be investigated. Importantly however, even when using the correct choice of cell lysis procedure and 16S rRNA gene primer pair, amplification artefacts (chimeras) are inevitably generated during PCR amplifications due to the presence of multiple PCR targets in a single reaction chamber. Such chimeras are generated independent of the sample type used. Importantly, the formation of these chimeric sequences can lead to erroneous taxonomic identifications and overestimated microbiota richness.[6] Further, although sequences can be filtered out of NGS results using specialized software,[7,8] the generation of chimeric products can still seriously reduce the amount of useful information obtained in a single sequencing run.[9] Importantly, and this is seldom appreciated by users of NGS technologies, PCR is a competitive reaction meaning that the presence of multiple PCR targets in a single amplification reaction may lead to the preferential amplification of a particular subset of 16S rRNA gene copies.[10] The results could then be biased by factors related to the amplification efficiency of particular 16S rRNA gene amplicons rather than the relative abundance of 16S rRNA genes in the test sample. To overcome these sample-independent limitations, we developed and evaluated a micelle-based amplification strategy targeting the 16S rRNA gene that greatly reduces chimera production during PCR amplification and prevents the formation of PCR competition products.

Micelle PCR (micPCR) is designed as a beadless emulsion PCR whereby a single molecule of template DNA is clonally amplified. Template DNA molecules are separated into a large number of physically distinct reaction compartments using water-in-oil emulsions. This compartmentalization per molecule reduces the probability of chimera formation and restrains PCR competition. For example, emulsion based amplification has been successfully applied for aptamer selection to reduce product-product and primer-product hybridizations.[11] Also, emulsion PCRs may be performed in BEAMing experiments, reliable and sensitive assays for the identification and quantification of variations in gene sequences and transcripts.[12] Finally, NGS platforms such as Ion Torrent (Life Technologies) and 454 (Roche) have adopted emulsion-based amplification strategies in their standard NGS workflows to clonally re-amplify DNA sequencing libraries, as their molecular detection methods are not sensitive enough for single molecule sequencing and to prevent mixed sequences.

RESULTS

To evaluate the ability of micPCR to increase the accuracy of 16S rRNA gene sequencing, universal 357F and 926R primers were used to amplify the 16S rRNA gene V3–V5 region from a synthetic microbial community containing equimolar 16S rRNA operon counts derived from 20 different bacterial species (HM-782D supplied by BEI Resources).[13] The protocol utilized a two-step micPCR protocol, as well as a two-step traditional PCR protocol – used for comparative purposes – for NGS library preparation.[14] Importantly, the final number of amplification cycles of a two-step PCR protocol is higher compared to a one-step PCR protocol, resulting in an increased formation of chimeric sequences, making it suitable for evaluating the micPCR.[15] Results of triplicate experiments showed that micPCR/NGS generated only 1.5% (± 1.2%) chimeric sequences in the synthetic community compared to 56.9% (± 1.7%) chimeras using traditional PCR/NGS (Supplementary Table 1). For the micPCR/NGS, the rarefaction analysis rapidly reached horizontal equilibrium at the expected 20 operational taxonomic units (OTUs), indicating a highly reliable calculation of richness (Figure 1). In contrast, the traditional PCR/NGS resulted in 72 OTUs in the synthetic community, with rarefaction analysis showing that the number of OTUs steadily increased as the number of sequence reads increased. It was found that the excess of OTUs consisted of chimeras of the sequences of the 20 bacterial species in the synthetic mix that had not been recognized as such by the mothur software package (https://www.mothur.org/).

Another important factor that influences NGS-related microbiota profiling is competition between different 16S rRNA gene molecules, resulting in unequal/preferential amplification rates for certain amplicon sequences. The result of competition can be an

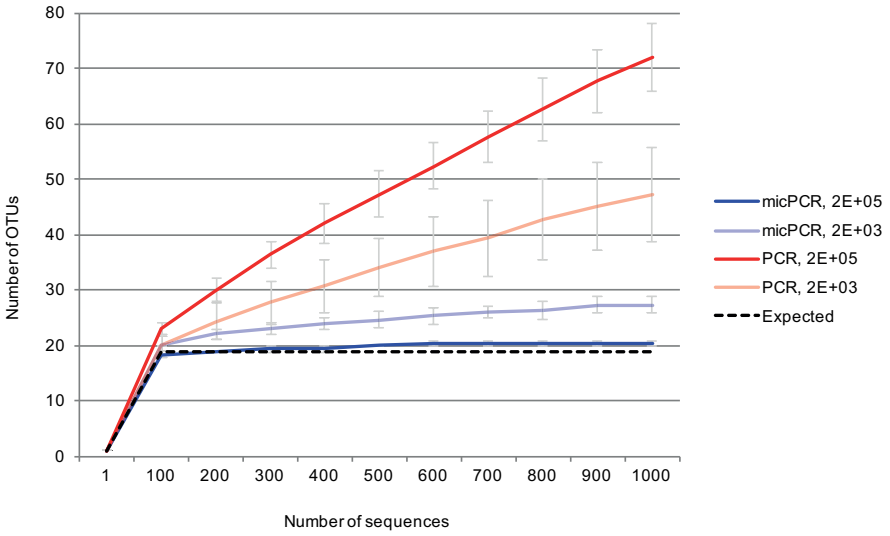


Figure 1. Comparison of rarefaction analyses between micPCR/NGS and traditional PCR/NGS using an equimolar, synthetic microbial community. The number of observed OTUs in the synthetic microbial community is shown as the function of the number of sequences obtained using micPCR/NGS reactions containing 2E+05 (dark blue) and 2E+03 (light blue) input molecules, and traditional PCR/NGS reactions containing 2E+05 (dark red) and 2E+03 (light red) input molecules. Data points represent average values from triplicate experiments and error bars show standard deviations. Rarefaction curves were generated using mothur[19] with an OTU defined at 97% similarity. Analysis was performed on a random 1,000-sequence subset from each sample. *Staphylococcus aureus* and *Staphylococcus epidermidis* present in the synthetic community could not be differentiated at a 97% similarity level, resulting in a maximum of 19 expected OTUs.

over- or underestimation of particular OTUs. For example, in our current experiments we utilized a synthetic community consisting of 20 bacterial species that are each present at an equimolar concentration of 5% of 16S rRNA genes. MicPCR/NGS data showed an average 0.85-fold difference from the 5% OTU frequency expected in the synthetic community, with a maximum overestimation of 1.73-fold for *Listeria monocytogenes* and a maximum underestimation of 0.28-fold for *Streptococcus pneumoniae* (Figure 2). In contrast, the OTU differences associated with PCR competition and traditional PCR/NGS were more extreme, yielding an average 0.65-fold difference in OTU frequency above the expected frequency, with an overestimated maximum of 2.31-fold for *Bacteroides vulgatus* and an underestimated maximum of 0.04-fold for *Helicobacter pylori*. These findings are in agreement with the previously reported consistent overestimation of *Bacteroides* spp. and underestimation of *Helicobacter* spp. in four different laboratories when investigating an identical synthetic community.[13]

In order to determine the usefulness of the micPCR/NGS protocol in determining the microbiota profiles of actual clinical and environmental samples, we evaluated the use of

Bacterial species	micPCR/NGS (input DNA molecules)						PCR/NGS (input DNA molecules)					
	2E+05	2E+05	2E+05	2E+03	2E+03	2E+03	2E+05	2E+05	2E+05	2E+03	2E+03	2E+03
Listeria monocytogenes	0,86	0,78	0,97	0,94	0,68	0,53	0,82	0,38	0,51	0,26	0,26	0,60
Clostridium beijerinckii	0,60	0,51	0,91	0,33	0,08	0,49	1,36	1,36	1,21	0,68	1,29	0,64
Bacillus cereus	0,53	0,40	0,75	-0,12	-0,25	0,06	-0,89	-0,89	-0,74	0,11	-0,56	-0,32
Streptococcus agalactiae	0,51	0,38	0,26	0,06	0,24	0,26	-0,12	-0,22	0,29	0,44	0,16	0,44
Propionibacterium acnes	0,49	0,40	-0,56	1,03	0,80	0,78	-0,18	-0,29	-0,09	0,80	-0,09	-0,15
Bacteroides vulgatus	0,16	-0,03	-0,03	-0,22	0,21	-0,06	0,94	1,19	1,16	1,33	1,26	1,37
Streptococcus mutans	0,16	0,55	0,11	0,26	0,51	0,19	0,46	0,40	0,29	0,38	0,75	0,42
Escherichia coli	0,14	0,08	-0,84	0,14	0,24	0,21	-1,32	-1,25	-1,64	-0,74	-1,12	-1,06
Staphylococcus aureus/epidermidis	0,14	0,45	0,62	0,16	0,11	0,43	-0,09	-0,18	-0,20	-0,22	-0,17	0,06
Neisseria meningitidis	0,11	-0,09	0,53	-0,06	0,33	0,26	0,64	1,06	0,99	0,53	0,82	1,00
Pseudomonas aeruginosa	0,11	0,38	-1,00	-0,32	0,31	-0,69	-0,69	-0,74	-1,18	-0,06	-0,94	-1,40
Lactobacillus gasseri	-0,06	-0,32	0,38	0,33	-0,84	-0,09	0,03	0,40	0,51	0,31	0,53	-0,03
Deinococcus radiodurans	-0,22	-0,89	-0,84	-0,15	-0,43	-0,25	0,19	0,00	0,19	0,33	0,03	0,29
Enterococcus faecalis	-0,36	-0,40	0,06	-0,15	-0,18	0,11	-1,12	-1,32	-1,56	-2,06	-1,06	-1,84
Acinetobacter baumanni	-0,40	-0,47	-1,00	-0,47	-0,29	-0,43	-1,00	-1,32	-0,84	-0,94	-0,06	-1,94
Actinomyces odontolyticus	-1,12	-1,47	-3,06	-0,89	-1,64	-2,32	-3,06	-2,47	-4,64	-3,64	-4,06	-2,47
Rhodobacter sphaeroides	-1,47	-1,06	-1,25	-2,06	-0,69	-1,06	-0,64	-0,64	-0,64	-1,25	-1,40	-0,84
Helicobacter pylori	-1,84	-1,25	-0,64	-1,56	-1,40	-1,74	-4,64	-4,64	-5,64	-4,06	-5,64	-2,47
Streptococcus pneumoniae	-2,18	-1,74	-1,47	-1,94	-1,94	-1,64	-2,64	-2,18	-2,06	-3,32	-2,47	-1,74

Figure 2. Quantitative accuracy of micPCR/NGS compared to traditional PCR/NGS from synthetic microbial community 16S rRNA gene profiling. The observed species-level frequency data, corrected for the expected species-level frequency ratio for each of the synthetic community members, is shown as a heatmap using a binary logarithm scale. The expected frequency ratio is based on the reported equimolar 16S rRNA operon counts derived from 20 bacterial species. Blue shades indicate an overestimation of species frequency and red colours an underestimation of species frequency. Data from triplicate experiments are presented individually.

micPCR/NGS to determine the microbiota profiles for samples possessing a low diversity of bacteria (nose swabs), a medium bacterial diversity (faeces), and samples containing a high diversity of bacteria (sludge). Results for three independent samples per sample type (nose, faeces, sludge) revealed that chimeric sequences were reduced in all samples from an average of 38.0% (± 15.7%) using traditional PCR/NGS to an average of 1.2% (± 1.3%) using micPCR/NGS (Supplementary Tables 2-4). The reduction of chimera formation resulted in decreased richness values among all samples, particularly among the bacterially diverse faeces and sludge specimens in which micPCR/NGS generated 212 (± 30) OTUs less per 1,000 normalized sequences per sample than the traditional PCR/NGS protocol (Supplementary Figures 1-3). In addition, differences were also observed in the quantitative OTU composition between individual clinical and environmental samples when comparing the micPCR/NGS results to the results obtained using traditional PCR/

NGS. The maximum relative difference between identical OTUs within the same sample obtained by micPCR/NGS compared to traditional PCR/NGS was 17.0, 6.1, and 7.6% measured for the nose, faeces, and sludge samples respectively (Supplementary Tables 5-7).

Finally, single molecule amplification using micPCR actually prevented the generation of chimeric products, due to the fact that we found an increase in chimeric sequences in the micPCR/NGS as the amount of template DNA molecules in micPCRs was increased (Supplementary Table 1). Importantly, the total template DNA molecules in a micelle PCR/NGS protocol should be kept below 10% of total micelle count to avoid any detectable chimera formation due to individual micelles hosting more than one template molecule. Therefore, the final numbers of target DNA molecules have to be carefully adjusted for each micPCR/NGS project to balance reaction yield and reaction specificity according to the experimental requirements.

DISCUSSION

In this report, we show that the use of micelle PCR is particularly suitable for 16S rRNA gene microbiota profiling experiments and strongly reduces the formation of the chimeric 16S rRNA gene amplicons that are a major source of unidentifiable OTUs in microbiome studies. The authors developed and evaluated the use of a micelle-based amplification strategy for 16S rRNA gene profiling of complex samples. Micelle or emulsion based amplification strategies have been successfully applied for a variety of DNA-targeted enzymatic reactions.[11,17] Most notably, Williams et al. published a protocol in 2006 describing the use of emulsion PCR to amplify complex gene libraries that reduce such amplification biases as chimeric sequences and competition between fragments of different lengths.[17] However, standardized commercial kits are now available to buy, which our micPCR protocol used, to offer a straightforward, easy and reproducible method to perform 16S rRNA gene micelle PCR.

Our results show that the use of micelle PCR/NGS greatly reduces chimera formation without the reliance on complex computational methods, resulting in improved microbial diversity estimates. An often-used approach to circumvent the overestimation of richness is to restrict analysis to OTUs that are found more than once, though the accompanying cost is a loss of sequencing sensitivity and accuracy due to the potential removal of singletons that are genuinely very low abundant representatives of their taxa within the total microbiota being profiled.[16] Further, it is true that the confidence of identifying: 1) truly low abundant OTUs and 2) singleton chimeric OTUs, increases as the number of sequence reads per sample is increased when using traditional PCR/NGS. This is because there is an increased chance of detecting multiple (>1) low abundant OTUs as the number of sequence reads increases. However, the researcher has more

confidence that singletons obtained using a micPCR/NGS protocol actually originate from low abundance bacterial species. This is because the number of chimeras formed using micPCR/NGS is very low and independent of the depth of sequencing.

The compartmentalization of template DNA molecules using micPCR/NGS prevents amplicon competition in PCRs, resulting in the generation of more accurate quantitative microbiota profiles. In addition to the standardized synthetic community experiments, different quantification values were also obtained from micPCR/NGS compared to traditional PCR/NGS performed on actual clinical and environmental samples. This results in different interpretations of sample composition and inter-sample variation. For example, micPCR/NGS showed a 3.3-fold reduction in *Staphylococcus* abundance among nose sample 1 compared to nose sample 3 (2.4% vs. 7.8%), whereas traditional PCR/NGS showed a 4.7-fold increase in *Staphylococcus* abundance among nose sample 1 compared to nose sample 3 (12.2% vs. 2.6%). Although the actual composition of these samples is unknown, the quantitative microbiota profiles obtained using micPCR/NGS likely represents a more accurate reflection of the true microbiota profiles as indicated previously using the synthetic community. Therefore, the use of micPCR/NGS will improve and help standardize microbiota profiling during comparative studies (e.g. longitudinal surveys). However, it should be noted that possible effects of sample handling, cell lysis and primer specificity on the final results of these microbiota profiles still exist. These factors should still be optimized for each type of test sample the researcher is investigating.

Taken together, our results show that micPCR/NGS increases the accuracy of 16S rRNA gene microbiota profiling when compared to traditional PCR/NGS, and its use should be recommended for future NGS projects due to the fact that chimera formation and PCR amplicon competition can potentially affect the accuracy of current microbiota profiling results.

METHODS

Sample collection and DNA extraction

Genomic DNA from microbial mock community B (even, low concentration), v5.1 L, catalogue no. HM-782D for 16S rRNA gene microbiota profiling was obtained from BEI Resources, NIAID, NIH as part of the Human Microbiome Project and consists of genomic DNA from 20 bacterial strains with equimolar ribosomal RNA operon counts (100,000 copies per organism per µL). The microbial mock community contains species with different 16 rRNA gene copy numbers in their genomes, ranging from two for *Helicobacter pylori* to 14 for *Clostridium beijerinckii*. Nose swabs and faecal samples were collected from healthy adult volunteers. DNA was extracted from both types of samples using

the QIAsymphony instrument (Qiagen) according to the manufacturer's instructions. DNA was extracted from three sludge samples from river bed, using the Powersoil DNA isolation kit (MO BIO Laboratories, Inc.). The total number of 16S rRNA genes within each sample was quantified as described previously.[18] Prior to use as template for micelle and traditional PCR amplification, the samples were normalized to 1E+03 16S rRNA genes/µL (nasal swabs) or 1E+05 16S rRNA genes/µL (faeces and sludge samples).

Micelle PCR amplification

The micPCR consisted of two PCR rounds of micPCR amplification. This was necessary, because micPCR only yields a limited number of amplicons per template molecule, which is a consequence of the limited reaction volume contained in a single micelle. We estimated that after a micPCR only 1E+04 amplicon molecules were formed in a single micelle starting with a single genomic DNA fragment carrying a 16S rRNA gene copy. This low number of amplicon molecules is not sufficient for NGS of samples containing low amounts of bacterial DNA, such as nose swabs. However, using a second round of micPCR allowed us to increase the number of amplicon molecules for NGS, as well as allowing the addition of molecular identification (MID) sequences and Roche 454 specific A and B sequences. In the first step, micPCR was performed using modified 357F and 926R primers that amplified the V3-V5 regions of 16S rRNA genes and which incorporated universal sequence tails at their 5' ends. In the second step, a micPCR was again used, but to amplify micPCR amplicons obtained from the first step micPCR. The second step micPCR utilized primers containing complementary sequences to the universal tails and included additional 454 sequencing-specific nucleotides, and specimen-specific MIDs. For both amplification steps, water-in-oil emulsions were prepared using the Micellula DNA Emulsion Kit (Roboklon). The oil phase comprised ~73% Emulsion component 1, ~7% Emulsion component 2, and 20% Emulsion component 3, which was mixed for 5 minutes in a cold room as described by the manufacturer. The aqueous phase was a PCR mix comprising 0.01 mg/mL BSA, 2 µM of each primer, 200 µM dNTP mix, and 2.5 U Taq polymerase with 1x PCR Buffer B (EURx). Template DNA and water were added to give a final volume of 50 µL for each sample. Water-in-oil emulsions were prepared by adding 50 µL of pre-cooled PCR mix to 300 µL of pre-cooled oil phase. The first round of micPCR was carried out using the following cycling conditions: 95°C for 2 minutes followed by 25 cycles of PCR, with cycling conditions of 15 seconds at 95°C, 30 seconds at 55°C, and 60 seconds at 72°C, and a final extension at 72°C for 7 minutes. Emulsions were broken by the addition of 1 mL 2-butanol, and 400 µL of Orange-DX buffer (Roboklon) was added to the broken emulsion solution. This solution was centrifuged for phase separation. For the purification of DNA within the water phase, NucliSENS EasyMAG reagents (Biomérieux) were used according to the manufacturer's instructions. To normalize DNA concentration and reduce the number of template molecules for the second round of

amplification, the purified DNA was diluted 1E+04 or 1E+02-fold for high and low inputs, respectively, during the first micPCR. The second round of micPCR was performed under the following conditions: initial denaturation at 95°C for 2 minutes followed by 25 cycles of PCR, with cycling conditions of 15 seconds at 95°C, 30 seconds at 50°C and 60 seconds at 72°C. During the first 10 cycles of PCR, the annealing temperature was increased by 0.5°C per cycle to an annealing temperature of 55°C. The PCR was stopped after a final extension at 72°C for 7 minutes. Again, emulsions were broken using 2-butanol, and DNA was purified using NucliSENS EasyMAG reagents (Biomérieux).

Traditional PCR amplification

PCRs were performed in 10 µL volumes using the FastStart High Fidelity Reaction Kit (Roche) with the addition of 0.5 µM of each PCR primer. Resolight Dye (Roche) was added to measure DNA amplification in real-time using a LightCycler 480 instrument (Roche). The 16S rRNA gene V3-V5 regions were amplified by PCR using modified 357F and 926R primers to allow for a two-step amplification strategy, using the following cycling conditions: initial denaturation at 95°C for 2 minutes followed by 35 cycles of PCR, with cycling conditions of 30 seconds at 95°C, 30 seconds at 55°C, and 60 seconds at 72°C. After PCR amplification, the amplicons were purified from unincorporated dNTPs, primers, primer dimers and salts using magnetic AMPure XP beads (Agencourt). The purified 16S rRNA gene amplicons were re-amplified to incorporate 454 sequencing-specific nucleotides and specimen-specific MIDs. All PCRs were performed in 10 µL reaction volumes using the FastStart High Fidelity Reaction Kit with the addition of 0.5 µM of each PCR primer and the Resolight Dye. The PCRs were performed on a LightCycler 480 instrument, but under modified conditions: initial denaturation at 95°C for 2 minutes followed by 35 cycles of PCR, with cycling conditions of 30 seconds at 95°C, 30 seconds at 50°C, and 60 seconds at 72°C. During the first 10 cycles of PCR, the annealing temperature was increased by 0.5°C per cycle to an annealing temperature of 55°C. Bar-coded amplicons were mixed in equimolar concentrations and the complete pool was purified by gel extraction using the QIAquick Gel Extraction Kit (Qiagen), followed by a second purification with magnetic AMPure XP beads.

Quantification of 16S rRNA gene molecules

In preparation for 454 sequencing (Roche), the concentration of purified amplicons obtained by micPCR and traditional PCR was measured using a 16S rRNA gene quantitative PCR (qPCR). The qPCRs were performed in 10 µL reaction volumes using the LightCycler FastStart DNA Master SYBR Green I Kit (Roche) with the addition of 0.5 µM of amplification primer 357F and 926R without the universal tails. The PCRs were performed on a LightCycler 1.0 instrument (Roche), under the following conditions: initial denaturation at 95°C for 10 minutes followed by 45 cycles of PCR, with cycling conditions of 1 second

at 95°C, 5 seconds at 55°C, and 30 seconds at 72°C. The concentration of purified amplicons obtained by micPCR and traditional PCR were normalized to 1E+05 molecules/µL using a serial dilution of a standard solution containing 16S rRNA genes derived from a highly bacterial diverse sludge sample that was calibrated using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies).

Data analysis

The composition of microbiota was determined by sequencing 16S rRNA genes using the 454 GS Junior Sequencer platform (Roche) according to the manufacturer's instructions. NGS-data were automatically processed using the 'Full Processing Amplicon' pipeline available through the Run Wizard on the GS Junior Attendant PC (Roche). FASTA-formatted sequences were extracted from the .sff data files and processed using modules implemented in the mothur v. 1.33.0 software platform.[19] Primer sequences were trimmed and sequences with length smaller than 400 were removed from the analysis. In addition, only the first 450 bases of each sequence were used for further analysis. In order to characterize the number of chimeric sequences more precisely, no additional quality filtering was applied. Unique sequences were aligned using the 'align.seqs' command and an adaptation of the Bacterial SILVA SEED database release 119 as a template (available at: https://www.mothur.org/wiki/Silva_reference_ files). Potentially chimeric sequences were detected and removed with the Uchime source code, using firstly the sequences as their own reference and sequentially the SILVA alignment version of the gold database (available at: https://www.mothur.org/wiki/Silva_reference_files) as reference. The remaining aligned sequences were classified using a naïve Bayesian classifier with the SILVA SEED database release 119 and clustered into OTUs defined by 97% similarity. To reduce the effects of uneven sampling, all nose swab samples were rarefied to 500 sequences per sample and all other samples, including the synthetic community, faeces, and sludge samples, were rarefied to 1,000 sequences per sample. For all samples, rarefaction curves were plotted and the inverse Simpson's diversity index and Good's coverage were calculated. Finally, OTUs corresponding to the *Streptococcus* genus within the synthetic community were determined at species-level by checking the representative sequences against the reference sequences using BioNumerics version 5.10 (Applied Math).

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu). The following reagent was obtained through BEI Resources, NI-

AID, NIH as part of the Human Microbiome Project: Genomic DNA from microbial mock community B (even, low concentration), v5.1L, for 16S RNA gene sequencing, HM-782D.

SUPPLEMENTARY DATA

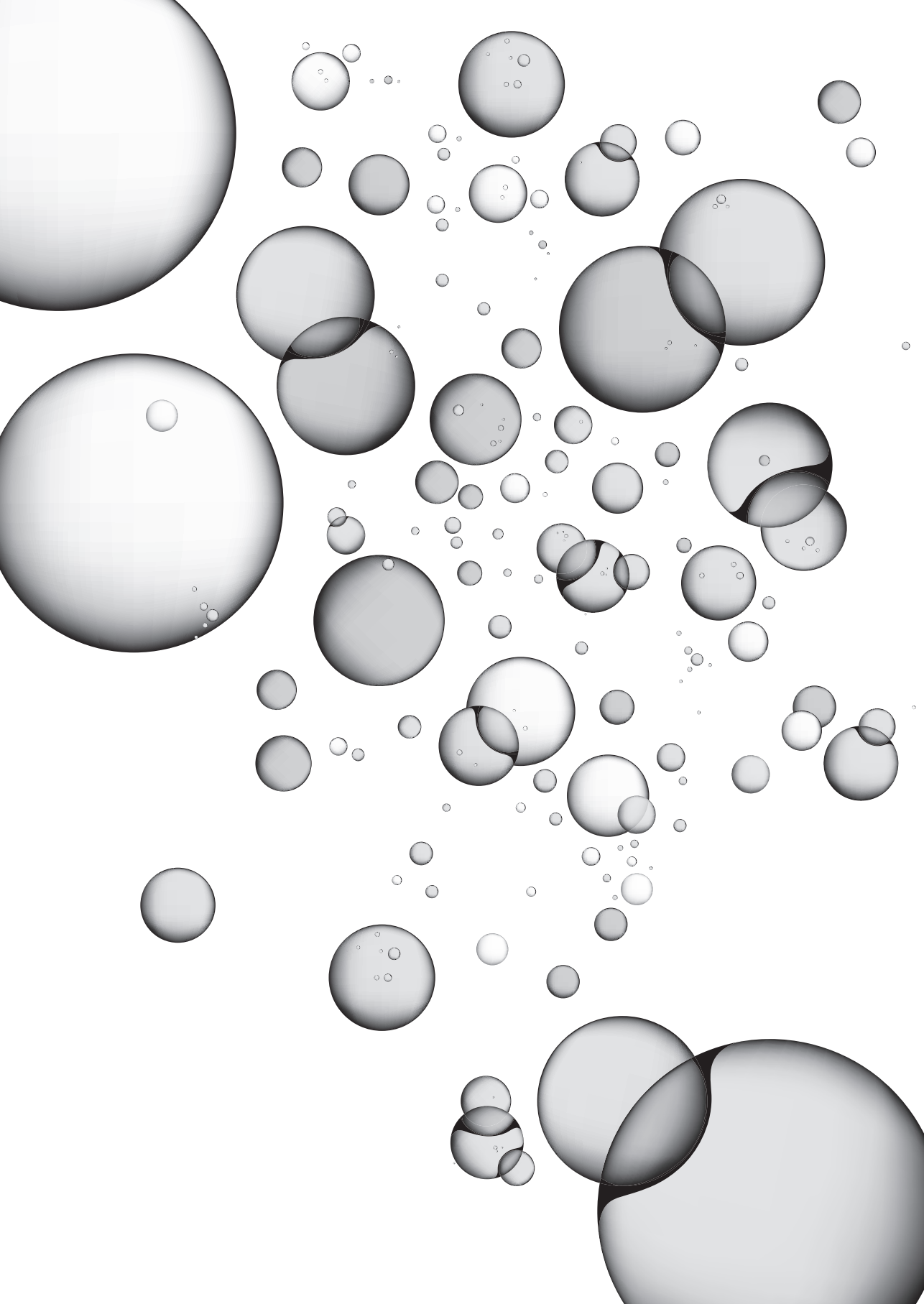
Supplementary information accompanies this paper at https://www.nature.com/articles/srep14181.

REFERENCES

1. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. *Nature* 2014; **509**: 357-360.
2. Weinstock GM. Genomic approaches to studying the human microbiota. *Nature* 2012; **489**: 250-256.
3. Goodrich JK, Di Rienzi SC, Poole AC, et al. Conducting a microbiome study. *Cell* 2014; **158**: 250-262.
4. 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.
5. Turroni F, Peano C, Pass DA, et al. Diversity of bifidobacteria within the infant gut microbiota. *PloS one* 2012; **7**: e36957.
6. 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.
7. 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.
8. Edgar RC, Haas BJ, Clemente JC, et al. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; **27**: 2194-2200.
9. 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.
10. Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* 1998; **64**: 3724-3730.
11. Shao K, Ding W, Wang F, et al. Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection. *PloS one* 2011; **6**: e24910.
12. Dressman D, Yan H, Traverso G, et al. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* 2003; **100**: 8817-8822.
13. 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.
14. Berry D, Ben Mahfoudh K, Wagner M, et al. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl Environ Microbiol* 2011; **77**: 7846-7849.
15. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, et al. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* 2005; **71**: 8966-8969.
16. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 2013; **10**: 996-998.
17. Williams R, Peisajovich SG, Miller OJ, et al. Amplification of complex gene libraries by emulsion PCR. *Nat Methods* 2006; **3**: 545-550.

18. Yang S, Lin S, Kelen GD, et al. Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. *J Clin Microbiol* 2002; **40**: 3449-3454.

19. 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.



Chapter 4

Novel micelle PCR-based method for accurate, sensitive and quantitative microbiota profiling

Stefan A. Boers
John P. Hays
Ruud Jansen

Sci Rep 2017; 7: 45536.

ABSTRACT

In the last decade, many researchers have embraced 16S rRNA gene sequencing techniques, which has led to a wealth of publications and documented differences in the composition of microbial communities derived from many different ecosystems. However, comparison between different microbiota studies is currently very difficult due to the lack of a standardized 16S rRNA gene sequencing protocol. Here we report on a novel approach employing micelle PCR (micPCR) in combination with an internal calibrator that allows for standardization of microbiota profiles via their absolute abundances. The addition of an internal calibrator allows the researcher to express the resulting operational taxonomic units (OTUs) as a measure of 16S rRNA gene copies by correcting the number of sequences of each individual OTU in a sample for efficiency differences in the NGS process. Additionally, accurate quantification of OTUs obtained from negative extraction control samples allows for the subtraction of contaminating bacterial DNA derived from the laboratory environment or chemicals/reagents used. Using equimolar synthetic microbial community samples and low biomass clinical samples, we demonstrate that the calibrated micPCR/NGS methodology possess a much higher precision and a lower limit of detection compared with traditional PCR/NGS, resulting in more accurate microbiota profiles suitable for multi-study comparison.

INTRODUCTION

The number of microbiota studies has rapidly increased since the introduction of next-generation sequencing (NGS) technologies and opened numerous new fields of research. This research often studies microbial changes as a result of various kinds of interventions and focusses on changes in the proportional microbial composition rather than actual microbial quantities. However, obtaining accurate and quantitative microbiota profiles makes high demands on the analytical sensitivity, specificity and reproducibility of standard NGS methodologies and requires a careful consideration of potential biases and bacterial DNA contaminations that can be introduced during the many steps of sample processing and sequencing.[1,2]

Recently, the authors published a micelle PCR/NGS (micPCR/NGS) methodology that limits both chimera formation and PCR competition, thereby reducing the introduction of PCR amplification biases into microbiota profiles.[3] However, 16S rRNA gene sequencing techniques still remain semi-quantitative methods, where the results are presented as proportional abundances of operational taxonomic units (OTUs), rather than absolute abundances of OTUs. This restriction lowers the reproducibility of microbiota profiling between different laboratories and does not reveal differences in absolute abundances of specific OTUs. For example, Hiergeist et al. observed high inter-laboratory deviations from an external quality assessment of 16S rRNA gene sequencing methods and concluded that there is an urgent need to develop microbiota profiling methods with an increased cross-study comparability.[4] This is of particular importance for the increasing implementation of 16S rRNA gene sequencing methods in the field of medical diagnostics, which requires high quality demands on the methods used. Further, the validity of 16S rRNA gene sequencing results are threatened by the presence of contaminating bacterial DNA derived from the laboratory environment and the consumables used in the experimental set-up.[5] Such contamination is particularly relevant for an accurate analysis of microbiota composition of low biomass samples (e.g. skin swabs).[6,7] These contaminating DNA molecules can be derived from two sources: 1) contaminating bacterial DNA present within the sample-processing environment, and 2) contaminating bacterial DNA already present in the reagents/consumables used during sample processing. The introduction of contaminating bacterial DNA derived from the processing environment will occur randomly within the samples being processed and can be recognized by their non-reproducibility in multiple determinations of the microbiota of a particular sample. In contrast, the variety of contaminating bacterial DNA from reagents/consumables will be dependent on the manufacturer and batch numbers of the actual reagents/consumables used and will be present in all samples processed using these particular reagents/consumables.

In this report, we present a novel approach that employs micPCR/NGS in combination with an internal calibrator (IC) to determine the composition and absolute quantity of microbial genera. The IC used for this study consisted of quantified genomic DNA from a *Synechococcus* bacterium species that is absent in the natural microbial flora of the samples under investigation and allows the researcher to express the resulting OTUs as a measure of 16S rRNA gene copies by the use of a correction factor (sample OTU copies = sample OTU reads x (initial IC copies/IC OTU reads)). We utilized this calibrated micPCR/NGS approach to process a range of samples in triplicate in order to increase accuracy and to correct for contaminating bacterial DNA derived from the laboratory environment. In addition, contaminating bacterial DNA derived from the reagents/consumables used during micPCR/NGS were subtracted from samples via the use of a negative extraction control (NEC). To validate the calibrated micPCR/NGS approach (including the two-step strategy for removing contaminating bacterial DNA – as described above), we used a series of dilutions of an equimolar synthetic microbial community (SMC) sample and compared the results obtained against traditional PCR/NGS. Additionally, we evaluated the performance of our method to generate accurate quantitative microbiota profiles from actual low biomass clinical samples.

RESULTS

In order to determine the accuracy (trueness and precision) of the calibrated micPCR/NGS methodology, we utilized a 10-fold dilution series of a SMC sample containing equimolar 16S rRNA gene copies of *Clostridium perfringens*, *Staphylococcus aureus*, *Haemophilus influenzae*, and *Moraxella catarrhalis* (ranging from 2.5 to 2,500 16S rRNA gene copies per (mic)PCR of each bacterial species). Prior to amplification, *Synechococcus* DNA was added in such a concentration that each SMC sample contained 10% of IC 16S rRNA gene copies with a minimum of 50 copies for samples that contained less than 500 16S rRNA gene copies. The 16S rRNA gene V3-V4 regions were amplified in triplicate, using the same SMC/IC sample for each replicate, and sequenced using both micPCR/NGS and traditional PCR/NGS (as comparator). As shown in Supplementary Table 1, we obtained an average of 4,201 (± 2,398) QC-passed sequence reads per SMC sample using both methods, of which the percentage of chimeric sequences was much lower for micPCR/NGS compared to traditional PCR/NGS (0.01% ± 0.03% vs. 4.56% ± 2.97%). Further, Table 1 shows the results obtained from triplicate experiments and indicates the accuracy for both methodologies used. The micPCR/NGS data, as well as the traditional PCR/NGS data, showed a median value of only a 1.3-fold difference between the measured 16S rRNA gene copies (average of triplicate experiments) and the expected 16S rRNA gene copies. Although these data suggest a similar and good trueness among both methods,

Table 1. Accuracy of 16S rRNA gene copy determination using synthetic microbial community (SMC) samples comparing the results of micPCR/NGS to traditional PCR/NGS.

micPCR/NGS						
OTU	**Expected**	**Replicate 1**	**Replicate 2**	**Replicate 3**	**Trueness**	**Precision**
Clostridium	2500	6,735	3,840	4,347	2.0	0.3
Staphylococcus	2500	4,776	3,147	4,875	1.7	0.2
Haemophilus	2500	4,082	3,133	2,611	1.3	0.2
Moraxella	2500	3,714	2,213	1,056	0.9	0.6
Clostridium	250	487	631	641	2.3	0.1
Staphylococcus	250	486	579	375	1.9	0.2
Haemophilus	250	238	225	183	0.9	0.1
Moraxella	250	225	363	214	1.1	0.3
Clostridium	25	28	31	52	1.5	0.3
Staphylococcus	25	57	47	52	2.1	0.1
Haemophilus	25	10	9	54	1.0	1.1
Moraxella	25	19	5	29	0.7	0.7
Clostridium	2.5	0	4	0	0.6	1.6
Staphylococcus	2.5	1	15	19	4.6	0.8
Haemophilus	2.5	1	0	9	1.3	1.6
Moraxella	2.5	1	3	0	0.6	1.1
traditional PCR/NGS						
OTU	**Expected**	**Replicate 1**	**Replicate 2**	**Replicate 3**	**Trueness**	**Precision**
Clostridium	2500	953	6,638	7,340	2.0	0.7
Staphylococcus	2500	403	3,793	4,075	1.1	0.7
Haemophilus	2500	370	2,483	2,509	0.7	0.7
Moraxella	2500	36	3,034	3,604	0.9	0.9
Clostridium	250	736	513	497	2.3	0.2
Staphylococcus	250	302	210	284	1.1	0.2
Haemophilus	250	281	226	261	1.0	0.1
Moraxella	250	240	188	231	0.9	0.1
Clostridium	25	119	29	36	2.4	0.8
Staphylococcus	25	15	27	50	1.2	0.6
Haemophilus	25	112	28	28	2.3	0.9
Moraxella	25	116	2	0	1.6	1.7
Clostridium	2.5	0	11	1	1.7	1.5
Staphylococcus	2.5	243	6	1	33.3	1.7
Haemophilus	2.5	0	0	10	1.3	1.7
Moraxella	2.5	0	7	0	1.0	1.6

The expected and measured values (Replicate 1-3) represent the number of 16S rRNA gene copies obtained for each individual bacterial species at four different input DNA concentrations (2,500, 250, 25 and 2.5 16S rRNA gene copies). The trueness shows the closeness of measurement results to the true (expected) value and was calculated by dividing the number of 16S rRNA gene copies measured (as an average of triplicate results) to the expected number of 16S rRNA gene copies present in the calibrated synthetic microbial community (SMC). The precision shows the coefficient of variation that was calculated by dividing the standard deviation obtained from triplicate results to the average number of 16S rRNA gene copies measured.

the dispersal of replicate results varied greatly using traditional PCR/NGS. As shown in Figure 1, the dispersal of replicate results obtained using micPCR/NGS was much smaller than using traditional PCR/NGS, indicating the higher precision of the micPCR/NGS methodology. For example, traditional PCR/NGS resulted in a coefficient of variation of 0.8, 0.5, 0.9 and 3.0 for the SMC samples containing 2,500, 250, 25 and 2.5 16S rRNA gene copies per bacterial species, respectively. In contrast, identical experiments using micPCR/NGS showed a coefficient of variation of only 0.4, 0.4, 0.6 and 1.4, respectively. The higher precision of the micPCR/NGS methodology lowers the number of random errors within 16S rRNA gene measurements and increases the repeatability of microbiota profiling results (Paired Wilcoxon signed-rank test; $p < 0.01$). As expected, the accuracy for both methods decreases as the number of template DNA molecules decreased due to the limited chance of successfully generating 16S rRNA gene amplicons at very low starting concentrations of DNA.

Microbiota analysis is prone to the introduction of contaminating bacterial DNA molecules during sample processing. This is illustrated by the finding of 114 distinct OTUs that are represented by at least two or more sequence reads in our SMC experiments (Supplementary Tables 2-5). As expected, the number of distinct OTUs was maximal in the samples containing low amounts of input DNA, or no input DNA (NEC), leading to an unintended overestimation of microbial diversity. To correct for random bacterial DNA contamination from the laboratory environment, we processed each sample in triplicate and removed all OTUs that were not present in all of the three datasets obtained. To

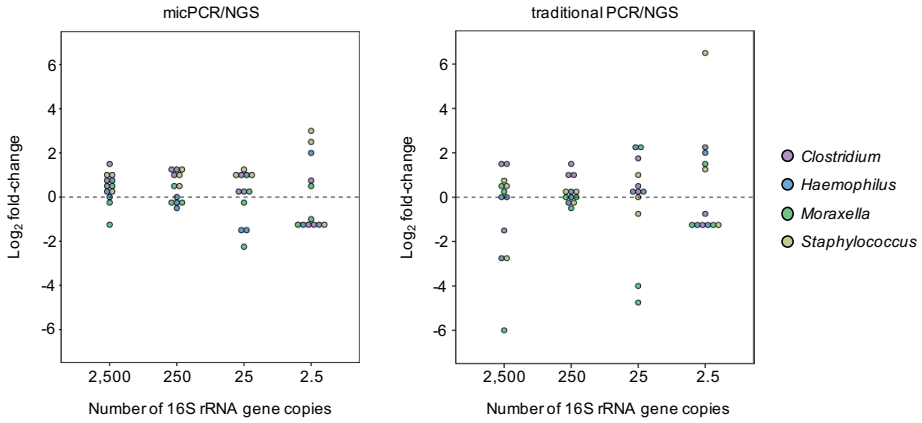


Figure 1. Precision of 16S rRNA gene copy determination using synthetic microbial community samples comparing the results of micPCR/NGS to traditional PCR/NGS. The synthetic microbial community (SMC) samples tested contained equimolar 16S rRNA gene copy numbers derived from four different bacterial species and ranged from 2,500 to 2.5 16S rRNA gene copies per species. Coloured data points represent the individual measurements per bacterial OTU from triplicate experiments, corrected for the number of expected 16S rRNA gene copies and plotted using a binary logarithmic scale.

correct for bacterial DNA contamination from the reagents/consumables used, we subtracted the contribution of the NEC from each sample. The bacterial DNA contamination from reagents/consumables was calculated as the mean plus three standard deviations of 16S rRNA gene copies per OTU that were present in all three independent NEC measurements (Supplementary Tables 6 and 7). The quantified microbiota profiles obtained using micPCR/NGS and traditional PCR/NGS, before and after the correction of contaminating bacterial DNA, are shown in Figure 2. Correcting for both types of bacterial DNA contamination resulted in the complete removal of contaminating bacterial DNA from

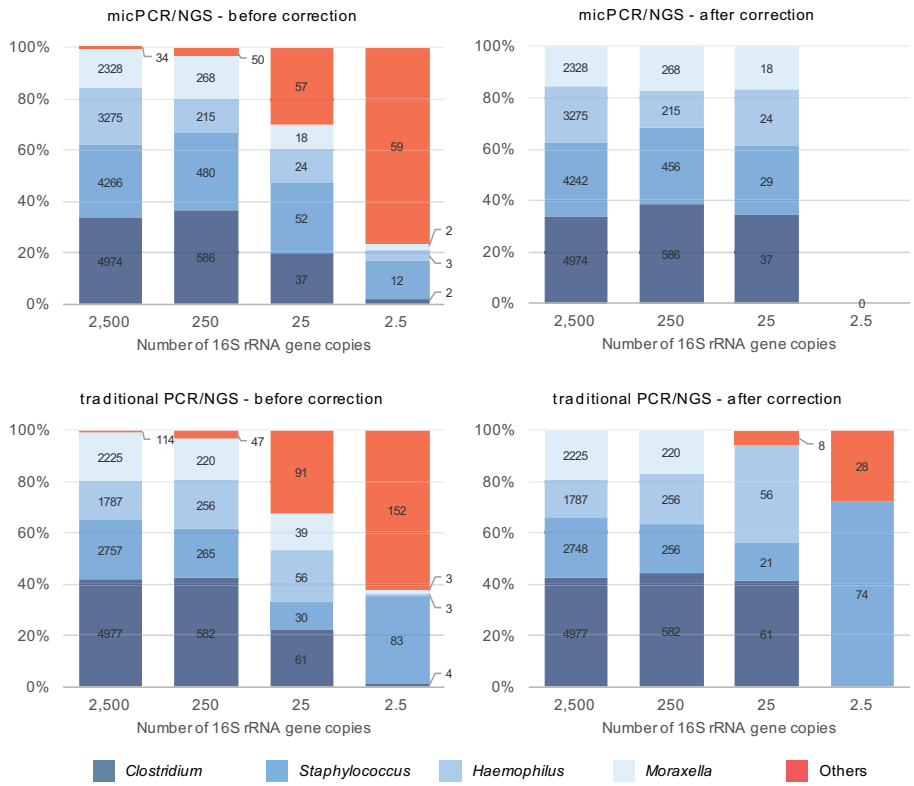


Figure 2. 16S rRNA gene microbiota profiles obtained from synthetic microbial community samples comparing the results of micPCR/NGS to traditional PCR/NGS before and after correction for contaminating bacterial DNA. The synthetic microbial community (SMC) samples tested comprised equimolar 16S rRNA gene copies derived from *C. perfringens*, *S. aureus*, *H. influenzae*, and *M. catarrhalis* and ranging from 2,500 to 2.5 16S rRNA gene copies per bacterial species. Averages of triplicate micPCR/NGS and triplicate traditional PCR/NGS results are shown in 100% stacked bars before and after correction for contaminating bacterial DNA. The correction of contaminating bacterial DNA comprises two steps: 1) eliminating OTUs that could not be reproducibly measured in triplicate experiments, and 2) subtracting 16S rRNA gene copies that were also quantified in triplicate measurements of a negative extraction control (NEC) sample. Values within bars represent the calculated number of 16S rRNA gene copies per bacterial OTU.

SMC samples generated using the micPCR/NGS methodology. In contrast, traditional PCR/NGS results still showed contaminating bacterial DNA present at 25 and 2.5 16S rRNA gene copies per organism, even after correction. This finding illustrates the higher accuracy of the micPCR/NGS methodology to quantify contaminating bacterial DNA from NEC samples and its requirement for the accurate subtraction of contaminating bacterial DNA from actual samples. In addition, the traditional PCR/NGS methodology failed to recover 25 16S rRNA gene copies of *M. catarrhalis* and hugely overestimated the abundance of *S. aureus* 16S rRNA gene copies in the SMC samples containing 2.5 16S rRNA gene copies per bacterium. Additionally, micPCR/NGS successfully detected all four bacterial species at 2,500, 250 and 25 16S rRNA gene molecules per bacterium, but failed to detect any of these species in triplicate experiments using 2.5 16S rRNA gene copies. We therefore estimated that the analytical sensitivity, indicated as the limit of detection (LOD), for micPCR/NGS is 25 16S rRNA gene molecules per OTU, whereas the LOD of traditional PCR/NGS was estimated as 250 16S rRNA gene copies per OTU.

In order to investigate the performance of the calibrated micPCR/NGS protocol using clinical samples, we selected four human skin swab samples that contained a low number of 16S rRNA gene copies (range: 64-604 16S rRNA gene copies/µL). The four samples, including an NEC, were processed in triplicate using micPCR/NGS and traditional PCR/NGS in parallel. Using micPCR/NGS, we obtained an average of 7,126 (± 1,702) QC-passed sequence reads per sample, of which only 2 (± 3) sequences per sample were identified and removed as chimeric sequences (Supplementary Table 8). Next, we were able to detect 3 to 13 OTUs in the samples, of which *Staphylococcus* and *Neisseria* species were commonly found and could be confirmed by bacterial culture. In addition, the micPCR/NGS method also detected the skin inhabitants *Streptococcus*, *Paracoccus*, *Enhydrobacter*, *Gemella*, *Sphingomonas*, *Alloprevotella*, *Propionibacterium*, *Brevundimonas*, *Roseomonas*, *Rothia*, *Granulicatella*, *Rhodococcus*, and *Flavobacteriaceae* species.[8,9] Importantly, correcting for contaminating bacterial DNA using the two-step strategy as described above removed 92% (range 89 – 96%) of the resultant OTUs found in the skin swab samples (Supplementary Table 9). Finally, we measured a high concordance between the total bacterial biomass when estimated indirectly using calibrated micPCR/NGS, without the correction for contaminating bacterial DNA, compared to the total 16S rRNA gene copies obtained when estimated directly using a 16S rRNA gene qPCR according to Yang et al.,[10] with an average of only a 1.03-fold difference (± 0.34), demonstrating the accuracy of the micPCR/NGS method. In contrast to micPCR/NGS, traditional PCR/NGS was not able to generate any 16S rRNA gene amplicons from these low biomass skin swab samples as this method only generated non-specific, low molecular weight amplicons. These non-specific PCR products were most like generated by the relative excess of human DNA – compared to bacterial DNA – in the samples, as the PCR/NGS methodology successfully generated 16S rRNA gene amplicons in SMC samples (where

only bacterial DNA and no human DNA was present). This result indicates that the LOD for the traditional PCR/NGS methodology is even higher for actual clinical samples than was previously estimated using our SMC samples.

DISCUSSION

In this report, we show that a micelle PCR/NGS methodology (micPCR/NGS), in combination with a unique internal calibrator (IC) generates robust and accurate quantitative microbiota profiles. Micelle PCR is characterized by the clonal amplification of template DNA by the physical separation of reaction ingredients into a large number of reaction compartments. In contrast to traditional PCR, which is performed in a single reaction volume, the compartmentalization during micPCR allows accurate quantification of target sequences due to a lower susceptibility to variations in PCR amplicon efficiencies and amplification biases such as chimera formation, false priming and primer dimer formation.[3] Therefore, our calibrated micPCR/NGS method allows for the utilization of just a single correction factor, obtained from a single IC, to convert 16S rRNA gene sequence reads to 16S rRNA gene copies for each individual OTU present within a sample. In contrast, optimized traditional PCR amplification protocols,[11] or alternative spike-in approaches that employ traditional PCR amplification methods, such as SCML,[12] remain vulnerable to template-specific variations in PCR efficiencies that could easily result in error-derived microbiota profiles. However, it is true that the number of 16S rRNA gene copies produced using micPCR depends on the actual volume of individual micelles and that possible quantitation biases might be introduced due to differences in micelle sizes. This may be particularly relevant for the accurate quantification of low abundant taxa that are more vulnerable to the possible stochastic distribution of template DNA molecules into unevenly shaped micelles. However, the randomness of micelle size frequency distributions generated in independent experiments will tend to average out any possible quantification bias generated due to differences in the distribution of micelle sizes between independent experiments, as indicated by the results obtained using our synthetic microbial community (SMC).

The absolute quantification of OTUs using micPCR/NGS in combination with an internal calibrator improves the standardization of microbiota profiling results by removing the susceptibility to compositional effects. For example, traditional 16S rRNA gene sequencing methods require specific tools and methods that properly account for the statistical implications from the compositional structure of the data obtained.[13] But despite the growing interest and recent efforts to develop these sophisticated methods, the problems of spurious correlations in compositional data remain as yet unsolved.[14] In contrast, our results show that the use of the calibrated micPCR/NGS strategy greatly

improves standardization of microbiota research without the reliance on (complex) compositional data analysis. However, these results remain vulnerable to contaminating bacterial DNA molecules derived from the sample processing environment. In order to correct for contaminating bacterial DNA, we eliminated OTUs that could not be reproducibly measured in triplicate experiments and subtracted 16S rRNA gene copies that were also quantified from negative extraction control (NEC) samples. The comparison of the micPCR/NGS and traditional PCR/NGS results showed that both methods possess a similar trueness when profiling the microbiota of synthetic microbial community (SMC) samples. However, the precision of the micPCR/NGS was much higher compared to the traditional PCR/NGS method. The low precision of the traditional PCR/NGS methodology resulted in unpredictable random errors within the microbiota profiles obtained, including those obtained from NEC samples. The increased random errors observed using traditional PCR/NGS makes the accurate subtraction of bacterial DNA contamination unreliable, whilst in contrast, the high precision of micPCR/NGS resulted in highly accurate quantification of (contaminating) 16S rRNA gene copies resulting in improved quantitative microbiota profiles that were free of contaminating bacterial DNA from environmental sources.

Using micPCR/NGS, we determined the quantitative microbiota profiles of low biomass skin swab samples. As expected, most OTUs (> 89%) obtained from these clinical samples could not be reliably determined in triplicate, or the quantified number of 16S rRNA gene copies did not exceed the quantified number of the same 16S rRNA gene copies determined within NEC samples, and were removed accordingly. This finding stresses the importance of removing contaminating bacterial DNA from microbiota profiles obtained using low biomass samples. Additionally, traditional PCR/NGS was not able to generate any useful 16S rRNA gene sequencing data using the same clinical samples. This result is probably caused by other sample components than bacterial DNA, such as human DNA, that interferes with traditional PCRs (via inhibition or competition). This finding also illustrates the specific nature of PCR in micelles. Since all sample components (both 16S rRNA gene templates and non-templates) are limited to a single micelle, micPCRs are not affected by inhibiting or competing components and are still able to generate 16S rRNA gene amplicons successfully. However, it is important to note that possible effects of sample storage conditions,[15] cell lysis,[16] and primer specificity[17] on the final results of these microbiota profiles still exist.

In summary, a combination of micPCR/NGS and an internal calibrator generates robust and accurate quantitative microbiota profiles. The high accuracy and low limit of detection of the calibrated micPCR/NGS, makes this method the preferential method to determine accurate and quantitative microbiota profiles for low biomass samples that are hampered by contaminating bacterial DNA. The general adoption of this approach

by microbiota investigators will greatly improve the standardization of microbiota profiling results between individual experiments, laboratories and scientific publications.

METHODS

Synthetic microbial community samples

The DNA used to create the SMC samples was extracted from four independently cultured bacterial strains; *Moraxella catarrhalis* (ATCC 25240), *Staphylococcus aureus* (ATCC 43300), *Haemophilus influenzae* (ATCC 10211), and *Clostridium perfringens* (ATCC 12915), using a phenol/bead-beating protocol combined with the AGOWA mag Mini DNA Isolation Kit (LGC) as described previously.[6] In addition, DNA from elution buffer BL (LGC) was extracted as a negative extraction control (NEC) sample at the same time to assess the composition of contaminating bacterial DNA in the experimental methodologies. In order to generate an equimolar mixture of 16S rRNA gene targets from the four bacterial DNA extracts, and to normalize the *Synechococcus* sp. (ATCC 27264D-5) DNA used as IC, the total dsDNA concentration from each DNA isolate was determined individually using the Quant-iT PicoGreen dsDNA assay Kit (Life Technologies) and normalized for genome sizes and 16S rRNA gene copy numbers based on bacterial whole-genome sequences that are publicly available at the NCBI database. Next, a 10-fold dilution series of the equimolar SMC sample was made, ranging from 2,500 to 2.5 16S rRNA gene copies per organism. Prior to amplification by either micPCR or traditional PCR, 1,000, 100, 100, 50 and 50 16S rRNA gene copies of *Synechococcus* DNA was added as IC to the SMC DNA extracts containing 2,500, 250, 25, 2.5 16S rRNA gene copies per bacterial species and the NEC DNA extract, respectively.

Skin swab samples

An acknowledged national ethics committee from the Netherlands (Medisch Ethische Toetsingscommissie Noord-Holland, http://www.metc.nl) approved the study protocol (M015–021) and all experiments were performed in accordance with the relevant guidelines and regulations. Skin swab samples were collected from patients with atopic dermatitis after written, informed consent was obtained from all subjects. Skin swab samples were collected using E-swabs (490CE, Copan) by gently rubbing the dry flocked swab over the dermatitis lesion (~2 cm^2) for 10 seconds after which the entire sample was eluted upon contact with 1 mL liquid Amies preservation medium. All samples were cultured according to standard laboratory protocols performed in our laboratory and stored at −80°C for subsequent 16S rRNA gene sequencing analysis. The routine culture methods included aerobic overnight culture at 35°C on CAP, TSASB and CLED agar plates after which Matrix-Assisted Laser Desorption Ionization Time-Of-Flight (MALDI-TOF)

mass spectrometry was used for the identification of cultured bacterial species. For 16S rRNA gene sequencing analysis, DNA was extracted from the skin swab samples using the High Pure PCR Template Preparation Kit (Roche) according to the manufacturer's instructions. In addition, DNA from Amies medium (490CE, Copan) was extracted as an NEC sample at the same time to allow contaminating bacterial DNA subtraction after NGS processing. Note that in this study we did not focus on DNA extraction efficiencies and the DNA extraction kit used may not be 100% efficient for determining microbial communities from skin swab samples. The total number of 16S rRNA gene copies within each DNA extract was measured using a 16S rRNA gene quantitative PCR (qPCR) according to Yang et al.[10] For this, CT-values were related to a serial dilution of the previously calibrated and normalized SMC sample and ranged from a total of 100 to 10,000 16S rRNA gene copies per PCR. Prior to amplification by either micPCR or traditional PCR, 50 16S rRNA gene copies of *Synechococcus* DNA was added as IC to the skin swab DNA extracts and the NEC DNA extract.

Micelle PCR and traditional PCR amplification

16S rRNA gene amplicon library preparation using micPCR and traditional PCR was performed as previously published,[3] but with a slight modification. In this study, both amplification strategies were performed using modified 341F (5'-GAC ACT ATA GCC TAC GGG RSG CAG CAG-3') and 806R (5'-CAC TAT AGG GAC TAC NVG GGT WTC TAAT-3') primers that amplified the V3-V4 regions of 16S rRNA genes and which incorporated universal sequence tails at their 5' ends to allow for a two-step amplification strategy. Also, both micPCR and traditional PCR were performed using the same PCR reagents (except for the oil phase used to generate the micelles) and PCR conditions following the Micelle PCR Amplification protocol as previously published.[3] Finally, both micPCR/NGS and the traditional PCR/NGS methodology utilized the same amplicon purification steps to synchronize experimental conditions.

16S rRNA gene sequencing and data analysis

Bidirectional sequencing of the 16S rRNA gene amplicon libraries was performed using the 454 Genome Sequencer (GS) Junior platform (Roche), with FASTA-formatted sequences being extracted from the GS Junior machine and further processed using the mothur v. 1.33.0 software package.[18] Primer sequences were trimmed and sequences that had an ambiguous base call (N) in the sequence or with lengths smaller than 400 were removed from the analysis. Unique sequences were then aligned against a customized reference alignment based on the SILVA reference alignment release 119 (available at: https://www.mothur.org/wiki/Silva_reference_files). The reference sequences were trimmed to only include the V3-V4 region of the 16S rRNA gene using the pcr.seqs command. Sequences that did not align to this region were culled from further

analysis and the alignments were trimmed so that the sequences fully overlapped the same alignment coordinates. Potentially chimeric sequences were removed using Uchime, as implemented in mothur. The remaining sequences were classified using the classify.seqs command with the customized SILVA alignment release 119 as reference. Next, sequences were clustered into OTUs at 97% similarity using the default settings of the dist.seq and cluster commands respectively. The classify.otu algorithm was used to get a consensus taxonomy for each OTU. Finally, all SMC samples were rarefied to 1,000 sequences per sample and all skin swab samples were rarefied to 5,000 sequences per sample. The sequencing data that are connected to this article are uploaded to the Sequence Read Archive database with accession number SRP076831.

Statistical analyses

The Kolmogorov-Smirnov test was used to check the normality of data distribution. Precision analyses were performed by calculating the coefficient of variation for each of the four OTUs obtained from the SMC samples. The paired Wilcoxon signed-rank test was used to compare the coefficients of variation obtained using micPCR/NGS and traditional PCR/NGS (SPSS version 23, IBM Corporation).

ACKNOWLEDGEMENTS

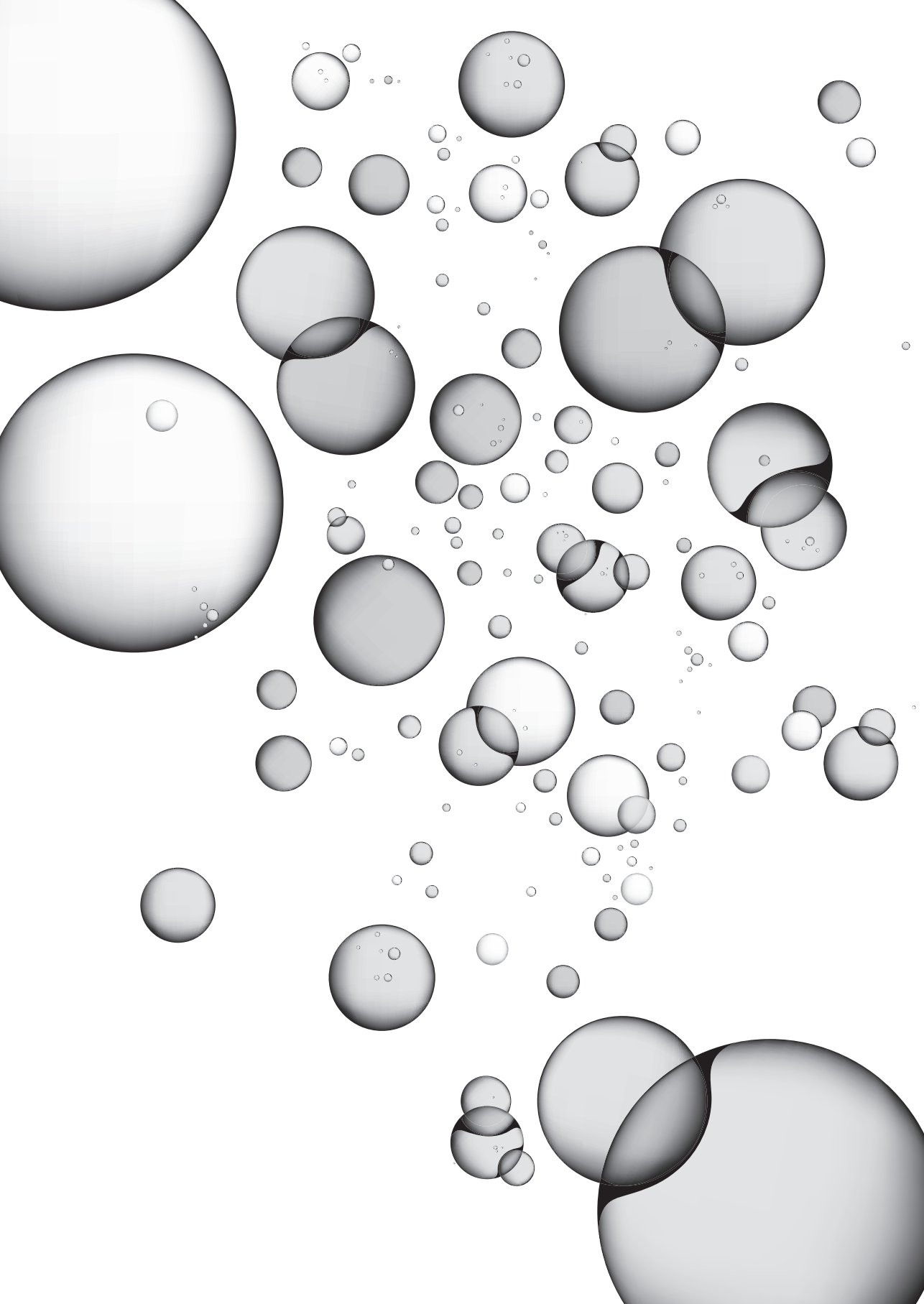
We would like to thank Sjoerd M. Euser, PhD and Wil A. van der Reijden, PhD for their help with the statistical analysis. We are also thankful to Bjorn L. Herpers, MD, PhD for organizing the collection of skin swab samples included in this study and Paul Badoux for characterizing the bacterial communities of these samples using bacterial culture. This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu).

SUPPLEMENTARY DATA

Supplementary information accompanies this paper at https://www.nature.com/articles/srep45536.

REFERENCES

1. Boers SA, Jansen R, Hays JP. Suddenly everyone is a microbiota specialist! *Clin Microbiol Infect* 2016; **22**: 581-582.
2. Brooks JP, Edwards DJ, Harwich MD Jr, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol* 2015; **15**: 66.
3. Boers SA, Hays JP, Jansen R. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Sci Rep* 2015; **5**: 14181.
4. 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.
5. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12**: 87.
6. Biesbroek G, Sanders EA, Roeselers G, et al. Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PloS one* 2012; **7**: e32942.
7. Meisel JS, Hannigan GD, Tyldsley AS, et al. Skin microbiome surveys are strongly influenced by experimental design. *J Invest Dermatol* 2016; **136**: 947-956.
8. Gao Z, Tseng CH, Pei Z, et al. Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci USA* 2007; **104**: 2927-2932.
9. van Rensburg JJ, Lin H, Gao X, et al. The human skin microbiome associates with the outcome of and is influenced by bacterial infection. *MBio* 2015; **6**: e01315-01315.
10. Yang S, Lin S, Kelen GD, et al. Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. *J Clin Microbiol* 2002; **40**: 3449-3454.
11. 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.
12. Stammler F, Glasner J, Hiergeist A, et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 2016; **4**: 28.
13. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014; **10**: e1003531.
14. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol* 2016; **26**: 330-335.
15. Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol* 2012; **12**: 158.
16. 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.
17. Mao DP, Zhou Q, Chen CY, et al. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol* 2012; **12**: 66.
18. 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.



Chapter 5

Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur

Saskia D. Hiltemann *
Stefan A. Boers *
Peter J. van der Spek
Ruud Jansen
John P. Hays
Andrew P. Stubbs

* These authors contributed equally to this study.

Submitted for publication.

ABSTRACT

Background

The determination of microbial communities using the mothur tool suite (https://www.mothur.org) is well established. However, mothur requires bioinformatics-based proficiency in order to perform calculations via the command-line. Galaxy is a project dedicated to providing a user-friendly web interface for such command-line tools (https://usegalaxy.org).

Results

We have integrated the full set of 125+ mothur tools into Galaxy as the Galaxy mothur Toolset (GmT) and provided a set of workflows to perform 'end-to-end' 16S rRNA gene analyses and integrate with third-party visualization and reporting tools. We demonstrate the utility of GmT by analysing the mothur MiSeq standard operating procedure (SOP) data set (https://www.mothur.org/wiki/MiSeq_SOP).

Conclusions

GmT is available from the Galaxy Tool Shed, and a workflow definition file and full Galaxy training manual for the mothur SOP have been created. A Docker image with a fully configured GmT Galaxy is also available.

KEY POINTS

- GmT provides a user-friendly interface to mothur by implementing mothur software in Galaxy
- A Galaxy workflow and full training manual for the mothur SOP are provided
- GmT provides integration with third-party visualization and reporting tools

FINDINGS

Introduction

16S rRNA gene profiling analysis can be achieved using an extensive array of sophisticated software including mothur,[1] QIIME,[2] MG-RAST,[3] and many more.[4] Whilst some of these applications have a graphical user interface (GUI) to provide access to these technologies for the research scientist, their use remains complex for non-bioinformaticians. In this respect, the Galaxy project was developed in order to simplify the use of complex command-line software tools.[5] Galaxy offers extensive support for both 16S rRNA gene-based and broader metagenomic analyses, with over 100 tools in the metagenomics section of the Galaxy Tool Shed, including QIIME,[2] KRONA,[6] PyNAST,[7] PICRUSt,[8] Kraken,[9] MetaPhlAn2,[10] HUMAnN2,[11] PrinSEQ,[12] Nonpareil,[13] VEGAN,[14] and many more.

mothur is an open-source application that was designed as a single piece of software capable of analysing and comparing microbial communities from 16S rRNA gene data derived from next-generation sequencing (NGS). The creators of mothur did not only provide an extensive set of tools, but also a collection of standard operating procedures (SOPs) that detail the recommended analytical protocol for different types of input data.

The latest version of mothur consists of over 125 components, lending it great flexibility, but at the same time, great complexity. To address this challenge, we have integrated the full set of 125+ mothur components into Galaxy that are collectively called the 'Galaxy mothur Toolset (GmT)'. To simplify usage of GmT we provide the full workflow definition files, usage of which shields the end-user from the full complexities of the analysis. By simultaneously providing access to all the individual components present in mothur as separate tools, expert users and bioinformaticians retain the ability to utilize the full flexibility of mothur by creating custom workflows or by modifying or extending our workflows to fit their use-case.

GmT also leverages Galaxy's collections framework to enable easy analysis of large numbers (many thousands) of samples at once. Many mothur components support parallel computing, and the Galaxy tools will utilize the maximum amount of processing power allotted to them by the instance administrator. As part of GmT, datatypes were also contributed to the Galaxy core codebase to facilitate the handling of mothur-specific datatypes within Galaxy. Furthermore, a Galaxy data manager was also created for the automatic installation and configuration of reference datasets utilized by the mothur tool suite. And lastly, a Galaxy interactive environment (GIE)[15] for Phinch[16] was also developed.[17]

GmT includes tools to produce standard file formats, such as the BIOM format[18] to facilitate interoperability with these downstream analysis components. Where no clear file standards exist, GmT provides custom tools for conversion of mothur datatypes to other tools (e.g. the taxonomy-2-krona tool). This allows for integration with third-party

tools such as PICRUSt for prediction of functional content, or visualisation tools such as Phinch, KRONA, and certain QIIME components. The mothur tools also natively support incorporation of some third-party analysis tools, such as UCHIME and ChimeraSlayer for chimera detection or VSEARCH for clustering, which are also available in GmT.

The Galaxy Training Network (GTN) is a network of people and groups that present Galaxy and Galaxy-based training around the world. The GTN has created a central repository for Galaxy training materials.[19] In order to further facilitate the use of GmT to end-users, we have contributed training materials to the GTN that illustrate how to run mothur's MiSeq SOP within Galaxy.[20] This work has also been incorporated in a larger-scale framework to easily and quickly explore microbiota data in a reproducible and transparent environment.[21]

Purpose of this work

The work performed and described in this technical note has four objectives. First to provide end-users and bioinformaticians with easy access to all the mothur tools as the GmT. Second is to provide open-access online training material to demonstrate/complete the mothur SOP in Galaxy. Third is to deliver an 'end-to-end' workflow for the mothur SOP in Galaxy that is available for upload to any Galaxy that has the GmT installed. Fourth is to provide a summarization of results in a web report using the iReport Galaxy tool.[22] Our aim is to provide 16S rRNA gene NGS analysis tools and awareness on how to use them in a format that supports FAIR data principles.[23]

Worked example

To illustrate the utility of our toolkit, we present results on example data below. GmT is designed to take short-read 16S rRNA gene NGS data as input and to output a dynamic web report for prokaryotic taxonomical classification using the Galaxy platform. A GmT workflow follows essentially a four-step process:

(i.) *Data upload*. The Galaxy platform provides the users with standard data upload functionality for single and multi-sample datasets.

(ii.) *Collection creation*. For multi-sample and/or paired-end datasets a Galaxy collection must be created in the Galaxy interface. Here datasets can also be assigned to groups. Galaxy will make intelligent suggestions for pairings of datasets based on the file names.

(iii.) *16S rRNA gene analysis*. mothur has been wrapped as a tool suite in Galaxy. Required steps included for a full 'end-to-end' 16S rRNA gene sequencing analysis consist of read-pair merging (mothur command: make.contigs), trimming of primer sequences (trim.seqs), additional quality control (screen.seqs), alignment of sequences to a (customized) reference alignment (align.seqs, screen.seqs), removal of chimeric sequences (chimera.uchime), classifying sequences using a Bayesian

classifier in combination with a reference database such as SILVA or GreenGenes (classify.seqs), and clustering of sequences into operational taxonomic units (OTUs) at a predefined percentage – usually 97 percent – of similarity (dist.seqs, cluster, and classify.otu) (Figure 1).

(iv.) *Experimental summary and reporting*. iReport in combination with KRONA[6] is used to deliver an HTML report in Galaxy. The iReport consists of multiple tabs to group results topically (e.g. taxonomy, rarefaction, diversity, quality control) and is highly customizable and easily tailored to an end-user's specific use-case. The entire report may be downloaded from the Galaxy interface to be viewed or shared offline. To compare the output from a single experiment or across multiple experiments we utilized Phinch,[16] a dynamic web application which uses BIOM-formatted files to explore and analyse biological patterns in 16S rRNA gene NGS datasets.

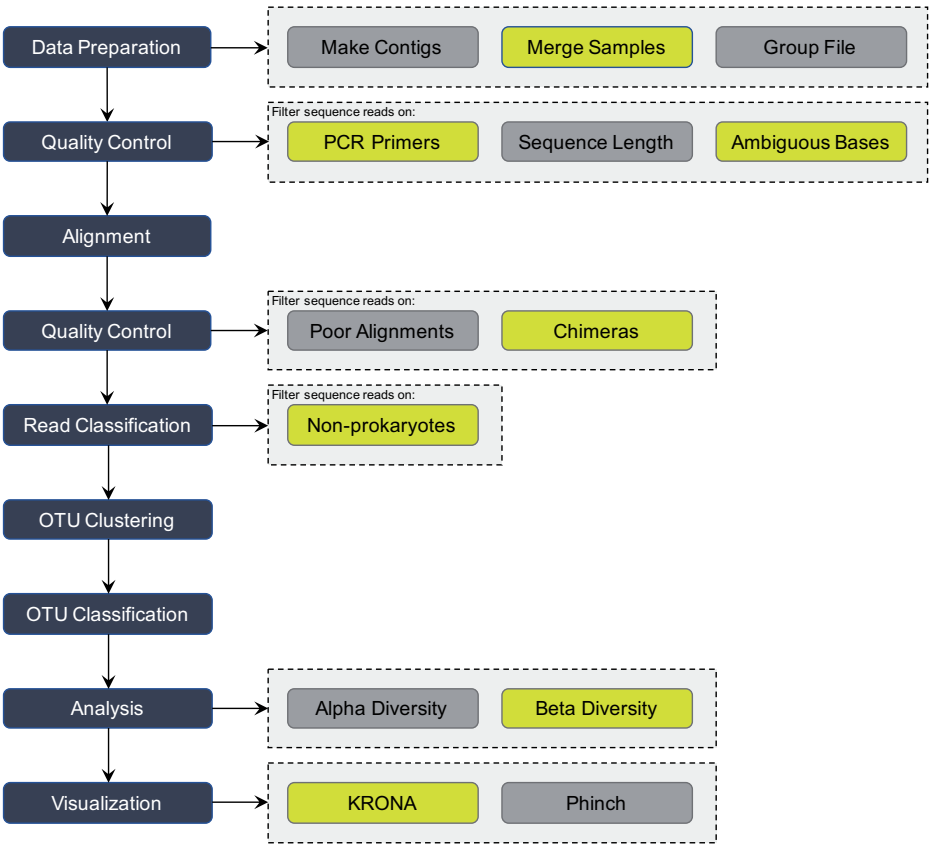


Figure 1. Conceptual view of the GmT mothur MiSeq SOP pipeline.

METHODS

Handling large datasets

Large-scale analyses have become the norm in the field, both large in disk space as in the number of files, and this can pose a challenge for analysis. For large files, Galaxy offers the option of uploading via FTP rather than web transfer. The introduction of the concept of 'collections' in Galaxy has enabled users to analyse datasets consisting of a large number of files (>100K) as easily as they would a single file.

Galaxy mothur Toolset

Many mothur components support parallelization, and our Galaxy wrappers will run these components with the maximum number of CPUs allotted to them by the Galaxy administrator. In order to diagnose potential failures, Galaxy outputs the full standard and error logs, which the users can inspect. Furthermore, we have contributed mothur datatype definitions to the Galaxy core code, meaning that the users will be protected from inputting the wrong datasets and thus reduce the number of errors they will make with the tools. All tools in GmT use only conda dependencies, making their installation in Galaxy a painless experience that requires nothing more than a single press of a button.

The mothur tool wrappers have been submitted to the Intergalactic Utilities Commission (IUC) tool repository,[24] and are available from the Galaxy Tool Shed.[25] The IUC is a group of community members dedicated to developing and upholding Galaxy tool development best practices and guidelines, thus by contributing our tools to this repo we ensure that the tools will be well-maintained. A metagenomics Galaxy flavour is available which contains all components presented here.[26] The full mothur suite has also been installed to Galaxy's main server.[27]

KRONA visualization

KRONA is a data viewer which provides the ability to interactively explore hierarchical data.[6] A Galaxy KRONA wrapper that works directly on mothur data formats was developed for this project.

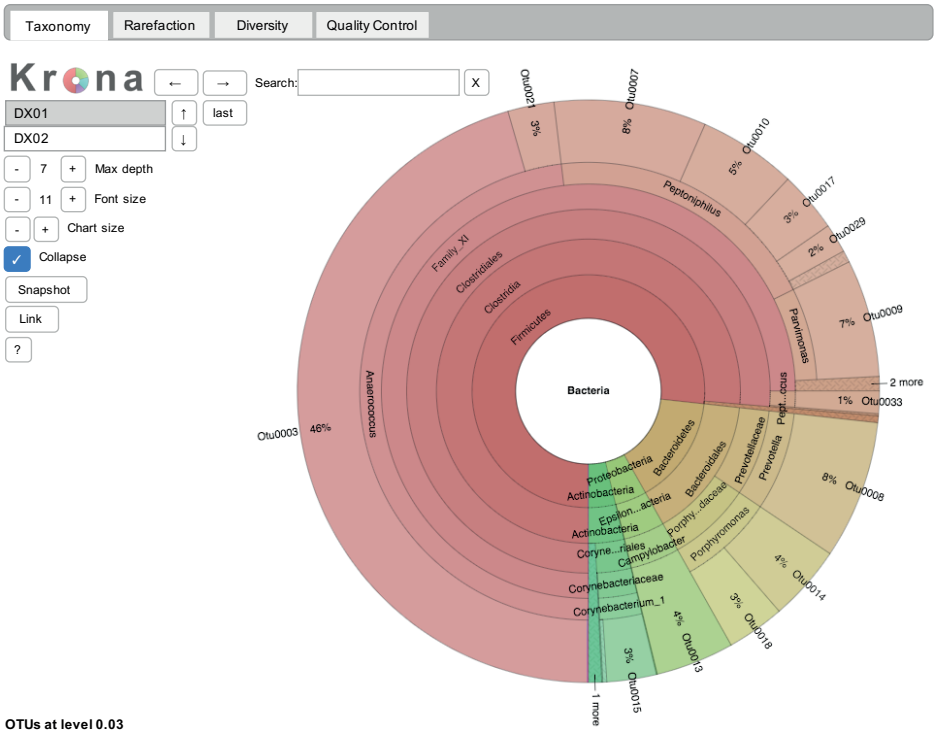
Phinch visualization

Galaxy offers integration with Phinch BIOM format viewer[16] in two ways; as a Galaxy interactive environment (GIE) developed in the context of this project,[17] and more recently also as an external display application hosted by the Galaxy team.

iReport summarization

To facilitate the evaluation of 16S rRNA gene sequencing analysis results, integration with the iReport tool are also provided.[22] This tool creates a web report to present the

analysis results in an organized fashion and provides links to external resources such as BLAST searches (Figure 2).



OTUs at level 0.03

Show 10 entries Search:

OTU	DX01	DX02	Kingdom	Phylum	Class	Order	Family	Genus	BLAST
Otu0002	0	8,265	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Morganella	send to BLAST
Otu0003	6,565	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Anaerococcus	send to BLAST
Otu0004	0	6,607	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella	send to BLAST
Otu0005	0	3,868	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	send to BLAST
Otu0006	0	502	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	send to BLAST
Otu0007	1,219	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus	send to BLAST
Otu0008	1,108	0	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	send to BLAST
Otu0009	939	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Parvimonas	send to BLAST
Otu0010	787	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus	send to BLAST
Otu0011	16	0	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	send to BLAST

Showing 1 to 10 of 37 entries Previous Next

Figure 2. Example iReport. This web report contains the interactive KRONA visualization, the (multi-sample) OTU table, rarefaction plots, diversity calculations, differential abundance analysis, and an extensive overview of the quality control measurements taken during the analysis. iReports are highly customizable and can be easily tailored to fit specific use-cases and end-user needs.

AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

- Project name: Galaxy mothur Toolset (GMT)
- Project home page: https://github.com/erasmusmc-bioinformatics/galaxy-mothur-toolset
- Tool shed repository: https://toolshed.g2.bx.psu.edu/view/iuc/suite_mothur/768c2e48b706
- Training manual: http://galaxyproject.github.io/training-material
- GmT Docker image: https://quay.io/shiltemann/galaxy-mothur-toolset:16.07
- Galaxy Metagenomics Docker Flavour (Docker): https://quay.io/repository/shiltemann/galaxy-metagenomics, https://github.com/shiltemann/galaxy-metagenomics
- Phinch interactive environment: https://github.com/shiltemann/phinch-galaxy-ie
- Operating system(s): Unix (Platform independent with Docker)
- License: GNU GPL v3

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

The data presented here to illustrate our work is the same data used in the training manual, and is available from Zenodo.[28]

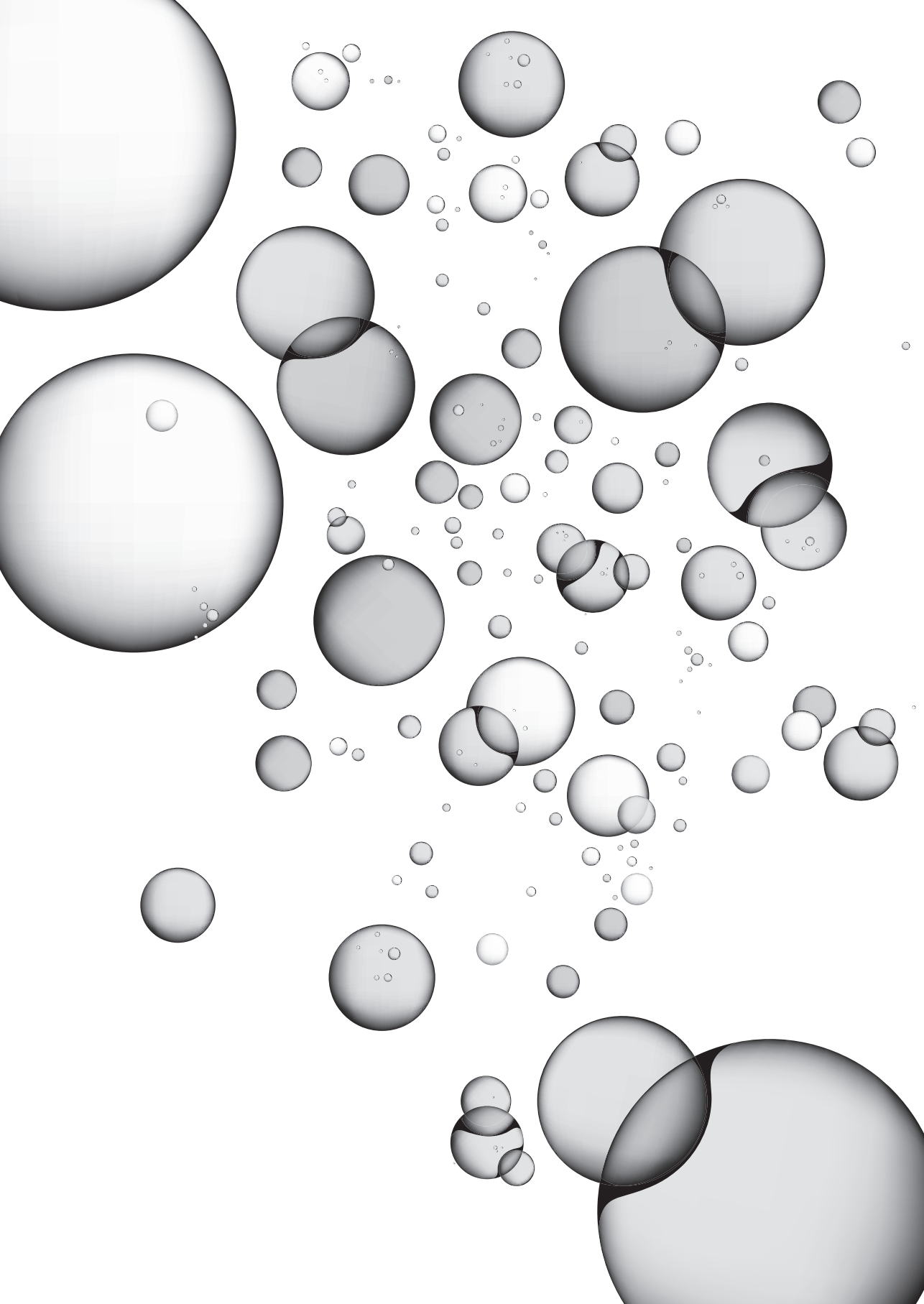
ACKNOWLEDGEMENTS

The authors would like to thank Jim Johnson and the many other contributors and reviewers of the mothur tool wrappers, including everybody who contributed to the development of these tools within the context of the Galaxy metagenomics contribution fest organized by the Galaxy community's Intergalactic Utilities Commission (IUC), a group of community members dedicated to developing and upholding Galaxy tool development best practices and guidelines.[29] We would also like to thank the Galaxy Training Network for providing the infrastructure and valuable feedback to share our training materials. This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu).

REFERENCES

1. 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.
2. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 2010; **7**: 335-336.
3. Glass EM, Wilkening J, Wilke A, et al. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* 2010; **1**: pdb–prot5368.
4. Oulas A, Pavloudi C, Polymenakou P, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform Biol Insights* 2015; **9**: 75.
5. Afgan E, Baker D, Van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016; **44**: W3–W10.
6. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomics visualization in a web browser. *BMC Bioinformatics* 2011; **12**: 385.
7. Caporaso JG, Bittinger K, Bushman FD, et al. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 2009; **26**: 266-267.
8. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013; **31**: 814.
9. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15**: R46.
10. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015; **12**: 902.
11. Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 2012; **8**: e1002358.
12. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011; **27**: 863-864.
13. Rodriguez-R LM, Konstantinidis KT. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 2013; **30**: 629-635.
14. Dixon P. VEGAN, a package of R functions for community ecology. *J Veg Sci* 2003; **14**: 927-930.
15. Eric R, Bjorn G, John C, et al. Galaxy Interactive Environments – a new way to interact with your data. In: Galaxy Community Conference; 2015.
16. Bik HM, Pitch Interactive. Phinch: an interactive, exploratory data visualization framework for -omic datasets. *bioRxiv* 2014: 009944.
17. Phinch Galaxy Interactive Environment: https://github.com/shiltemann/phinch-galaxy-ie.
18. The Biological Observation Matrix (BIOM) format: http://biom-format.org.
19. GTN Training Materials: https://training.galaxyproject.org.
20. Training materials for using GmT to run the mothur miseq SOP: https://galaxyproject.github.io/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html.
21. Batut B, Gravouil K, Defois C, et al. ASaiM: a Galaxy-based framework to analyze raw shotgun data from microbiota. *bioRxiv* 2017: 183970

22. Hiltemann S, Hoogstrate Y, van der Spek P, et al. iReport: a generalised Galaxy solution for integrated experimental reporting. *Gigascience* 2014; **3**: 19.
23. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3**: 160018.
24. IUC tool repository: https://github.com/galaxyproject/tools-iuc.
25. Galaxy Tool Shed: https://toolshed.g2.bx.psu.edu.
26. Metagenomics Galaxy Flavour: https://github.com/shiltemann/galaxy-metagenomics.
27. Galaxy Main server: https//usegalaxy.org.
28. Mothur MiSeq SOP Galaxy Tutorial Data: https://zenodo.org/record/800651.
29. Intergalactic Utilities Commission: https://galaxyproject.org/iuc.



Chapter 6

Development and evaluation of a culture-free microbiota profiling platform (MYcrobiota) for clinical diagnostics

Stefan A. Boers *
Saskia D. Hiltemann *
Andrew P. Stubbs
Ruud Jansen
John P. Hays

* These authors contributed equally to this study.

Eur J Clin Microbiol Infect Dis 2018; 37: 1081-1089.

ABSTRACT

Microbiota profiling has the potential to greatly impact on routine clinical diagnostics by detecting DNA derived from live, fastidious and dead bacterial cells present within clinical samples. Such results could potentially be used to benefit patients by influencing antibiotic prescribing practices, or to generate new classical-based diagnostic methods e.g. culture or PCR. However, technical flaws in 16S rRNA gene next-generation sequencing (NGS) protocols, together with the requirement for access to bioinformatics, currently hinders the introduction of microbiota analysis into clinical diagnostics. Here we report on the development and evaluation of an 'end-to-end' microbiota profiling platform (MYcrobiota), which combines our previously validated micelle PCR/NGS (micPCR/NGS) methodology with an 'easy-to-use', dedicated bioinformatics pipeline. The newly designed bioinformatics pipeline processes micPCR/NGS data automatically and summarizes the results in interactive, but simple web reports. In order to explore the utility of MYcrobiota in clinical diagnostics, 47 clinical samples (40 'damaged skin' samples and 7 synovial fluids) were investigated using routine bacterial culture as comparator. MYcrobiota confirmed the presence of bacterial DNA in 37/37 culture-positive samples and detected bacterial taxa in 2/10 culture-negative samples. Moreover, 36/38 potentially relevant aerobic bacterial taxa and 3/3 mixtures of anaerobic bacteria were identified using culture and MYcrobiota, with the sensitivity and specificity being 95%. Interestingly, the majority of the 448 bacterial taxa identified using MYcrobiota were not identified using culture, which could potentially have an impact on clinical decision-making. Taken together, the development of MYcrobiota is a promising step towards the introduction of microbiota analysis into clinical diagnostic laboratories.

INTRODUCTION

The detection and identification and further characterization of pathogenic microorganisms is the major step in establishing appropriate (antibiotic) treatment for infectious diseases. However, the causative microorganism of an infection may not always be detected using current 'gold standard' culturing techniques. Further, most molecular-based detection methods e.g. PCR, require *a priori* knowledge of the potential pathogen before a test is performed. To overcome these limitations, the bacterial composition can be detected and genera identified using a culture-free, broad-range PCR strategy that targets the prokaryotic 16S rRNA gene followed by next-generation sequencing (NGS).[1] However, to date, 16S rRNA gene NGS methods to profile microbial compositions have been focussed on research questions mostly, with only a few studies having evaluated the utility of 16S rRNA gene NGS methods for clinical microbiology.[2,3] Currently, the utilisation of 16S rRNA gene NGS methods within routine clinical diagnostics has been hindered by issues relating to the generation of PCR artefacts (e.g. chimera formation and PCR competition) and the susceptibility of 16S rRNA gene NGS methods to DNA contamination that is derived from the laboratory environment and/or the reagents/consumables used. These limitations hinder the standardization of current 16S rRNA gene NGS methods to such an extent that non-identical microbiota results may be obtained when repeatedly analysing the same sample.[4]

Recently, the authors published a micelle PCR/NGS (micPCR/NGS) methodology that limits the formation of chimeric sequences and prevents PCR competition via the clonal amplification of targeted 16S rRNA gene molecules.[5] In addition, the micPCR/NGS methodology allows for the utilization of an internal calibrator (IC) to calculate the number of 16S rRNA gene copies for each individual operational taxonomic unit (OTU) present within a (clinical) sample, which conveniently enables the subtraction of contaminating bacterial DNA via the quantification of 16S rRNA gene copies within negative extraction control (NEC) samples. The authors showed that the microbiota results obtained using micPCR/NGS possess a much higher accuracy (precision and trueness) compared to traditional 16S rRNA gene NGS protocols and that the ability to determine and subtract contaminating 16S rRNA gene copies, results in contamination-free quantitative microbiota profiles – with a limit of detection (LOD) of only 25 16S rRNA gene copies per OTU.[6] This low LOD allows for the detection of bacterial OTUs at very low abundances, or can confirm the absence of 16S rRNA gene copies in culture-negative results. Based on these findings, the authors suggested that the micPCR/NGS protocol could possess distinct advantages when processing clinical samples for microbiota profiling compared to traditional (semi-quantitative) 16S rRNA gene NGS methods that remain vulnerable to false-positive results (e.g. chimeric sequences or contaminant DNA) and inaccurate measurements of the OTU relative abundances in polymicrobial clinical samples due

to template-specific variations in PCR efficiencies (i.e. PCR competition). However, the analysis of 16S rRNA gene NGS data depends on the use of bioinformatics tools that are complex for non-bioinformatics educated technicians/clinicians to utilize and the required bioinformatics skills are nowadays mostly absent in clinical diagnostic laboratories.

In this publication, we designed an 'easy-to-use' bioinformatics pipeline to determine bacterial taxa from 16S rRNA gene sequences that together with the micPCR/NGS strategy is part of an 'end-to-end' microbiota profiling platform (MYcrobiota). The bioinformatics pipeline enables the full analyses of the NGS data obtained, from raw sequence files to final web reports that summarizes the quantitative microbiota profiles, without the knowledge of command-line scripts that would normally be required by 16S rRNA gene NGS users. As a proof of principle, we explored the utility of MYcrobiota for use in the clinical diagnostic laboratory by processing a total of 47 clinical samples and then comparing the results to conventional 'gold standard' culture results. The samples tested included 40 specimens that were obtained from a variety of damaged skin conditions for which a polymicrobial biomass was expected, and an additional 7 specimens, obtained from patients who were suspected of having (prosthetic) joint infections, for which a low bacterial biomass was expected.

MATERIALS AND METHODS

Ethics statement

An acknowledged national ethics committee from the Netherlands (Medisch Ethische Toetsingscommissie Noord-Holland, http://www.metc.nl) approved the study protocol (M015–021) and all experiments were performed on leftover material of the included clinical samples in accordance with the relevant guidelines and regulations. The national ethics committee waived the need for participant consent as all data were anonymized and analysed retrospectively under code.

Sample collection and study design

This study was performed retrospectively using 47 clinical samples obtained from 47 subjects. The results obtained by routine bacterial culturing methods had been used to guide patient treatment and care. In this study, we re-analysed these samples using MYcrobiota and compared the results to the initial outcome of the culture results. The 47 samples included in this study were derived from wounds (22), ulcers (10), abscesses (5), puss (1), erysipelas (1), erythema (1), and 7 synovial fluids obtained from patients with suspected (prosthetic) joint infections.

Routine bacterial culture

All samples were cultured according to standard laboratory protocols performed in our laboratory and stored at −80°C for subsequent MYcrobiota analysis. The routine bacterial culture methods included a 48-h incubation at 35°C on tryptic soy agar plates with 5% sheep blood (TSASB, Oxoid), colistin aztreonam blood agar plates (CAP, Oxoid) and cystine lactose electrolyte deficient agar plats (CLED, Oxoid) under aerobic conditions, a 48-h incubation at 35°C on chocolate agar plates with Vitox supplement (CHOCV, Oxoid) under 5% CO_2 conditions, and a 48-h incubation at 35°C on TSASB under anaerobic conditions. All Gram-negative rods, beta-haemolytic streptococci, *Staphylococcus aureus*, *Staphylococcus lugdunensis* and anaerobic bacteria cultured were reported as potentially relevant bacteria, of which the identification of aerobic bacteria was obtained using MALDI-TOF mass spectrometry (Bruker). Note that in this study we did not focus on optimizing culturing methods to increase the sensitivity of the culture results and the routine bacterial culture methods used may not 100% efficient for culturing the bacteria that were detected with MYcrobiota.

Micelle PCR and NGS

DNA was extracted from all 47 samples using the High Pure PCR Template Preparation Kit (Roche) according to the manufacturer's instructions. In addition, DNA from the accompanying elution buffer was extracted as a NEC at the same time in order to allow the subtraction of contaminating bacterial DNA after NGS processing. The total number of 16S rRNA gene copies within each DNA extract was measured using a 16S rRNA gene quantitative PCR (qPCR) according to Yang et al.,[7] after which each DNA extract was normalized to contain either 10,000, or <1,000 16S rRNA gene copies per microliter. A synthetic microbial community (SMC) sample, containing 10,000 16S rRNA gene copies of *Moraxella catarrhalis* (ATCC 25240), *Staphylococcus aureus* (ATCC 43300), *Haemophilus influenzae* (ATCC 10211), and *Clostridium perfringens* (ATCC 12915), was processed with each batch of clinical samples as a positive control (PC) sample. Prior to amplification by micPCR, 1,000, or 100 16S rRNA gene copies of *Synechococcus* DNA were added respectively as IC to the normalized DNA extracts containing 10,000, or <1,000 16S rRNA gene copies per microliter. One hundred 16S rRNA gene copies of *Synechococcus* DNA were also added to the NEC DNA extract. The IC was used to express the resulting OTUs as a measure of 16S rRNA gene copies by the use of a correction factor (sample OTU copies = sample OTU reads x (initial IC copies/IC OTU reads)) as previously validated elsewhere.[6]

16S rRNA gene amplicon library preparation using micPCR was performed as previously published,[6] but we utilized a different micPCR primer set that made it possible to replace the former Roche 454 NGS platform with the Illumina MiniSeq platform. In this study, micPCR amplification was performed using modified 515F (5′-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG TGY CAG CMG CCG CGG TAA-3′) and 806R (5′-

GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGA CTA CNV GGG TWT CTA AT-3') primers that amplified the V4 regions of 16S rRNA genes as recommended for Illumina NGS and which incorporated universal sequence tails at their 5' ends to allow for a two-step amplification strategy. During the second round of amplification, dual indices and Illumina sequencing adapters were attached using the Nextera XT Index kit (Illumina). Paired-end sequencing of the 16S rRNA gene amplicon library was performed using the MiniSeq system in combination with the 2x150 bp MiniSeq System High-Output Kit (Illumina), after which FASTQ-formatted sequences were extracted from the MiniSeq machine for downstream analysis. We utilized the micPCR/NGS approach to process all samples, including the NEC and the PC, in triplicate in order to increase accuracy and to correct for contaminating bacteria DNA derived from the laboratory environment as previously described.[6]

Bioinformatics pipeline

The bioinformatics pipeline designed during this study consists of 23 well-established mothur tools (v.1.36)[8] and an additional 9 custom-made tools developed by the authors that have been integrated and combined in Galaxy as a full analysis service to deliver 16S rRNA gene analysis for micPCR/NGS experiments. Essentially, we have incorporated the functionality of mothur in Galaxy, which is a project dedicated to simplify the use of complex command-line bioinformatics tools (such as mothur) using a user-friendly web interface,[9-11] and added new calculator tools to allow for a completely automatic processing of quantitative micPCR/NGS data. Importantly, the bioinformatics pipeline presents the microbiota results, together with an extensive overview of the quality control measurements performed during the micPCR/NGS data analysis, to the user in an organized fashion via an interactive web report. The complete workflow of the bioinformatics pipeline is visualized in Figure 1. All the tools required for the bioinformatics pipeline can be found in Galaxy's Tool Shed (https://toolshed.g2.bx.psu.edu). A workflow definition file can be downloaded from GitHub (https://github.com/ErasmusMC-Bioinformatics/MYcrobiota) and may be imported to any Galaxy platform, thereby offering the required set of bioinformatics tools. For more information on how to install and use this pipeline, please refer to the documentation in GitHub (https://github.com/ErasmusMC-Bioinformatics/MYcrobiota).

Quantitative PCR methods

The total bacterial biomass within each DNA extract was measured using a 16S rRNA gene qPCR that targets the 16S rRNA gene V5-V7 region, which is a different region of the 16S rRNA gene compared to MYcrobiota.[7] Therefore, the 16S rRNA gene qPCR is a complementary technique that enables the validation of the MYcrobiota process when determining the total number of 16S rRNA gene copies. For this, CT-values were related

to a serial dilution of the previous calibrated and normalized SMC sample that contained mixed and equimolar concentrations of four bacterial species and ranged from a total of 100 to 10,000 16S rRNA gene copies per PCR. In addition, the *S. aureus* specific biomass was assessed within each DNA extract using a *S. aureus* qPCR that employs a *S. aureus* specific marker as described by Martineau et al.[12] Here, CT-values were related to a serial dilution of only the calibrated *S. aureus* (ATCC 43300) DNA stock that ranged from a total of 10 to 10,000 copy numbers of the Martineau fragment. The PCRs were performed in 10 µL reaction volumes using the LightCycler 480 Probes Master (Roche) with the addition of 0.5 µM and 1.0 µM of each PCR primer for the 16S rRNA gene and *S. aureus* qPCRs respectively. Also, 0.25 µM of a Fam-labelled probe was added for the real-time detection of the 16S rRNA gene amplification and 1x Resolight Dye (Roche) was added to the *S. aureus* qPCR in order to measure the *S. aureus* DNA amplification. All PCRs were performed on a LightCycler 480 instrument (Roche), using the following conditions: initial denaturation at 95°C for 5 minutes followed by 45 cycles of PCR, with cycling conditions of 5 seconds at 95°C, 10 seconds at 55°C, and 30 seconds at 72°C.

Availability of data and materials

The datasets generated and analysed during the current study are available in the Sequence Read Archive repository with accession number SRP109023.

RESULTS

Development of an 'easy-to-use' bioinformatics pipeline

In order to analyse 16S rRNA gene NGS data obtained using micPCR/NGS, we designed a Galaxy-based bioinformatics pipeline for use in clinical diagnostics. This workflow is largely based on the well-established standard operating procedure (SOP) defined by the creators of mothur.[13] We have adapted the SOP to our specific use-case by integrating several custom-made tools that allows for the subsampling of large datasets, the averaging over multiple technical replicates, converting the number of obtained sequence reads per OTU to 16S rRNA gene copies per OTU via the use of an IC, and correction for contaminating bacterial DNA via the use of NECs. All results are presented to the user as a single, interactive web report in Galaxy using the iReport tool.[14] The iReport was designed to visualize the resultant microbiota profiles using KRONA,[15] list quantitative microbiota profiles in OTU tables (with the microbial load per OTU reported as 16S rRNA gene copies), summarize results of diversity calculators, and provide an extensive overview of the quality control measurements during the analysis. Importantly, the iReport is relatively small in size (~ 6 MB per sample for our datasets) that enables easy sharing and storage of 16S rRNA gene NGS results (Figure 1).

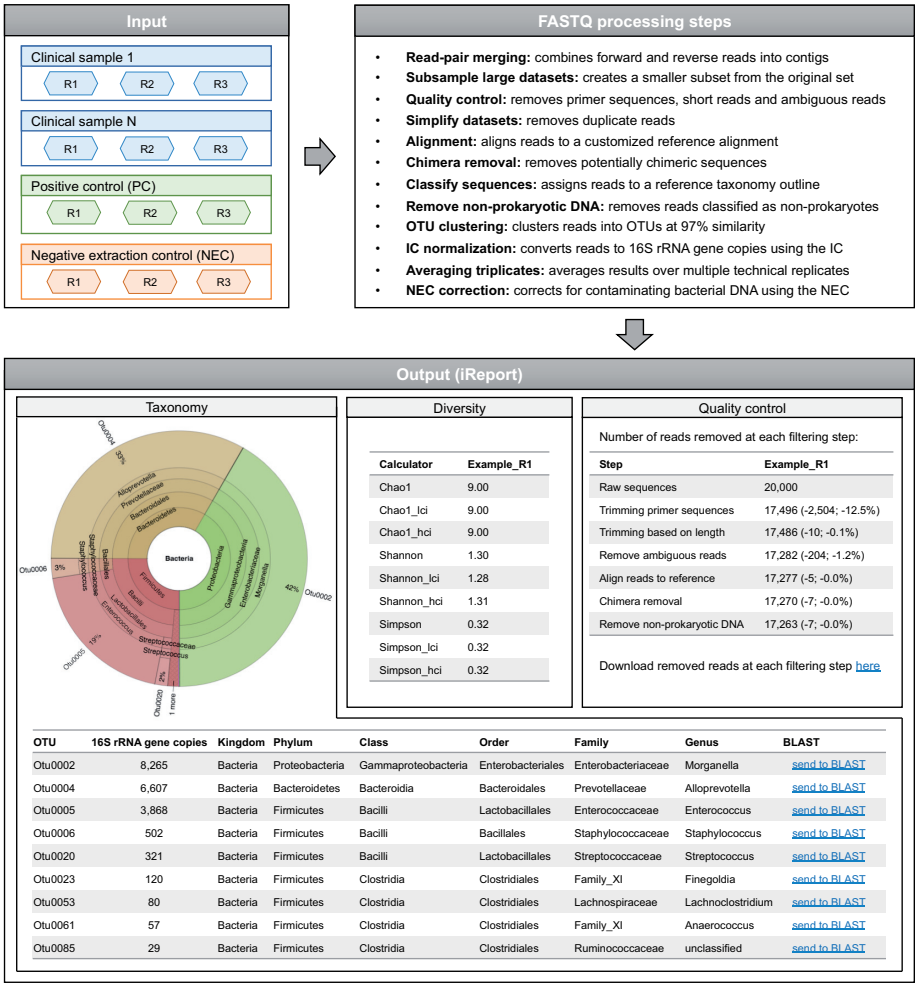


Figure 1. Schematical overview of the bioinformatics pipeline. FASTQ-formatted sequences obtained from triplicate experiments using micPCR/NGS (R1, R2 and R3) are automatically processed via the use of 32 (mothur) tools that have been integrated and combined in Galaxy as an 'end-to-end' analysis service. The results obtained per sample (average of triplicate results) are presented to the user in a single, interactive iReport that consist of 3 tabs. The taxonomy tab visualizes and lists the resultant microbiota profiles. The diversity tab summarizes the results of 3 diversity calculators (Chao1, Shannon and Simpson). The quality control tab provides an extensive overview of the quality control measurements during the analysis.

Validation of the MYcrobiota process

As shown in Figure 2, MYcrobiota results obtained from the PC that was profiled in three independent experiments showed a median value of only a 1.3-fold (± 0.2) difference between the measured 16S rRNA gene copies per bacterial species and the expected 10,000 16S rRNA gene copies per bacterial species present in the PC. In addition, comparisons between the measured 16S rRNA gene copies determined in actual clinical

samples using MYcrobiota compared to qPCR results revealed an average of only a 1.5-fold (± 0.5) and a 1.3-fold (± 0.4) difference for the total bacterial biomass and the *Staphylococcus* OTU-specific biomass respectively. Of note, 10 of the 47 clinical samples included in this study resulted in culture-negative results and the absence of bacterial DNA in these samples was confirmed with both qPCR and MYcrobiota methods. Also, one discrepant sample was detected that showed a 20-fold higher abundance of staphylococci detected by MYcrobiota compared to qPCR. This result can be explained by the presence of *S. aureus* and *S.* non-*aureus* within this sample. In fact, the *S. aureus* qPCR showed a 100% specificity compared to *S. aureus* culture-positive results and indicates the presence of *S.* non-*aureus* bacteria within 7 additional samples in which the *Staphylococcus* OTU was detected using MYcrobiota but no *S. aureus* could be cultured. Taken together, these data demonstrate the accuracy of the MYcrobiota process and the ability to incorporate quantitative results obtained from additional (species-specific) qPCRs.

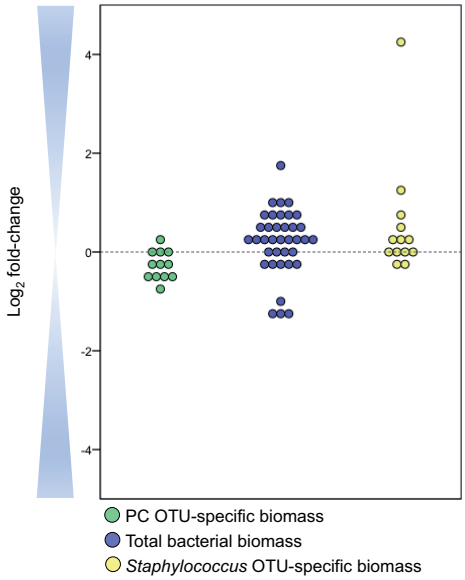


Figure 2. Accuracy of 16S rRNA gene copy determination using MYcrobiota. The expected number of 16S rRNA gene copies within the positive control (PC) was compared to the measured number of 16S rRNA gene copies using MYcrobiota (green dots). The PC contained 10,000 16S rRNA gene copies of four different bacterial species and was processed in three independent MYcrobiota experiments. The indirect estimation of the total bacterial biomass within 37 clinical samples using MYcrobiota was compared to the total 16S rRNA gene copies measured directly using a 16S rRNA gene qPCR (blue dots). The *Staphylococcus* OTU-specific biomass from 13 *S. aureus* culture-positive samples was compared to the *S. aureus* biomass detected directly using a *S. aureus*-specific qPCR (yellow dots). In order to compare the number of *S. aureus* genome copies estimated using qPCR to the number of 16S rRNA gene copies detected using MYcrobiota, the estimated *S. aureus* genome copies were first multiplied by a factor of 6 to correct for differences in copy numbers of the Martineau fragment and the 16S rRNA gene present on the *S. aureus* genome. The calculated differences between methods were plotted using a binary logarithmic scale.

Comparing MYcrobiota results to routine bacterial culture

In order to explore the utility of MYcrobiota in the field of clinical diagnostics, we processed a total of 47 clinical samples and compared the results to routine bacterial culture. All bacterial genera detected using culture and MYcrobiota are reported per sample in Table 1. Using standard bacterial culture techniques, our laboratory detected a total of 38 potentially relevant aerobic bacterial genera within 25 clinical samples and obtained a positive culture of a mixture of anaerobic bacteria in 3 samples. No bacteria were cultured from 10 samples and an additional 10 samples resulted in the growth of bacteria that were all presumed to be commensal flora. In contrast, using MYcrobiota we detected a total of 448 bacterial operational taxonomic units (OTUs) in 39 samples of which 337 OTUs (75%) could be identified as anaerobic bacterial genera that were detected in 21 samples. No bacterial DNA was measured in 8 out of 10 culture-negative samples. The sensitivity for bacterial culture detection by MYcrobiota was determined at 100% and the specificity at 83% using culture as 'gold standard'.

The majority of bacterial genera identified with culture were also identified using MYcrobiota. As shown in Table 2, MYcrobiota detected 36 of all 38 aerobic bacteria cultured on a genus-level taxonomy and confirmed the growth for anaerobic bacteria in 3 samples (sensitivity 95%; specificity 95%). Important to note, the two discrepant bacterial genera were measured using the micPCR/NGS strategy, but below the technique's LOD of 25 16S rRNA gene copies per OTU. In contrast, the vast majority of bacterial genera identified with MYcrobiota were presumed to belong to the commensal flora using culture or were not cultured at all (Table 1). These additional taxa include potential pathogens such as the *Kingella* OTU that was detected from a synovial fluid sample obtained from a juvenile patient that was not detected using culture and was confirmed using a *Kingella kingae*-specific PCR.

Table 1. Bacterial genera identified from 47 clinical samples using routine bacterial culture and MYcrobiota.

Sample	Routine bacterial culture	MYcrobiota
01_U	Commensal flora (1+)	**Anaerobic bacteria (346,300)**, *Corynebacterium* (10,725)
02_U	Commensal flora (2+)	*Staphylococcus* (941)
03_W	Commensal flora (1+)	**Anaerobic bacteria (263)**, *Streptococcus* (33), *Staphylococcus* (25)
04_U	*Pseudomonas* (3+), *Staphylococcus* (2+)	*Pseudomonas* (4,706), *Staphylococcus* (848), ***Enterococcus* (135)**, **Anaerobic bacteria (102)**
05_U	*Proteus* (2+), *Enterobacteriaceae** (2+), *Streptococcus* (1+), Commensal flora (1+)	**Anaerobic bacteria (8,271)**, *Proteus* (3,510), *Streptococcus* (632), *Enterobacteriaceae** (333)
06_U	Commensal flora (1+)	*Moraxella* (8,947), *Corynebacterium* (734)
07_W	*Enterobacteriaceae* (1+)	*Enterobacteriaceae** (5,386), ***Bacillus* (44)**

Table 1. Bacterial genera identified from 47 clinical samples using routine bacterial culture and MYcrobiota. (continued)

Sample	Routine bacterial culture	MYcrobiota
08_W	Negative	Negative
09_W	Commensal flora (1+)	**Anaerobic bacteria (523)**, *Staphylococcus* (31)
10_A	Anaerobic bacteria (2+), *Pasteurella* (2+), *Streptococcus* (2+)	Anaerobic bacteria (3,704,750), *Pasteurella* (242,250), *Streptococcus* (28,625)
11_W	*Enterobacteriaceae** (3+), *Staphylococcus* (3+)	*Enterobacteriaceae** (3,420,786), ***Acinetobacter* (1,126,632)**, *Staphylococcus* (32,760)
12_A	*Enterobacteriaceae** (2+), *Streptococcus* (2+)	*Enterobacteriaceae** (18,046), *Streptococcus* (6,409), ***Enterococcus* (67)**
13_Es	Commensal flora (1+)	*Staphylococcus* (344), **Anaerobic bacteria (150),** *Dermabacteraceae** (93), *Haemophilus* (64), *Corynebacterium* (53)
14_W	Commensal flora (1+)	*Staphylococcus* (31)
15_U	*Staphylococcus* (4+)	*Staphylococcus* (17,035)
16_U	*Enterobacteriaceae** (3+), *Stenotrophomonas* (2+), Commensal flora (2+), *Proteus* (1+), ***Pseudomonas* (1+)**	*Enterobacteriaceae** (828,310), *Proteus* (250,670), *Stenotrophomonas* (11,760)
17_W	*Staphylococcus* (1+), Commensal flora (1+)	*Staphylococcus* (4,886)
18_W	*Staphylococcus* (3+), Commensal flora (2+)	*Staphylococcus* (141,120), *Corynebacterium* (4,959)
19_W	*Streptococcus* (2+), *Staphylococcus* (1+)	*Streptococcus* (114,257), *Staphylococcus* (44,772), ***Corynebacterium* (8,749), Anaerobic bacteria (897)**
20_W	*Enterobacteriaceae** (3+), ***Staphylococcus* (2+)**	*Enterobacteriaceae** (4,574,310)
21_U	Commensal flora (2+)	*Moraxella* (1,066,608), *Acinetobacter* (142,155), *Pseudomonas* (30,051), **Anaerobic bacteria (30,051),** *Corynebacterium* (23,976), *Alkanindiges* (2,187)
22_W	*Staphylococcus* (2+), Commensal flora (1+)	*Staphylococcus* (105,648)
23_Et	*Staphylococcus* (2+), Commensal flora (2+)	*Staphylococcus* (14,803), *Corynebacterium* (66)
24_U	*Staphylococcus* (3+), *Streptococcus* (3+), Commensal flora (2+)	*Staphylococcus* (231,756), **Anaerobic bacteria (96,740),** *Streptococcus* (15,904), *Enterococcus* (1,680)
25_W	*Staphylococcus* (3+), Commensal flora (2+)	*Staphylococcus* (23,175), *Corynebacterium* (15,488), **Anaerobic bacteria (1,271)**
26_W	Commensal flora (1+)	*Staphylococcus* (4,142), **Anaerobic bacteria (101)**, *Corynebacterium* (94), *Streptococcus* (47)
27_A	*Streptococcus* (1+)	**Anaerobic bacteria (1,062,060),** *Streptococcus* (5,490), ***Treponema* (3,435), *Gemella* (1,425), *Mycoplasma* (870), *Tannerella* (720)**
28_W	Commensal flora (2+), *Streptococcus* (2+)	**Anaerobic bacteria (114,004),** *Streptococcus* (43,208)
29_A	*Streptococcus* (1+)	*Streptococcus* (12,225), **Anaerobic bacteria (3,384), *Gemella* (299), *Enterococcus* (295), *Haemophilus* (221), *Capnocytophaga* (156), *Granulicatella* (122), *Neisseria* (119), *Rothia* (52), *Lautropia* (35)**
30_W	Negative	Negative

Table 1. Bacterial genera identified from 47 clinical samples using routine bacterial culture and MYcrobiota. (continued)

Sample	Routine bacterial culture	MYcrobiota
31_U	*Acinetobacter* (2+), *Enterobacteriaceae** (2+), Commensal flora (2+)	*Acinetobacter* (518,396), *Stenotrophomonas* (423,320), *Enterobacteriaceae** (12,046), *Corynebacterium* (5,928), *Bordetella* (4,636), *Brevibacterium* (988)
32_W	*Staphylococcus* (2+), Commensal flora (1+)	**Anaerobic bacteria (251,692)**, *Streptococcus* (30,408), *Staphylococcus* (8,960)
33_W	*Staphylococcus* (1+), Commensal flora (1+)	*Staphylococcus* (466), **Anaerobic bacteria (171),** *Streptococcus* (105), *Acinetobacter* (84), *Corynebacterium* (41)
34_W	*Staphylococcus* (3+)	*Staphylococcus* (218,141)
35_W	Commensal flora (2+)	*Staphylococcus* (5,121), **Anaerobic bacteria (769)**, *Roseomonas* (40)
36_W	Anaerobic bacteria (3+), Commensal flora (1+)	Anaerobic bacteria (493,183), *Streptococcus* (1,045)
37_W	*Streptococcus* (2+)	*Streptococcus* (11,457)
38_W	Anaerobic bacteria (3+)	Anaerobic bacteria (830,531)
39_P	*Streptococcus* (2+)	*Streptococcus* (10,277,376)
40_A	Negative	**Anaerobic bacteria (94,633)**, ***Enterobacteriaceae**** **(2,944)**, ***Streptococcus* (44)**, ***Thalassospira* (36)**
41_S	Negative	Negative
42_S	Negative	Negative
43_S	Negative	Negative
44_S	Negative	Negative
45_S	Negative	Negative
46_S	Negative	Negative
47_S	Negative	***Kingella* (25)**

Samples were derived from wounds (W), ulcers (U), abscesses (A), puss (P), erysipelas (Es), erythema (Et), and suspected joint infections (S). Cultured bacteria other than Gram-negative rods, beta-haemolytic streptococci, *S. aureus*, *S. lugdunensis* and anaerobic bacteria were reported as commensal flora. The semi-quantitative culture results are presented as 1+, 2+, 3+, or 4+, depending on which quadrants demonstrate bacterial growth. The presence of anaerobic bacteria was reported as either a positive or negative result. Bacterial species and OTUs detected using culture and MYcrobiota respectively are grouped at the genus level to compare results. Red shades indicate bacterial genera that were only identified by culture and blue shades indicate bacterial genera that were only identified by MYcrobiota (with 'commensal flora' culture results representing a positive detection signal for any kind of aerobic bacterial OTU identified by MYcrobiota). The number of 16S rRNA genes measured using MYcrobiota are indicated between brackets. (*) Several bacterial genera that belong to the *Enterobacteriaceae* and *Dermabacteraceae* families could not be differentiated at a 97% similarity level using MYcrobiota.

Table 2. Comparison of the cultured bacterial taxa to MYcrobiota results.

Bacterial taxa	Number of positive samples		Sensitivity	Specificity
	Routine bacterial culture	MYcrobiota		
Acinetobacter	1	4	100%	98%
*Enterobacteriaceae**	7	8	100%	98%
Pasteurella	1	1	100%	100%
Proteus	2	2	100%	100%
Pseudomonas	2	2	67%	100%
Staphylococcus	14	20	93%	100%
Stenotrophomonas	1	2	100%	100%
Streptococcus	10	16	100%	97%
Anaerobic bacteria	3	21	100%	71%
Total	**41**	**76**	**95%**	**95%**

The culture results are restricted to genus-level classifications in order to compare the OTUs detected using MYcrobiota to the culture-based results. The presence of anaerobic bacteria was reported as either a positive or negative result. 'Commensal flora' culture results were interpreted as a positive detection signal for any kind of aerobic bacterial OTU identified by MYcrobiota to perform specificity calculations. (*) Several bacterial genera that belong to the *Enterobacteriaceae* family could not be differentiated at a 97% similarity level using MYcrobiota.

DISCUSSION

In this study, we developed and explored the utility of an 'end-to-end' microbiota profiling platform (MYcrobiota) – consisting of our previously published 16S rRNA gene sequencing methodology (micPCR/NGS) in combination with an 'easy-to-use' bioinformatics pipeline – to investigate human samples for the clinical diagnostics laboratories. The bioinformatics pipeline designed during this study allows for a fully automated sequence interpretation of 16S rRNA gene NGS data that is obtained using the validated micPCR/NGS protocol without the need for advanced bioinformatics skills that are often unavailable in the clinical diagnostic laboratories. The MYcrobiota results are presented using (interactive) visualizations and tables, including an overview of all removed sequences during the analysis that allows for a manual evaluation of the quality measurements pre-installed within the bioinformatics pipeline. Moreover, connections of OTU representative sequences to the external NCBI database are available and can be used to ensure that the taxonomic identification of bacterial genera is correct.[16] Importantly, the summarizing reports are relatively small in size and storage of these files enables the traceability of patient test results that is required for clinical diagnostic laboratories according to quality requirements.

Using MYcrobiota, we processed a total of 47 clinical samples and compared the results to routine bacterial culture. Our results showed that the majority of bacteria

identified with culture were also identified with MYcrobiota, but the majority of bacterial taxa identified with MYcrobiota were not identified using culture. Many of the additional bacterial taxa identified using MYcrobiota are obligate anaerobes that were commonly detected as a large component of the microbial population in samples obtained from damaged skin sites, which is consistent with previous studies.[17,18] Indeed, it is well-known that anaerobic bacteria are able to cause serious and life-threatening infections but are often overlooked due to their requirement for appropriate methods of collection, transportation and cultivation.[19] Therefore, the culture-free MYcrobiota profiling platform can play an important role in the identification of the bacteriological aetiology of anaerobic infections, or any other infections caused by fastidious microorganisms. Of note, it could be argued that the development of extensive culture techniques (so-called 'culturomics') may eventually facilitate the successful culture of supposedly 'non-culturable' microbial isolates.[20]

In addition to the accurate detection and identification of bacterial OTUs within clinical samples, MYcrobiota also provide the relative abundances in combination with the absolute abundances for each detected bacterial OTU. This feature allows clinicians to obtain a comprehensive overview of the microbial composition of the clinical sample so that each quantified bacterial OTU, as well as the bacterial community as a whole, might be taken into account in clinical decision-making. Additionally, MYcrobiota allows for the removal of contaminating DNA from environmental sources in order to accurately and reliably investigate very low bacterial biomass, or no bacterial biomass, clinical samples.[6] For example, MYcrobiota confirmed the absence of 16S rRNA gene copies in 8 of the 10 samples that generated culture-negative results. The two discrepant samples contained either anaerobic bacteria or low amounts of the fastidious *Kingella* bacterium respectively. The ability to confirm culture-negative results improves the reliability of culture-negative diagnostic results. Additionally, the ability of MYcrobiota to detect bacterial OTUs at very low abundances makes MYcrobiota a suitable method to investigate normally sterile body sites, such as synovial fluids, cerebrospinal fluids, blood samples, etc. It should be noted however, that the authors are aware of the fact that the development of MYcrobiota is only a first step in the transition of microbiota research into actual clinical diagnostics. Extensive clinical and financial validation studies will be needed in order to validate and justify the routine introduction of molecular microbiota profiling methods into clinical diagnostic laboratories.

In conclusion, the stepwise development of MYcrobiota paves the way to introduce quantitative microbiota profiling into the clinical diagnostic laboratory. The method provides an highly accurate and comprehensive overview of the microbial composition of clinical samples or, alternatively, confirms the absence of 16S rRNA gene copies in culture-negative samples, using a standardized and validated 16S rRNA gene NGS workflow. Despite some shortcomings e.g. lack of species identification and the inabil-

ity to provide detailed information on antibiotic susceptibility, our data illustrates that MYcrobiota has promising applications in the field of clinical diagnostics and warrants investment in future studies to accurately evaluate the clinical relevance of 16S rRNA gene NGS results in clinical samples.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu).

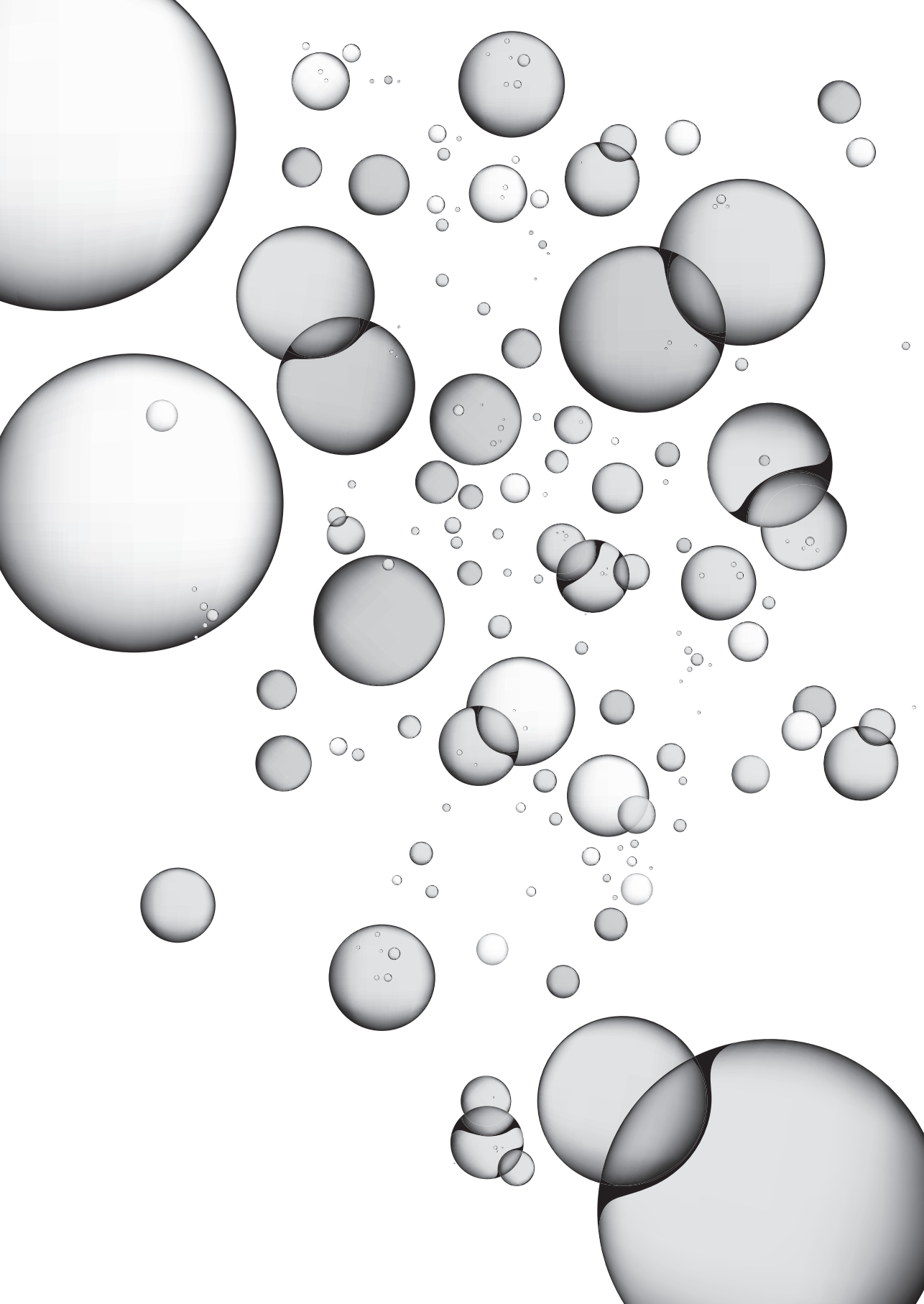
REFERENCES

1. Fournier PE, Raoult D. Prospects for the future using genomics and proteomics in clinical microbiology. *Annu Rev Microbiol* 2011; **65**: 169-188.
2. Rhoads DD, Wolcott, RD, Sun Y, et al. Comparison of culture and molecular identification of bacteria in chronic wounds. *Int J Mol Sci* 2012; **13**: 2535-2550.
3. Salipante SJ, Sengupta DJ, Rosenthal C, et al. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PloS one* 2013; **8**: e65226.
4. 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.
5. Boers SA, Hays JP, Jansen R. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Sci Rep* 2015; **5**: 14181.
6. Boers SA, Hays JP, Jansen R. Novel micelle PCR-based method for accurate, sensitive and quantitative microbiota profiling. *Sci Rep* 2017; **7**: 45536.
7. Yang S, Lin S, Kelen GD, et al. Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. *J Clin Microbiol* 2002; **40**: 3449-3454.
8. 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.
9. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005; **15**: 1451-1455.
10. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* 2010; **89**: 19.10.1-19.10.21.
11. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010; **11**: R86.
12. Martineau F, Picard FJ, Roy PH, et al. Species-specific and ubiquitous-DNA-based assays for rapid identification of *Staphylococcus aureus*. *J Clin Microbiol* 1998; **36**: 618-623.
13. 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.
14. Hiltemann S, Hoogstrate Y, van der Spek P, et al. iReport: a generalised Galaxy solution for integrated experimental reporting. *Gigascience* 2014; **3**: 19.
15. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 2011; **12**: 385.
16. Boers SA, Jansen R, Hays JP. Suddenly everyone is a microbiota specialist. *Clin Microbiol Infect* 2016; **22**: 581-582.
17. Price LB, Liu CM, Melendez JH, et al. Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota. *PloS one* 2009; **4**: e6462.

18. Smith K, Collier A, Townsend EM, et al. One step closer to understanding the role of bacteria in diabetic foot ulcers: characterising the microbiome of ulcers. *BMC Microbiol* 2016; **16**: 54.

19. Brook I. Clinical review: bacteremia caused by anaerobic bacteria in children. *Crit Care* 2002; **6**: 205-211.

20. 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.



Chapter 7

Detection of bacterial DNA in septic arthritis samples using the MYcrobiota platform

Stefan A. Boers
Linda Reijnen
Bjorn L. Herpers
John P. Hays
Ruud Jansen

J Clin Rheumatol 2018; in press.

ABSTRACT

Background/Objective

Bacterial septic arthritis is considered a medical emergency that may lead to disability or death. While the majority of these infections are described to be caused by gram-positive bacteria, clinicians should be aware of less common bacterial causes of septic arthritis that are not detected by routine bacterial culture strategies. Therefore, we investigated 23 joint fluid samples that were obtained from 19 patients with suspected bacterial septic arthritis using a culture-free 16S rRNA gene next-generation sequencing (NGS) platform (MYcrobiota) and compared the results to routine diagnostic testing.

Methods

In this cross-sectional descriptive study, all samples were collected over a period of three months in 2017 and processed retrospectively using our previously validated and published MYcrobiota platform without prior knowledge of their culture, PCR or traditional sequencing results.

Results

All joint fluid samples tested were found to be culture-negative and MYcrobiota confirmed the absence of bacterial operational taxonomic units (OTUs) in 13/23 samples. However, MYcrobiota detected the presence of either an *Enterococcus*, *Kingella*, *Parvimonas*, *Prevotella*, *Ruminococcus*, *Turicella*, or *Ureaplasma* OTU in the other 10 samples. Four out of seven OTUs detected by MYcrobiota confirmed the additional diagnostic test results (i.e. blood culture results and molecular test results) that has led to an effective targeted antibiotic treatment for four patients.

Conclusions

The accurate detection of bacterial OTUs using MYcrobiota greatly improves the identification of the aetiology of bacterial septic arthritis compared to routine diagnostic testing.

Bacterial septic arthritis is an inflammation in native (non-prosthetic) joints with an incidence that ranges between 4-29 cases per 100,000 people per year, depending on population variables and pre-existing structural joint abnormalities.[1] Most infections are introduced into the joints as a result of bacteraemia, though the joints may also become infected directly via trauma or an infection around the joint.[2] In routine procedures, the diagnosis of bacterial septic arthritis is confirmed by detection of bacteria in joint fluid samples using culture-based techniques. However, a substantial proportion of joint fluid samples are culture-negative, even from patients with typical signs of septic arthritis, suggesting a role of 'difficult-to-culture' microorganisms (e.g. anaerobic or fastidious bacteria) in such clinical presentations.[1,3] The failure of bacterial culture will delay an effective antimicrobial treatment of the septic arthritis patients who may suffer the destruction of the joint tissue.[4] With this in mind, we investigated to what extent a culture-independent 16S rRNA gene next-generation sequencing (NGS) platform (MYcrobiota) could add to the microbiological diagnosis obtained by culturing of joint fluid samples. For this, we tested 23 joint fluids that were obtained from 19 patients with suspected bacterial septic arthritis using MYcrobiota and compared the results to routine bacterial cultures.

MYcrobiota is a consolidated tool that includes a validated, quantitative micelle PCR/NGS methodology and a dedicated bioinformatics pipeline that was specifically designed for use in clinical diagnostic laboratories.[5] The MYcrobiota platform enables the detection, quantification and characterization of bacterial DNA derived from live, fastidious, and dead bacterial cells present within clinical samples, and takes account of possible bacterial DNA contamination derived from laboratory reagents and/or the laboratory environment. This is achieved by quantifying contaminating 16S rRNA gene copies within negative extraction controls and subtracting these 16S rRNA gene copies from the quantitative microbiota profiles obtained from actual samples.[6] Therefore, MYcrobiota possess a much lower limit of detection compared to (semi-quantitative) conventional 16S rRNA gene NGS methods, which allows the detection of bacterial operational taxonomic units (OTUs) at very low abundances, or alternatively, can reliably confirm the absence of bacterial DNA in culture-negative joint fluids.

The joint fluid samples were collected from the Regional Laboratory of Public Health Kennemerland and processed retrospectively using MYcrobiota, with no prior knowledge of specimen culture results. An acknowledged national ethics committee from the Netherlands (Medisch Ethische Toetsingscommisie Noord-Holland, http://www.metc.nl) approved the study protocol (M015-021) and all experiments were performed in accordance with relevant national guidelines and regulations. The national ethics committee waived the need for participant consent as all data were anonymized and analysed retrospectively under code. The joint fluid samples were cultured directly on a blood-based agar (under aerobic and anaerobic conditions) and chocolate-based agar (under 5% CO_2

conditions), and after thioglycolate enrichment. The left-over specimens were stored at −80°C for subsequent MYcrobiota analysis that was performed as previously described.[5]

Bacterial culture of all joint fluid samples included in the study generated culture-negative results. In contrast, MYcrobiota revealed the presence of bacterial OTUs in 10/23 (43%) joint fluids and confirmed the culture-negative results in the remaining 13 samples. As shown in Table 1, the bacterial OTUs detected using MYcrobiota were classified as *Enterococcus*, *Kingella*, *Parvimonas*, *Prevotella*, *Ruminococcus*, *Turicella*, and *Ureaplasma* and confirmed the clinician's suspicion of bacterial septic arthritis in seven patients that were found to be culture negative. From two patients, we received and processed multiple joint fluid samples that resulted in identical results from each sample. Specifically, four distinct joint fluid samples taken from one patient all tested positive for a *Parvimonas* OTU and two joint fluid samples obtained from a different patient confirmed the absence of bacterial DNA in both of these culture-negative joint fluids.

Table 1. Seven different bacterial OTUs were detected from 23 joint fluid samples using MYcrobiota and confirmed the suspicion of bacterial septic arthritis in 7 patients (including serial samples from individual patients). All joint fluids tested negative using routine bacterial culture techniques.

Bacterial OTUs	Patients (#)	Samples (#)	16S rRNA gene copies/µL	Clinical decision
Enterococcus	1	1	6,600	NA
Kingella	1	1	25	On the basis of additional molecular testing, treatment was started with amoxicillin/clavulanic acid, directed against *Kingella*.
Parvimonas (4/4)	1	4	1,500 – 51,000	NA
Prevotella	1	1	71,000	On the basis of additional blood culture results, treatment was switched from flucloxacillin/ gentamycin to intravenous amoxicillin/clavulanic acid and later oral clindamycin, directed against *Prevotella*
Ruminococcus	1	1	192,000	On the basis of additional molecular testing, treatment was switched from vancomycin/ rifampin to intravenous penicillin, directed against *Ruminococcus*
Turicella	1	1	50	NA
Ureaplasma	1	1	1,200	On the basis of additional molecular testing, clindamycin was switched to doxycycline, directed against *Ureaplasma*
Negative	12	13	0	Empirical antibiotic treatment was stopped in one case. In the other cases, MYcrobiota data were not available during treatment or no empirical antibiotics were given.

NA: MYcrobiota data were produced in retrospective analysis and were not available during the clinical treatment period.

In all bacterial OTU positive cases, only a single bacterial genus was identified per sample with bacterial loads that ranged from 25 to 192,000 16S rRNA gene copies per microliter of DNA extract used. Importantly, although 16S rRNA gene copies can provide some relevant information on bacterial load (tens of copies versus thousands of copies for example), it should be noted that it is not possible to accurately translate 16S rRNA gene copies into an actual number of bacterial genomes or cells present within a sample. This is because different bacterial species may carry different copy numbers of 16S rRNA genes in their genomes and copy numbers for all bacterial species are not known.

All bacterial genera detected using MYcrobiota, except *Turicella*, have been previously described as unconventional pathogens in bacterial septic arthritis.[7-12] Although *Turicella otitidis* (the only *Turicella* species described) is usually isolated from ear exudates, this bacterium has been previously identified as a cause for bacteraemia in at least two independent cases.[13] Further, the main discrepancies between routine bacterial culture and MYcrobiota results could be explained by a difficulty in culturing anaerobic bacteria e.g. *Parvimonas* spp., *Prevotella* spp., and *Ruminococcus* spp., as well as fastidious bacteria e.g. *Kingella* spp. and *Ureaplasma* spp. Interestingly, some culture-negative joint fluid samples were found to contain DNA from 'easily-culturable' bacteria, including *Enterococcus* spp. and *Turicella* spp.

Although all routine bacterial cultures from all joint fluid samples included in this study generated culture-negative results, additional diagnostic testing – including blood cultures, specific PCRs, and 16S rRNA gene PCR/Sanger sequencing – was performed to confirm the initial culture results. However, these additional diagnostic tests revealed the presence of four different bacterial pathogens (*Kingella*, *Prevotella*, *Ruminococcus* and *Ureaplasma*) in four joint fluid samples, which led to the subscription of an effective targeted antibiotics treatment for four patients with bacterial septic arthritis (Table 1). These patients received antibiotic treatment for six weeks in a routine clinical setting, including routine clinical follow up of treatment outcome that showed clinical and laboratory evidence of successful treatment. The additional diagnostic test results were later confirmed by our retrospective MYcrobiota analysis in all four cases. However, MYcrobiota detected an additional three OTUs (*Enterococcus*, *Parvimonas* and *Turicella*) in three joint fluids that were not identified using additional diagnostic testing. Unfortunately, we were unable to find an explanation for these discrepancies. Nonetheless, these data illustrate that MYcrobiota is an accurate diagnostic platform and its results can be used to start, or switch to, targeted therapeutic strategies for patients presenting with suspected bacterial septic arthritis. Importantly, MYcrobiota provides this clinical relevant information without the dependence on 'traditional' molecular tests that require *a priori* knowledge of the likely pathogen present within the sample (i.e. specific PCRs), or those that can only process monomicrobial samples with a bacterial load that

exceeds the inevitable background DNA contamination derived from the experimental set-up (i.e. 16S rRNA gene Sanger sequencing methods).

Joint infection is generally secondary to haematogenous dissemination of bacteria from other sites within the human body. Interestingly, for two patients, we were able to causally link the bacterial OTU detected by MYcrobiota to recent invasive surgery and dental procedures. The first patient underwent gastrointestinal surgery after which DNA derived from the *Ruminococcus* bacterium – that normally resides within the gastrointestinal tract – was detected within his joint, and in a second patient we identified DNA derived from the odontogenic *Parvimonas* bacterium following a tooth extraction. Taken together, these data further support the need to consider unconventional bacteria as causes of bacterial septic arthritis, especially in those patients who recently underwent surgical procedures of the digestive tract or in the mouth.

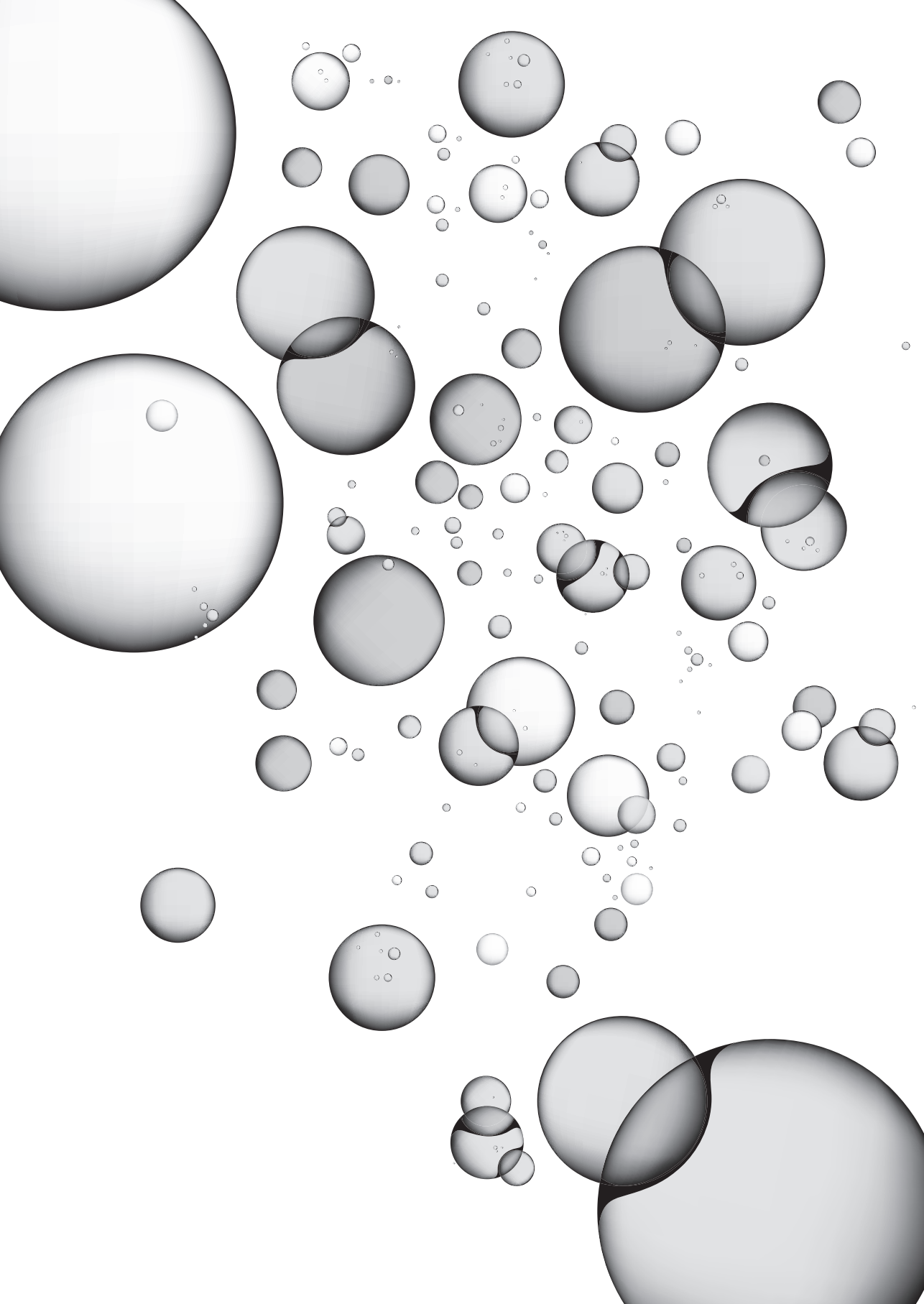
In conclusion, obtaining accurate quantitative microbiota profiles using MYcrobiota enables the identification of bacteria present within joint fluid samples that were not identified using routine bacterial culture strategies. Therefore, continued suspicion of bacterial septic arthritis despite culture-negative results should lead clinicians to consider the use of culture-free 16S rRNA gene NGS techniques, such as the MYcrobiota platform.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu).

REFERENCES

1. Mathews CJ, Weston VC, Jones A, et al. Bacterial septic arthritis in adults. *Lancet* 2010; **375**: 846-855.
2. Horowitz DL, Katzap E, Horowitz S, et al. Approach to septic arthritis. *Am Fam Physician* 2011; **84**: 653-660.
3. Ross JJ. Septic arthritis of native joints. *Infect Dis Clin North Am* 2017; **31**: 203-218.
4. Iliadis AD, Ramachandran M. Paediatric bone and joint infection. *EFORT Open Rev* 2017; **2**: 7-12.
5. Boers SA, Hiltemann SD, Stubbs AP, et al. Development and evaluation of a culture-free microbiota profiling platform (MYcrobiota) for clinical diagnostics. *Eur J Clin Microbiol Infect Dis* 2018; **37**: 1081-1089.
6. Boers SA, Hays JP, Jansen R. Novel micelle PCR-based method for accurate, sensitive and quantitative microbiota profiling. *Sci Rep* 2017; **7**: 45536.
7. Raymond NJ, Henry J, Workowski KA, Enterococcal arthritis: case report and review. *Clin Infect Dis* 1995; **21**: 516-522.
8. Williams N, Cooper C, Cundy P. *Kingella kingae* septic arthritis in children: recognising an elusive pathogen. *J Child Orthop* 2014; **8**: 91-95.
9. Baghban A, Gupta S. *Parvimonas micra*: a rare cause of native joint septic arthritis. *Anaerobe* 2016; **39**: 26-27.
10. Salman SA, Baharoon SA. Septic arthritis of the knee joint secondary to *Prevotella bivia*. *Saudi Med J* 2009; **30**: 426-428.
11. Titécat M, Wallet F, Vieillard MH, et al. *Ruminococcus gnavus*: an unusual pathogen in septic arthritis. *Anaerobe* 2014; **30**: 159-160.
12. George MD, Cardenas AM, Brinbaum BK, et al. *Ureaplasma* septic arthritis in an immunosuppressed patient with juvenile idiopathic arthritis. *J Clin Rheumatol* 2015; **21**: 221-224.
13. Greninger AL, Kozyreva V, Truong CL, et al. Draft genome sequence of *Turicella otitidis* TD1, isolated from a patient with bacteremia. *Genome Announc* 2015; **3**: e01060-15.



Chapter 8

Monitoring of microbial dynamics in a drinking water distribution system using the culture-free, user-friendly, MYcrobiota platform

Stefan A. Boers
Emmanuelle I. Prest
Maja Taučer-Kapteijn
Aleksandra Knezev
Peter G. Schaap
John P. Hays
Ruud Jansen

Sci Rep 2018; in press.

ABSTRACT

Drinking water utilities currently rely on a range of microbiological detection techniques to evaluate the quality of their drinking water (DW). However, microbiota profiling using culture-free 16S rRNA gene next-generation sequencing (NGS) provides an opportunity for improved monitoring of the microbial ecology and quality of DW. Here, we evaluated the utility of a previously validated microbiota profiling platform (MYcrobiota) to investigate the microbial dynamics of a full-scale, non-chlorinated DW distribution system (DWDS). In contrast to conventional methods, we observed spatial and temporal bacterial genus changes (expressed as operational taxonomic units - OTUs) within the DWDS. Further, a small subset of bacterial OTUs dominated with abundances that shifted across the length of the DWDS, and were particularly affected by a post-disinfection step. We also found seasonal variation in OTUs within the DWDS and that many OTUs could not be identified, even though MYcrobiota is specifically designed to reduce potential PCR sequencing artefacts. This suggests that our current knowledge about the microbial ecology of DW communities is limited. Our findings demonstrate that the user-friendly MYcrobiota platform facilitates culture-free, standardized microbial dynamics monitoring and has the capacity to facilitate the introduction of microbiota profiling into the management of drinking water quality.

INTRODUCTION

Drinking water distribution systems (DWDSs) are complex ecosystems where microorganisms can actively grow and reproduce. In fact, several studies have shown that the bacterial composition within drinking water (DW) samples are highly diverse,[1,2] with total cell concentrations that typically range from 1,000 to 100,000 bacterial cells per millilitre.[3] Importantly, the presence of (opportunistic) pathogenic bacteria within DW may present an emerging public-health risk.[4] Moreover, certain bacteria may also cause operational problems within DWDSs due to bacterial induced corrosion of iron pipes,[5] or produce metabolites that affect the taste, odour and colour of DW.[6] To reduce the risks of bacterial growth, DW utilities employ a combination of several treatment processes to minimize the number of microorganisms, as well as the microbial growth supporting nutrients, in the DW produced. These treatments commonly involve primary disinfection processes e.g. chlorination, ozonation, or UV/H_2O_2 advanced oxidation, which are also used for oxidation of natural organic matter, combined with filtration processes, such as active carbon, rapid or slow sand filtration, soil infiltration, or membrane filtration methods.[7] However, despite these efforts, eliminating all microorganisms and nutrients during treatment is impossible using current treatment processes, meaning that bacterial growth in the DWDS can occur and that DW utilities are compelled to monitor microbiological changes during DW treatment and distribution.

Current microbiological characterization of DW relies heavily on conventional culturing techniques such as heterotrophic plate counts (HPC) and selective plating for *Aeromonas* spp., *Legionella* spp. and faecal indicator bacteria. Importantly however, these culture-based methods are time-consuming and it is well-known that they often detect only a small proportion of the total microbial population present in DWDSs.[8] Therefore, multiple culture-independent methods have been developed over the past decade to overcome these limitations. Most notably, flow cytometry (FCM) has emerged as a promising tool for the rapid assessment of DW quality that enables the detection and quantification of relevant microbial dynamics throughout a DWDS with high sensitivity.[9] Although FCM is useful for counting the total and viable number of bacterial cells throughout a DWDS, it does not generate taxonomic information about the microbial composition within DWDSs. The identification and quantification of bacterial taxa is required in order to adequately evaluate the complex nature of microbial communities within DW samples and to determine whether potentially (opportunistic) pathogenic bacteria are present within the DWDS. In contrast to FCM, microbiota profiling methods using 16S rRNA gene next-generation sequencing (NGS) techniques are able to differentiate the composition of microbial communities on a taxonomic level and these methods have already been applied to DWDSs.[2,10,11] Importantly however, obtaining accurate 16S rRNA gene profiles requires careful consideration of (often overlooked) PCR

amplification biases and bacterial DNA contamination that can be introduced during the many steps of sample processing and sequencing.[12] These inevitable biases frustrate the accurate validation of current 16S rRNA gene NGS methods and, consequently, impedes the transition of these powerful tools from research into industrial diagnostic practice.

Recently, the authors published a validated micelle PCR/NGS (micPCR/NGS) methodology that significantly reduces PCR amplification biases in microbiota profiles via the clonal amplification of targeted 16S rRNA gene molecules.[13] The micPCR/NGS method drastically reduces chimera formation compared with traditional 16S rRNA gene NGS methods and prevents PCR competition due to unequal amplification rates of different 16S rRNA gene template molecules. This is of particular importance for the accurate analysis of microbial compositions within high-diversity samples, such as DW samples, as these samples are more vulnerable to chimera formation and PCR competition compared to low-diversity samples.[13] Further, by adding an internal calibrator (IC) to the micPCR/NGS methodology, we are able to quantify the absolute abundances of the bacterial operational taxonomic units (OTUs) detected within the samples under investigation, which enables the subtraction of any non-sample associated contaminating bacterial DNA that is invariably present in the laboratory environment and chemicals/reagents mixes used via the processing of negative extraction control (NEC) samples.[14] Therefore, the micPCR/NGS methodology possess a much higher accuracy and a lower limit of detection (LOD) compared with traditional 16S rRNA gene NGS methods that allows for the accurate detection of minor microbial variations within DWDSs. Another problem associated with the introduction of 16S rRNA gene NGS methodologies into industrial processes involves the complexity associated with establishing a 16S rRNA gene analysis workflow that can be operated by non-bioinformatics educated users. To overcome this limitation, we previously developed a dedicated bioinformatics pipeline that enables the full analyses of our quantitative micPCR/NGS data, without knowledge of command-line scripts that would normally be required. This user-friendly analytical workflow together with the validated micPCR/NGS strategy is part of the microbiota profiling platform named 'MYcrobiota', which has already shown promising applications in routine clinical microbiological diagnostics.[15]

In this study, the authors evaluated the utility of MYcrobiota for studying microbial dynamics within a non-chlorinated DWDS. We characterized microbial changes at consecutive locations along a full-scale DWDS in the Netherlands during a 5-month period with one-month intervals. The MYcrobiota results were compared to results of conventional microbiological analysis, including HPC, bacterial adenosine-triphosphate measurements (bATP), and intact bacterial cell concentrations assessed with FCM. Additionally, we were able to use the MYcrobiota results to detect and identify spatial and temporal bacterial dynamics within the DWDS under investigation that may be used to evaluate the success of DW treatment processes.

RESULTS

Total bacterial biomass variations observed within the DWDS

A total of 30 DW samples were collected from an operational DWDS. From July till November 2016, each month the DWDS was sampled at six consecutive locations and the DW at each location processed using MYcrobiota (Figure 1). The sequencing and quality parameters obtained are shown in Supplementary Table 1. In addition, HPC, bATP, and intact bacterial cell concentrations were measured and compared to the results obtained using MYcrobiota from all DW samples (Supplementary Table 2). These comparisons revealed that FCM and MYcrobiota were both able to detect microbiological changes within the DWDS that were not observed using HPC and bATP measurements. As shown in Figure 2, FCM and MYcrobiota detected a clear reduction of bacterial biomass between locations A and B (with the exception of DW samples taken in September), after which time, the bacterial biomass increased towards location C, followed by a more stable bacterial biomass concentration towards location F. However, the concentration of intact bacterial cells as detected by FCM was higher in 29 out of 30 DW samples than the concentration of 16S rRNA gene copies detected with MYcrobiota, with an average of a 4.9-fold (± 2.5) difference. In order to investigate the accuracy of the quantitative MYcrobiota data, we compared the number of 16S rRNA gene copies obtained using MYcrobiota with the number of 16S rRNA gene copies measured using a 16S rRNA gene quantitative PCR (qPCR).[16] This comparison revealed an average of only a 1.3-fold (± 0.3) difference between both quantitative methods, demonstrating the accuracy of MYcrobiota in determining the number of 16S rRNA gene copies in DW. In contrast, the accuracy of the FCM method could not be confirmed using complementary techniques in this study, although the standard error for FCM was previously estimated to be less than 5%.[17]

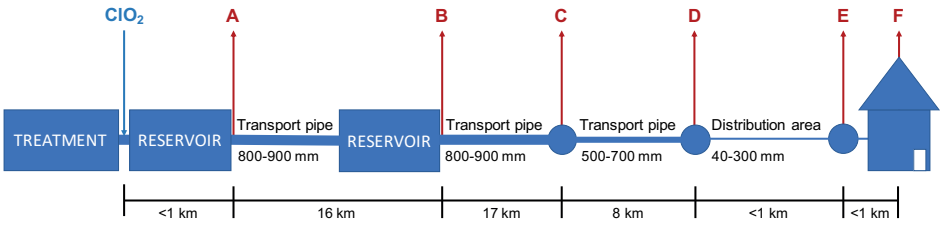


Figure 1. Sampling scheme. The sampling points, labelled A (effluent of water treatment plant, after chlorine dioxide dosage and after DW reservoir), B (effluent of intermediate pumping station and storage reservoir), C and D (transport pipelines), E (distribution pipeline), and F (tap water), are indicated by red arrows. The position of the chlorine dioxide (ClO_2) dosage is indicated by the blue arrow. The distance between sampling points are shown in kilometres (km).

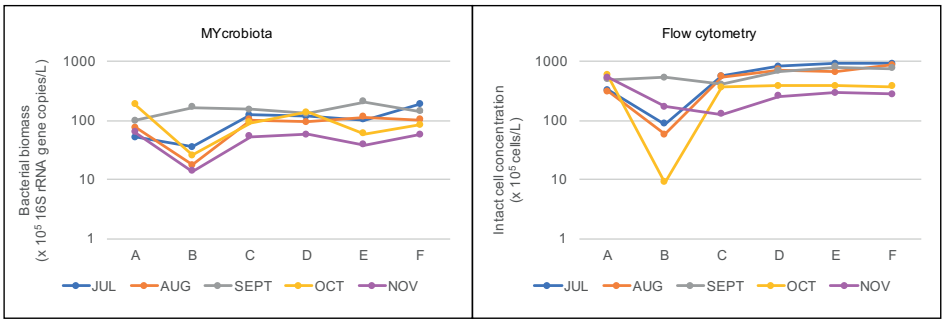


Figure 2. Variations of bacterial biomass measured at six consecutive locations within the DWDS. Coloured lines represent 16S rRNA gene copies (left) and intact bacterial cell concentrations (right) measured from 6 consecutive locations (A-F) over a 5-month period.

Spatial microbial dynamics observed within the DWDS

The use of MYcrobiota revealed that the DWDS investigated in this study consisted of a highly diverse bacterial environment with a median of 69 (± 16) OTUs – defined by 97% 16S rRNA gene sequence similarity – per DW sample (Supplementary Table 1). Only six of these OTUs were detected with a median relative abundance higher than 5% per sampled location. These dominating OTUs were classified as *Comamonadaceae*, *Deferrisoma*, *Gallionellaceae, Nitrospira*, *Parcubacteria*, and *Peribacterales*. Interestingly, DW samples taken at the start of the DWDS were dominated by the *Peribacterales* (median relative abundance: 23%), *Parcubacteria* (7%), and *Gallionellaceae* (7%) OTUs, whereas DW samples taken at the end of the DWDS were dominated by the *Comamonadaceae* (21%), *Deferrisoma* (11%), and *Nitrospira* (6%) OTUs as shown in Figure 3. In contrast, 138 distinct OTUs – representing 14 different prokaryotic phyla – were detected with a median relative abundance lower than 5% per sampled location. Only 30 of these OTUs (22%) could be classified to the genus taxonomic level using public 16S rRNA gene reference databases, indicating that the vast majority of prokaryotic species present within DWDSs remain unknown or poorly understood.

A shift in the microbial composition tended to occur between locations A and C within the DWDS. As shown in Figure 4, the number of 16S rRNA gene copies for the majority of OTUs decreased between locations A and B, including the dominating *Peribacterales*, *Parcubacteria*, and *Gallionellaceae* OTUs. In contrast, the number of 16S rRNA gene copies measured for most OTUs increased between locations B and C, including the dominant *Comamonadaceae*, *Deferrisoma*, and *Nitrospira* OTUs, but not the *Peribacterales, Parcubacteria*, and *Gallionellaceae* OTUs. Importantly, no significant increase of 16S rRNA gene copies per OTU was detected within the final stage of the DWDS (locations C to F), which includes the home plumbing system of the consumer – beyond the maintenance of the DW supply company.

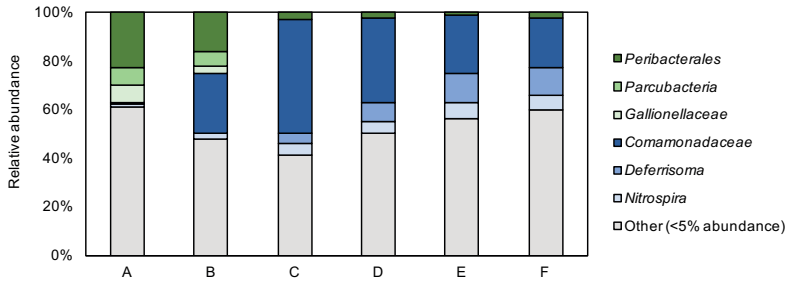


Figure 3. Variations in the relative abundances of six dominating OTUs detected at consecutive locations within the DWDS. The medians of 16S rRNA gene copies per OTU that were measured over a 5-month period are shown using 100% stacked bars for each consecutive location (A-F). OTUs with > 5% relative abundance at the start of the DWDS are shown in shades of green, whereas OTUs with > 5% relative abundance at the end of the DWDS are shown in shades of blue. All other OTUs with < 5% relative abundance per sampled location were grouped together to ease visualization and are shown in grey.

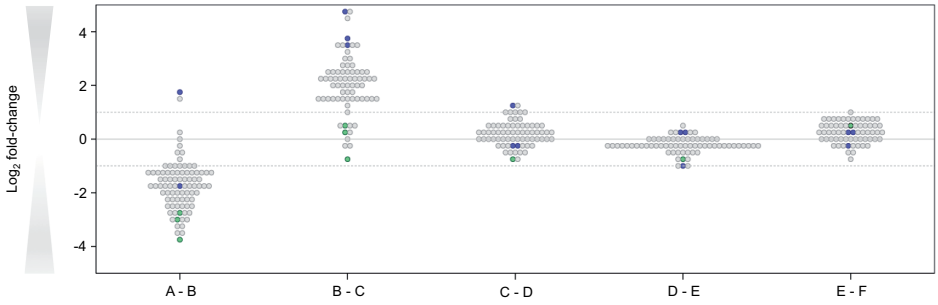


Figure 4. Differences in absolute abundances of OTUs detected at consecutive locations within the DWDS. OTU-level differences between DWDS locations (A-F) were calculated by dividing the number of 16S rRNA gene copies per OTU derived from two consecutive locations. For this, the medians of 16S rRNA gene copies per OTU that were measured over a 5-month period were used and their differences between two consecutive locations plotted using a binary logarithmic scale. OTUs with a relative abundance of < 5% per sampled location are shown as grey dots, whereas OTUs with > 5% relative abundance at location A (*Peribacterales, Parcubacteria*, and *Gallionellaceae*) are shown as green dots and OTUs with > 5% relative abundance at location F (*Comamonadaceae*, *Deferrisoma*, and *Nitrospira*) are shown as blue dots. Differences with more than a 2-fold increase/decrease (dotted lines) were considered as significant differences that cannot be explained by technical variations introduced during sample processing.

Temporal microbial dynamics observed within the DWDS

In order to investigate temporal trends in bacterial community structure, we compared the medians of 16S rRNA gene copies per OTU that were measured over the six consecutive DWDS locations to each month-of-sampling. From this dataset, only a single OTU – belonging to the genus of *Thioalkalispira* – could be detected during specific months only (July, August, and September, but not October and November). In contrast, the other six dominating OTUs described above were detected in all sampled months and followed the same overall dynamic trends regarding the increase and decrease in the ab-

solute abundance of these OTUs throughout the DWDS, although these trends became weaker as water temperatures decreased (Figure 5). Importantly, a single DW sample obtained from location B in September showed a sudden increase of *Comamonadaceae* 16S rRNA gene copies by a factor of 138 compared to location A. In fact, we measured an increase of 16S rRNA gene copies for another 42 out of 143 OTUs (30%) by at least a factor of 2 at location B in September, although their relative abundances remained below the threshold of 5%. These observations explain the deviating curve that was obtained when comparing the variations of total bacterial biomass within consecutive DW samples taken in September, namely an increase in bacterial abundance between locations A and B, while the abundance decreased at other months between these two locations (Figure 2).

OTU	Month / Temp.	A	B	C	D	E	F	R_s	Relative abundance (range)
Peribacterales	JUL / 17.9°C	123	52	64	11	10	9	-0.94 **	4% (1% - 26%)
	AUG / 19.4°C	72	7	6	0	0	0	-0.94 **	0% (0% - 11%)
	SEPT / 20.2°C	128	16	7	0	0	0	-0.94 **	0% (0% - 16%)
	OCT / 15.7°C	383	44	18	16	6	10	-0.94 **	2% (2% - 25%)
	NOV / 10.1°C	143	12	48	38	30	36	-0.37	11% (9% - 28%)
Parcubacteria	JUL / 17.9°C	73	35	24	5	0	0	-0.99 **	2% (0% - 16%)
	AUG / 19.4°C	45	5	0	0	0	0	-0.85 *	0% (0% - 7%)
	SEPT / 20.2°C	26	7	0	0	0	0	-0.85 *	0% (0% - 3%)
	OCT / 15.7°C	41	6	0	0	0	0	-0.85 *	0% (0% - 3%)
	NOV / 10.1°C	3	0	0	0	0	0	-0.66	0% (0% - 1%)
Gallionellaceae	JUL / 17.9°C	10	4	7	4	4	0	-0.82 *	1% (0% - 2%)
	AUG / 19.4°C	45	3	5	0	0	0	-0.88 *	0% (0% - 7%)
	SEPT / 20.2°C	37	4	0	0	0	0	-0.85 *	0% (0% - 5%)
	OCT / 15.7°C	40	2	0	0	0	0	-0.85 *	0% (0% - 3%)
	NOV / 10.1°C	5	0	0	0	0	0	-0.66	0% (0% - 1%)
Comamonadaceae	JUL / 17.9°C	7	25	274	318	252	512	0.83 *	33% (1% - 42%)
	AUG / 19.4°C	9	8	73	26	29	41	0.60	6% (1% - 12%)
	SEPT / 20.2°C	4	593	518	401	697	416	0.31	40% (1% - 47%)
	OCT / 15.7°C	11	12	535	224	104	111	0.49	23% (1% - 67%)
	NOV / 10.1°C	5	65	55	40	30	46	0.03	12% (1% - 52%)
Deferrisoma	JUL / 17.9°C	0	0	21	51	50	97	0.93 **	4% (0% - 8%)
	AUG / 19.4°C	0	0	30	64	91	80	0.93 **	8% (0% - 14%)
	SEPT / 20.2°C	0	0	31	50	95	61	0.93 **	4% (0% - 6%)
	OCT / 15.7°C	0	0	4	20	13	21	0.93 **	1% (0% - 4%)
	NOV / 10.1°C	0	0	0	3	2	3	0.83 *	0% (0% - 1%)

OTU	Month / Temp.	A	B	C	D	E	F	R_s	Relative abundance (range)
Nitrospira	JUL / 17.9°C	4	4	38	38	31	59	0.79	4% (1% - 5%)
	AUG / 19.4°C	8	2	48	29	41	30	0.54	5% (1% - 8%)
	SEPT / 20.2°C	17	8	31	36	50	31	0.70	3% (1% - 4%)
	OCT / 15.7°C	12	2	6	25	11	14	0.43	2% (1% - 3%)
	NOV / 10.1°C	3	0	4	4	3	4	0.53	1% (0% - 1%)
Thioalkalispira	JUL / 17.9°C	77	35	50	58	63	69	0.14	8% (5% - 16%)
	AUG / 19.4°C	24	2	5	3	6	5	-0.06	1% (1% - 4%)
	SEPT / 20.2°C	2	0	0	0	0	0	-0.66	0% (0% - 0%)
	OCT / 15.7°C	0	0	0	0	0	0	N/A	0% (0% - 0%)
	NOV / 10.1°C	0	0	0	0	0	0	N/A	0% (0% - 0%)

Figure 5. Detailed quantitative measurements of seven OTUs detected at six consecutive locations within the DWDS. The number of 16S rRNA gene copies that belong to the *Peribacterales, Parcubacteria, Gallionellaceae, Comamonadaceae*, *Deferrisoma*, *Nitrospira*, and *Thioalkalispira* OTUs were measured in 5 sample series and shown as a heatmap using white shades for low absolute abundances and red shades for high absolute abundances. Values within the coloured boxes represent the calculated number of 16S rRNA gene copies/100µL. Sample series are presented per month following consecutive locations within the DWDS (A-F) and includes the average water temperature that was measured at each sampled location. Spearman correlations (R_s) and its associated significances (**: p-value < 0.01; *: p-value < 0.05) were calculated to check for significant increases/decreases in the absolute abundance of these OTUs throughout the DWDS. The relative abundance represents the median percentage of the number of 16S rRNA gene copies belonging to each OTU measured per sample compared to the total number of 16S rRNA gene copies measured per sample with the range given between brackets.

DISCUSSION

In this study, we evaluated the use of the user-friendly MYcrobiota platform as a culture-free monitoring tool to be used for microbiological DW quality assessment. For this, a full-scale DWDS served as a model system to investigate the performance, advantages and drawbacks of MYcrobiota compared to conventional microbiological screening methods. Our results show that MYcrobiota facilitates the accurate quantification of the total bacterial biomass present in the DWDS, as the dynamic bacterial trends – based upon the total bacterial biomass estimated using MYcrobiota – are in close agreement with the dynamic bacterial trends observed using FCM – which measures intact bacterial cell concentrations. This result indicates that MYcrobiota is not hindered by the presence of cell-free bacterial DNA derived from dead bacterial cells. Surprisingly however, the number of 16S rRNA gene copies measured using MYcrobiota was lower compared to the number of intact cells detected using FCM within the same DW samples. These differences are likely caused by a systematic loss of bacterial cells and/or bacterial DNA during sample processing (e.g. filtration and DNA extraction) since the accuracies of both methods were demonstrated in this study or elsewhere.[17] Note however that the

differences in absolute quantification of intact bacterial cell concentrations and 16S rRNA gene copies did not affect the dynamic trends observed in this study.

The sensitive and accurate measurement of quantitative bacterial biomass allowed us to assess the changes in bacterial communities during water transport and distribution within a DWDS. This was illustrated by a small, but reproducible, decrease of bacterial biomass between locations A and B, which was not detected using conventional HPC and bATP measurements. The decrease of bacterial biomass is most likely the result of chlorine dioxide dosage before storage in the DW reservoir. Despite a residence time of 1 to 2 hours on average in the DW reservoir located between the chlorine dioxide dosage point and sampling location A, the disinfection process may still be happening in the subsequent transport section and intermediate reservoir located before sampling location B. This effect may be causing a decrease in bacterial biomass between locations A and B, with a decrease in intact bacterial cells, as well as a decrease in 16S rRNA gene copies of individual OTUs. Besides providing additional contact time with chlorine dioxide, the intermediate storage reservoir may also contribute to the decrease of bacterial biomass through e.g. settling, though the specific role of reservoirs was not specifically studied here. However, the disinfectant residual is not maintained in the transport and distribution system after location B, which results in an increase in bacterial biomass between locations B and C, indicating re-growth of bacteria. Importantly, further microbial re-growth could not be detected from location C to location F. The absence of changes in bacterial biomass concentration as well as community composition after location C is surprising. While locations C and D are located in large transport mains (500 to 900 mm diameter), sampling location E is situated in a small distribution pipe (152 mm diameter), and location F inside a consumer household. It has been shown earlier that small diameter pipes and low flow velocities or even stagnation have a great impact on bacterial growth and/or interaction between bulk water bacteria and biofilm.[18-20] Therefore, one would expect further growth and/or shifts in bacterial communities in these later sections. It should be noted however that DW samples were taken at location F after flushing of the taps, in order to avoid the household plumbing effect.

DW samples obtained from all locations sampled in September showed a similar number of 16S rRNA gene copies and intact bacterial cell concentrations, which suggested the presence of a stable bacterial biomass concentration across all of the locations sampled (Figure 2). However, the use of MYcrobiota showed that these apparently stable bacterial loads actually involved unexpected increases and expected decreases of 16S rRNA gene copies derived from multiple OTUs (Figure 5). Since the OTU dynamics between locations A and B in September were completely different compared to the general reduction of 16S rRNA gene copies measured between both locations for the other months, we speculate that this unexpected increase of bacterial biomass for certain OTUs in September was caused by deviation in the water quality leaving the treatment plant.

In this period, turbidity was 2 to 3 times higher than normal in the water at location A, which suggests that the water contained more particles and/or more microorganisms. The particles and microorganisms would react with the chlorine dioxide and accelerate the chlorine dioxide decay, which is even enhanced by higher temperatures – note that the highest temperatures were measured in September throughout the whole study. A faster disinfectant decay would allow earlier re-growth in the transport pipe and/or reservoir, which would result in higher numbers of 16S rRNA gene copies and intact bacterial cell concentrations than at other times of the year. Although the reduction of bacterial biomass between locations A and B was restored in the following months, these results highlight the ability of MYcrobiota to detect specific bacterial taxa that might be linked to possible problematic scenarios, such as (sudden) excessive bacterial growth, and opens the possibility of continual monitoring and the implementation of rapid action to maintain water quality using MYcrobiota.

The MYcrobiota results indicated only a small subset of bacteria dominating the DWDS, despite its high bacterial diversity. These bacteria comprised seven distinct OTUs (at a cut-off of 97% sequence similarity) that were classified as *Comamonadaceae*, *Deferrisoma*, *Gallionellaceae, Nitrospira*, *Parcubacteria* (also known as candidate phylum OD1), *Peribacterales* (candidate phylum PER), and *Thioalkalispira*. All of these bacterial taxa were previously detected within other DWDS samples,[2,11,21] or within other aquatic ecosystems.[22-24] Interestingly, the *Peribacterales*, *Parcubacteria* and *Gallionellaceae* OTUs dominated the start of the DWDS, but were 'replaced' by the *Comamonadaceae*, *Deferrisoma*, and *Nitrospira* OTUs at the end of the DWDS. This finding was independent of the month of sampling and suggests that temporal variations are minor compared to spatial variations within this DWDS during the 5-month period of sampling. It should be noted however that most bacterial OTUs detected within this DWDS could not be identified at the genus taxonomic level due to the basic lack of 16S rRNA gene reference sequences derived from members of DW microbial communities in large publicly accessible reference databases (e.g. SILVA, RDP, GreenGenes, or NCBI). It is very likely that these OTU sequences belong to novel lineages for which there are no available culturable representatives, as the MYcrobiota protocol has been specifically established to help prevent the formation of PCR artefacts (i.e. chimeric sequences). This suggests that our current knowledge about the ecology of DW microbial communities is currently limited and requires further investigation with respect to the presence of potential (as yet unknown) bacterial pathogens in DW. Further, it is impossible to link bacterial taxa identified by MYcrobiota to their function in DWDSs. Only additional functional testing, or the identification of genes with well-established functions will provide insights into the functionality of the DW microbiome. For example, Di Rienzi and colleagues elegantly demonstrated that groundwater samples harbour non-photosynthetic bacteria belong-

ing to a new candidate phylum sibling to *Cyanobacteria* via the use of metagenomic sequencing and metabolic reconstruction.[25]

Recently, there is a discussion about whether OTU-based strategies – such as implemented within MYcrobiota – should be replaced by newly developed amplicon sequence variants (ASV)-based methods as the standard to delineate microbial taxa.[26] These ASV-based methods avoid clustering sequences at arbitrary thresholds that define OTUs (e.g. 97%) by using only unique, identical marker gene sequences for downstream analysis. Unlike OTUs, ASVs can be resolved to single-nucleotide differences over the sequenced gene region using specialized 'de-noising' algorithms that is expected to increase taxonomic resolution.[27] However, Glassman and Martiny recently illustrated that OTU-based and ASV-based methods will often reveal similar ecological results when using the 16S rRNA gene.[28] This finding can be explained by the fact that 16S rRNA gene sequence types may not reflect ecologically or phylogenetically cohesive populations.[29] Therefore, in this study, we employed a traditional OTU-based strategy to simplify the identification of (sub-sets) of bacterial taxa that changed in abundance across the sampled locations. Of note, future versions of MYcrobiota could potentially include ASV-based analysis as part of its software, depending on the requirement to obtain fine-scale taxonomy results.

In conclusion, MYcrobiota is an alternative for the culture-free monitoring of bacteria that reside within DWDSs and allows DW providers to gain accurate insights into the spatial and temporal microbial dynamics within their DWDS that are not observed using conventional methods. Using information obtained by MYcrobiota facilitates the continual assessment of the desired 'biological stability' over the whole DWDS, thereby helping ensure that safe and high-quality DW reaches the consumer.

MATERIALS AND METHODS

Drinking water sample collection

A total of 30 DW samples were collected from a full-scale DW transport and distribution network in the Netherlands. The transported DW was produced from surface water, using coagulation/sedimentation, rapid sand filtration, advanced oxidation, and activated carbon filtration. Chlorine dioxide (ClO_2) was added to the water prior to storage at the DW reservoir. The residual ClO_2 after storage is 0.03 mg/L on average, but is not maintained during DW transport and distribution. DW samples were taken at six locations throughout the DWDS, from the treatment plant towards the studied distribution area, as follows: (A) effluent of water treatment plant, after ClO_2 dosage and after DW reservoir, (B) effluent of intermediate pumping station and storage reservoir after a first transport pipe (800-900 mm), (C and D) at sampling points in the second and third transport pipe

sections (800-900 mm and 500-700 mm), (E) at a sampling point within the distribution pipe (40 – 300 mm) and (F) at a household tap (Figure 1). High-density polyethylene (HD-PE) plastic bottles (Identipack BV) containing 2 mL L^{-1} of a mixed solution of sodium thiosulfate (20 g L^{-1}) and of nitrilotriacetic acid (25 g L^{-1}) were used to collect DW for HPC, bATP, FCM, and MYcrobiota analysis, as routinely used by accredited laboratories for DW analysis in the Netherlands. The DW samples were transported and stored at 4°C until analysis, and processed within 24 hours after sampling.

Conventional parameters

HPC was measured according to the Dutch standard procedure (NEN-EN-ISO 6222, 1999). In short, 1 mL per DW sample was transferred to a sterile Petri dish and mixed with 20 mL of yeast extract agar. The agar was kept at 44°C before plating. The samples were incubated at 22°C for 3 days. ATP was measured as described previously by Magic-Knežev and van der Kooij.[30] The ATP measurement is based on the emission of light resulting from the reaction between the ATP molecule and a luciferin/luciferase reagent (LuminATE, Celsis). For total ATP determination, ATP was first released from suspended microbial cells with nucleotide-releasing buffer (LuminEX, Celsis), while this step was not performed for the assessment of free ATP. The intensity of the emitted light was measured using a luminometer (Celsis AdvanceTM) that was calibrated with solutions of free ATP (Biotherma) in autoclaved tap water following the procedure given by the reagent manufacturer. Bacterial ATP concentrations were calculated by subtracting free ATP from total ATP concentrations. Finally, FCM analysis were performed following the protocol described by Prest et al.[17,31] In short, DW samples (100 µL) were pre-heated to 37°C for 4 minutes, stained with fluorescent dyes and incubated in the dark for 10 minutes at 35°C before measurement. Bacterial staining with 10 µL per mL of a working solution containing a mixture of SYBR Green I (1:100 dilution in DMSO; Molecular Probes) and propidium iodide (0.5 mg/mL) was used for the assessment of intact bacterial cell concentrations. FCM measurements were performed using a BD Accuri C6 FCM (BD Accuri Cytometers) equipped with a 50 mW laser emitting at a fixed wavelength of 488 nm. The FCM is equipped with volumetric counting hardware, calibrated to measure the number of particles in a 50 µL volume fraction of a 100 µL sample. Measurements were performed at a pre-set flow rate of 35 µL per minute. A threshold value of 450 a.u. was applied on the green fluorescence channel (FL1). Bacterial signals were selected and distinguished from inorganic particles and instrument background on the BD Accuri CFlow software using electronic gating on density plots of green fluorescence (FL1; 533 nm), and red fluorescence (FL3; >670 nm).[17]

MYcrobiota analysis

One litre of water from each DW sample was filtered through a 0.2 µm pore-size polycarbonate membrane filter (Sartorius) within 5 hours of sampling, using sterile (autoclaved) filtration units. The filters were stored at -20°C until processing. DNA was extracted from the collected biomass using the PowerBiofilm DNA Isolation Kit Sample (MO BIO Laboratories) according to the manufacturer's instructions. In addition, DNA from a sterile filter was extracted as a NEC at the same time in order to allow for the subtraction of contaminating bacterial DNA after NGS processing. Amplicon library preparation using micPCR that clonally amplified the V4 regions of 16S rRNA genes was performed as previously published,[15] but with a slight modification. In this study, the IC to determine the absolute quantity of 16S rRNA gene copies consisted of quantified genomic DNA from a *Campylobacter jejuni* bacterium (ATCC 700819). Importantly, prior to the addition of the IC, the absence of *Campylobacter* species within each DNA extract was established using a *Campylobacter* specific PCR according to Lund et al.[32] We utilized the micPCR/NGS approach to process all samples, including the NEC, in triplicate in order to increase accuracy and to correct for contaminating bacterial DNA derived from the laboratory environment as previously described.[14] FASTQ-formatted sequences were extracted after paired-end sequencing of the 16S rRNA gene amplicon library using the MiniSeq system (Illumina) and processed using our previously developed bioinformatics analysis service.[15] This bioinformatics pipeline consists of 23 well-established mothur tools (v.1.36)[33] and an additional 9 custom-made tools that have been integrated and combined in Galaxy,[34] and allows for a fully automated sequence interpretation of 16S rRNA gene micPCR/NGS data. The sequencing data that are connected to this article are uploaded to the Sequence Read Archive database with accession number SRP114562.

Quantification of 16S rRNA gene molecules

The total number of 16S rRNA gene molecules within each DNA extract was measured using a 16S rRNA gene quantitative qPCR as described previously.[16] For this, CT-values were related to a 10-fold dilution series of a synthetic microbial community (SMC) sample, containing 10,000 16S rRNA gene copies of *Moraxella catarrhalis* (ATCC 25240), *Staphylococcus aureus* (ATCC 43300), *Haemophilus influenzae* (ATCC 10211), and *Clostridium perfringens* (ATCC 12915). The qPCRs were performed in 10 µL reaction volumes using the LightCycler 480 Probes Master (Roche) with the addition of 0.5 µM of each PCR primer and 0.25 µM of a Fam-labelled probe for the real-time detection of the 16S rRNA gene amplification. All qPCRs were performed using the following conditions: initial denaturation at 95°C for 5 minutes followed by 45 cycles of PCR, with cycling conditions of 5 seconds at 95°C, 10 seconds at 55°C, and 30 seconds at 72°C.

Statistical analysis

Spearman's correlation coefficients were calculated to check for significant increases/decreases in the absolute abundance of OTUs measured over the DWDS (SPSS version 23, IBM Corporation).

SUPPLEMENTARY DATA

Supplementary information is available upon request from the authors.

REFERENCES

1. Belila A, El-Chakhtoura J, Otaibi N, et al. Bacterial community structure and variation in a full-scale seawater desalination plant for drinking water production. *Water Res* 2016; **94**: 62-72.
2. Shaw JL, Monis P, Weyrich LS, et al. Using amplicon sequencing to characterize and monitor bacterial diversity in drinking water distribution systems. *Appl Environ Microbiol* 2015; **81**: 6463-6473.
3. Proctor CR, Hammes F. Drinking water microbiology – from measurement to management. *Curr Opin Biotechnol* 2015; **33**: 87-94.
4. Wang H, Bedard E, Prevost M, et al. Methodological approaches for monitoring opportunistic pathogens in premise plumbing: a review. *Water Res* 2017; **117**: 68-86.
5. Sun H, Shi B, Lytle DA, et al. Formation and release behavior of iron corrosion products under the influence of bacterial communities in a simulated water distribution system. *Environ Sci Process Impacts* 2014; **16**: 576-585.
6. Srinivasan R, Sorial GA. Treatment of taste and odor causing compounds 2- methyl isoborneol and geosmin in drinking water: a critical review. *J Environ Sci* 2011; **23**: 1-13.
7. Prest EI, Hammes F., van Loosdrecht MCM, et al. Biological stability of drinking water: controlling factors, methods, and challenges. *Front Microbiol* 2016; **7**: 45.
8. Hoefel D, Grooby WL, Monis PT, et al. Enumeration of water-borne bacteria using viability assays and flow cytometry: a comparison to culture-based techniques. *J Microbiol Methods* 2003; **55**: 585-597.
9. Van Nevel S, Koetzsch S, Proctor CR, et al. Flow cytometric bacterial cell counts challenge conventional heterotrophic plate counts for routine microbiological drinking water monitoring. *Water Res* 2017; **113**: 191-206.
10. Pinto AJ, Schroeder J, Lunn M, et al. Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. *MBio* 2014; **5**: e01135-14.
11. Roeselers G, Coolen J, van der Wielen PWJJ, et al. Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. *Environ Microbiol* 2015; **17**: 2505-2514.
12. Boers SA, Jansen R, Hays JP. Suddenly everyone is a microbiota specialist! *Clin Microbiol Infect* 2016; **22**: 581-582.
13. Boers SA, Hays JP, Jansen R. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Sci Rep* 2015; **5**: 14181.
14. Boers SA, Hays JP, Jansen R. Novel micelle PCR-based method for accurate, sensitive and quantitative microbiota profiling. *Sci Rep* 2017; **7**: 45536.
15. Boers SA, Hiltemann SD, Stubbs AP, et al. Development and evaluation of a culture-free microbiota profiling platform (MYcrobiota) for clinical diagnostics. *Eur J Clin Microbiol Infect Dis* 2018; **37**: 1081-1089.
16. Yang S, Lin S, Kelen GD, et al. Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. *J Clin Microbiol* 2002; **40**: 3449-3454.
17. Prest EI, Hammes F, Kötzsch S, et al. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. *Water Res* 2013; **47**: 7131-7142.
18. Prévost M, Rompré A, Coallier J, et al. Suspended bacterial biomass and activity in full-scale drinking water distribution systems: impact of water treatment. *Water Res* 1998; **32**: 1393-1406.

19. Lipphaus P, Hammes F, Kötzsch S, et al. Microbiological tap water profile of a medium-sized building and effect of water stagnation. *Environ Technol* 2014; **35:** 620-628.

20. Ling F, Whitaker R, LeChevallier MW, et al. Drinking water microbiome assembly induced by water stagnation. *ISME J* 2018; **12:** 1520-1531.

21. Hwang C, Ling F, Andersen GL, et al. Microbial community dynamics of an urban drinking water distribution system subjected to phases of chloramination and chlorination treatments. *Appl Environ Microbiol* 2012; **78**: 7856-7865.

22. 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.

23. Slobodkina GB, Reysenbach AL, Panteleeva AN, et al. *Deferrisoma camini* gen. nov., sp. nov., a moderately thermophilic, dissimilatory iron(III)-reducing bacterium from a deep-sea hydrothermal vent that forms a distinct phylogenetic branch in the *Deltaproteobacteria*. *Int J Syst Evol Microbiol* 2012; **62**: 2463-2468.

24. Sorokin DY, Tourova TP, Kolganova TV, et al. *Thioalkalispira microaerophila* gen. nov., sp. nov., a novel lithoautotrophic, sulfur-oxidizing bacterium from a soda lake. *Int J Syst Evol Microbiol* 2002; **52**: 2175-2182.

25. Di Rienzi SC, Sharon I, Wrighton KC, et al. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to *Cyanobacteria*. *Elife* 2013; **1:** e01102.

26. Callahan BD, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017; **11:** 2639-2643.

27. Callahan BD, McMurdie PJ, Rosen MJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13:** 581-583.

28. Glassman SI, Martiny JBH. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *mSphere* 2018; **3:** e00148-18.

29. Berry MA, White JD, Davis TW, et al. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes. *Front Microbiol* 2017; **8:** 365.

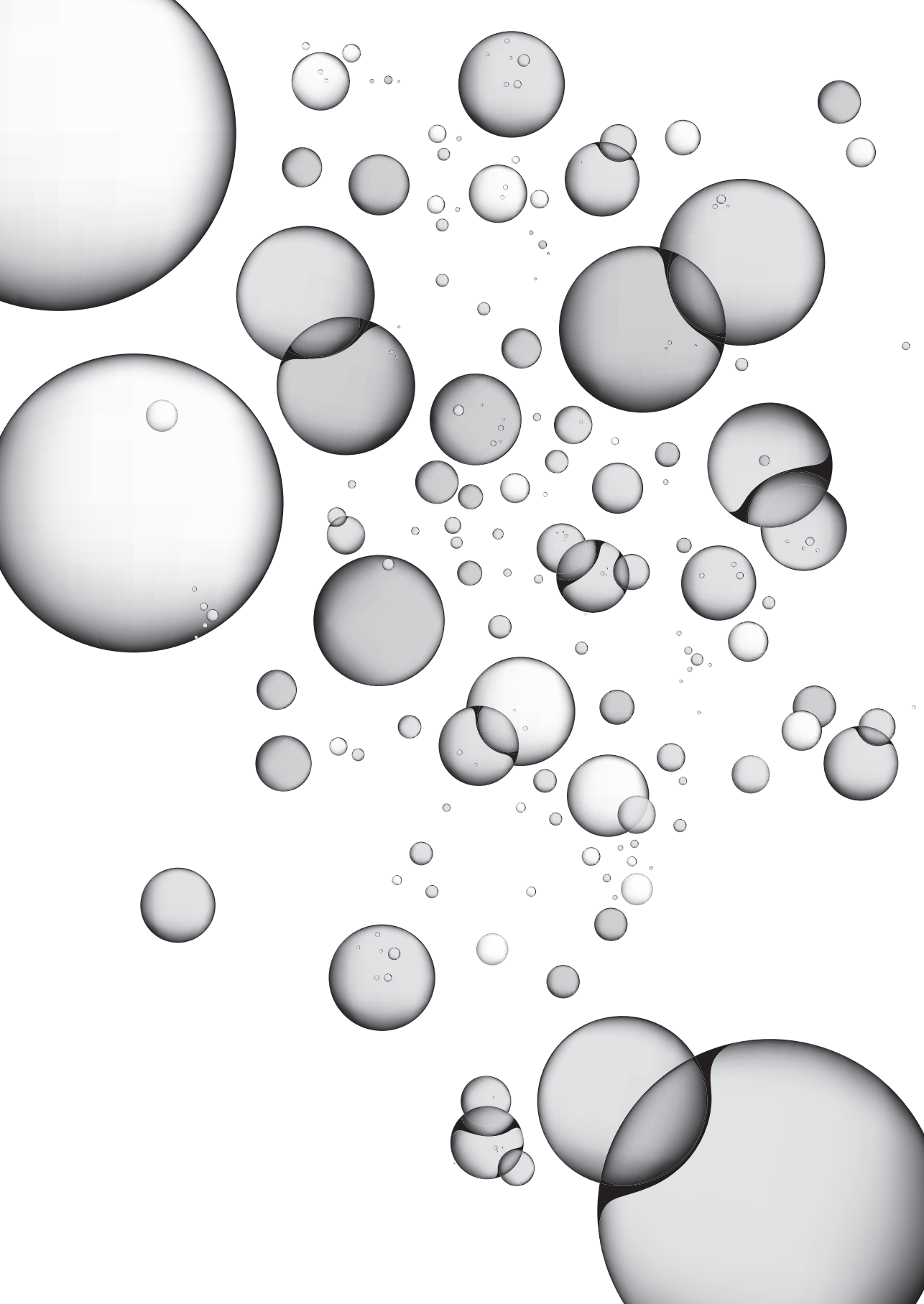
30. Magic-Knezev A, van der Kooij D. Optimisation and significance of ATP analysis for measuring active biomass in granular activated carbon filters used in water treatment. *Water Res* 2004; **38**: 3971-3979.

31. Prest EI, El-Chakhtoura J, Hammes F, et al. Combining flow cytometry and 16S rRNA gene pyrosequencing: a promising approach for drinking water monitoring and characterization. *Water Res* 2014; **63**: 179-189.

32. Lund M, Nordentoft S, Pedersen K, et al. Detection of *Campylobacter* spp. in chicken fecal samples by real-time PCR. *J Clin Microbiol* 2004; **42**: 5125-5132.

33. 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.

34. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010; **11**: R86.



Chapter 9

Summarizing discussion, conclusions, and future perspectives

SUMMARIZING DISCUSSION

As mentioned in the introduction to this thesis (**Chapter 1**), next-generation sequencing (NGS) and phylogenetic analysis of 16S ribosomal RNA (rRNA) genes provided the foundation for modern study of microbial communities, generating many hundreds of publications. However, there has been very little focus on the development and application of these culture-independent 16S rRNA gene NGS methods into routine diagnostics. Therefore, the aim of this thesis was to develop and apply an accurate 16S rRNA gene NGS protocol for use in routine (clinical) microbiological diagnostic laboratories. Specifically, this thesis describes the development of an 'end-to-end' microbiota profiling platform (MYcrobiota), consisting of a novel calibrated micelle-based PCR (micPCR) amplification strategy coupled to a dedicated and 'easy-to-use' bioinformatics pipeline, which was validated and evaluated using a range of samples, including synthetic microbial community (SMC) mixes, samples from the clinic, and environmental samples. The main findings of the research performed relating to: i) the development and validation of MYcrobiota, and ii) the evaluation of MYcrobiota for routine clinical and environmental microbiological diagnostic use, are summarized and discussed in the following thesis chapter. In addition, the main conclusions of this thesis are listed and recommendations for future research are made.

Development and validation of MYcrobiota

In order to advance 16S rRNA gene NGS methods to a more standardized and routine (clinical) microbiological diagnostic procedure, it is essential to establish uniform and validated standard operating procedures (SOPs) throughout laboratories for maintaining consistent test performance. With this in mind, MYcrobiota was designed to overcome the most important experimental pitfalls and biases of current 16S rRNA gene NGS methods that have previously hampered the introduction of these methods into the routine (clinical) microbiological diagnostic laboratories (Table 1, **Chapters 1** and **2**). Specifically, MYcrobiota was designed to: i) prevent chimera formation, ii) reduce PCR competition induced bias, iii) standardize microbiota profiling results via the absolute quantification of each microorganism present within microbial samples, iv) remove contaminating DNA derived from the experimental set-up, and v) simplify bioinformatics analysis that can be operated by non-bioinformaticians. In the following section, each of these technical improvements of the 16S rRNA gene NGS process that were realized with the development of MYcrobiota will be discussed in more detail.

Prevention of chimera formation. Current 16S rRNA gene NGS methods apply multi-template PCR strategies in order to amplify all 16S rRNA gene template molecules simultaneously within a single reaction tube. However, the use of multi-template PCRs

Table 1. Overview of the potential experimental pitfalls and biases generated using 'traditional' 16S rRNA gene NGS methods and the technical improvements that were realized with the development of MYcrobiota.

Experimental pitfalls and biases*	General remarks and MYcrobiota improvements
Step 1: Sampling collection	
- Sampling protocol	- The sampling protocol depends on the sample type to be investigated.
- Transport and storage conditions	- Optimal preservation of microbial samples involves immediate freezing followed by long-term storage at -80°C.
Step 2: DNA extraction	
- Lysis method	- The choice of the most efficient lysis method is dependent on the sample type and target microbial species under investigation.
Step 3: PCR amplification	
- Selection of PCR primers	- The most optimal PCR primer pair should be selected based on its primer binding capacity to the (expected) microbial species present within the investigated sample.
- PCR competition effects and chimera formation	- **MYcrobiota** utilizes a novel micelle PCR/NGS methodology that limits both the formation of PCR competition induced bias and chimera sequences (**Chapter 3**).
Step 4: Next-generation sequencing	
- Technical limitations of the NGS-platform used	- **MYcrobiota** targets the 16S rRNA gene V4 region, which allows for a large overlap of DNA sequences that are obtained from both ends of the PCR amplicon using Illumina's MiniSeq/MiSeq NGS-platforms. This strategy generates the lowest error rates compared to other 16S rRNA gene regions used with these NGS-platforms (**Kozich et al., 2013**).[75]
Step 5: Bioinformatics analysis	
- Choice of algorithms and their settings	- **MYcrobiota** uses a standardized and validated bioinformatics pipeline that allows for the automated sequence interpretation of 16S rRNA gene NGS data without the requirement for advanced bioinformatics skills (**Chapters 5** and **6**).
- Quality/completeness of reference databases	- **MYcrobiota** allows for a manual evaluation of the taxonomic identifications of bacterial genera (**Chapter 6**).
Miscellaneous	
- Contamination	- **MYcrobiota** determines the absolute quantity of microbial genera present within a sample, which improves the standardization of 16S rRNA gene NGS results and enables the removal of contaminating DNA derived from the experimental set-up via the processing of negative (extraction) control samples (**Chapter 4**).

*: The experimental pitfalls and biases are described in more detail in **Chapter 1** of this thesis.

results in the formation of chimeric sequences, which are composed of multiple distinct DNA sequences that are wrongfully joined together. The presence of chimeric 16S rRNA gene sequences artificially increases microbial diversity calculations, as these sequence artefacts are a major source of spurious operational taxonomic units (OTUs) in microbiota studies.[1,2] In addition, chimeric sequences that are comprised of 16S rRNA gene sequences derived from different taxonomic lineages could be misinterpreted as novel prokaryotic genera, thereby turning the microbiologist unwittingly into a microbial creationist.[3] Current strategies to reduce the amount of chimeric sequences generated by 16S rRNA gene NGS methods involve the use of optimized PCR protocols to reduce the chance of chimeric formation during PCR amplification,[4,5] as well as the use of complex computational algorithms that facilitate the removal of chimeric sequences after the PCR amplification and DNA sequencing processes are complete.[3,6-8] Unfortunately however, none of these methods has been shown to completely eliminate chimeric sequences from 16S rRNA gene NGS datasets entirely.[9]

Therefore, in order to try to completely eliminate the formation of chimeric sequences, we developed a novel micPCR amplification strategy whereby template DNA molecules are separated into a large number of physically distinct reaction compartments using water-in-oil emulsions (**Chapter 3**). This template DNA molecule compartmentalization drastically reduced the formation of chimera sequences due to the statistical presence of a single template DNA molecule in each droplet of the emulsion. For example, our results show that the use of micPCR followed by NGS (micPCR/NGS) reduced the formation of chimeric sequences by a factor of 71 (0.2% vs. 17.4%) compared with traditional PCR/NGS methods (**Chapters 3** and **4**). The fact that small numbers of chimeric sequences were still being detected even when using the micPCR/NGS method can be explained by the possibility that some micelles still host more than one template DNA molecule during micPCR amplification or through the generation of false positive chimera results by downstream computational methods. This observation shows that the number of template DNA molecules have to be carefully adjusted for each micPCR/NGS experiment in order to try to achieve a balance of one template DNA molecule per micelle. Nonetheless, inevitable sequence-errors introduced by PCR polymerases,[10] and NGS-platforms in general,[9] result in 'noisy' 16S rRNA gene NGS reads that can be misidentified as chimeric artefacts by the chimera checking software tool used. These false-positive chimera results could explain the very low numbers of chimeric sequences that are regularly detected within micPCR/NGS datasets, even when optimized numbers of template DNA molecules are used.

The vast reduction in chimeric sequences results in more accurate microbial diversity estimates. This can be illustrated using rarefaction analyses, in which the researcher plots the number of OTUs as a function of the number of 16S rRNA gene NGS reads obtained. Rarefaction curves generally grow rapidly at first, as the most common OTUs are found,

but the curves soon reach a plateau, as only the rarest OTUs remain to be sampled. As shown in **Chapter 3**, rarefaction curves rapidly reached a plateau phase using micPCR/NGS at the expected 20 OTU level when using a SMC sample that contained 20 different bacterial species. This result indicates that micPCR/NGS generates an accurate view of microbial diversity within these samples. In contrast, a traditional PCR/NGS method (used as comparator) resulted in 72 OTUs, with rarefaction analysis showing that the number of OTUs per sample steadily increased as the number of 16S rRNA gene NGS reads increased. Importantly, it was found that this excess of OTUs consisted of chimeric sequences that had not been recognized as chimeras by the chimera checking software tool used. In addition, samples obtained from human healthy volunteers (including low biomass nasal swab samples and high biomass faecal samples), as well as highly diverse environmental sludge samples, revealed that chimeric sequences were reduced in all samples when using micPCR/NGS, resulting in decreased diversity estimates among all samples compared to traditional PCR-based results (**Chapter 3**). Therefore, micPCR/NGS drastically reduces chimera formation without the reliance on complex downstream computational methods, resulting in more accurate microbial diversity estimates compared to traditional PCR/NGS methods.

Reduction of PCR competition induced bias. Another important factor that can impact 16S rRNA gene NGS results is PCR competition between different 16S rRNA gene template molecules in polymicrobial samples, resulting in unequal amplification rates for certain template DNA sequences. This biased amplification of 16S rRNA gene template molecules can lead to over- or underestimations of particular OTUs.[11,12] However, the clonal amplification of each template DNA molecule during micPCR prevents the generation of PCR competition artefacts because all template DNA molecules (as well as potentially competing non-template DNA molecules) are limited to a single micelle and amplified to the extend until all limited resources contained within this host micelle are depleted. As a result, equal amplicon yields are obtained for each targeted template DNA molecule (independently of template-specific PCR amplification efficiencies) that accurately represents the template ratios in the test sample. In contrast, the unequal amplification rate of certain template DNA molecules during multi-template PCRs could yield (unpredictable) amplicon ratio's that might not represent the original sample composition.[13] Therefore micPCR/NGS allows for a more accurate interpretation of the actual microbiota profile ratio compared to traditional PCR/NGS methods. For example, micPCR/NGS data showed an average of only a 0.85-fold difference (range: 0.28-fold - 1.73-fold difference) between the measured and expected relative abundances of OTUs obtained from an SMC sample comprising 20 different bacterial species, whereas the traditional PCR/NGS data revealed an average of 0.65-fold difference (range: 0.04-fold – 2.31-fold difference) using the same SMC sample (**Chapter 3**). Importantly, biased

PCR amplification efficiencies led to different interpretations of microbial content when using actual samples of unknown composition. For example, micPCR/NGS showed a 3.3-fold reduction in *Staphylococcus* abundance among two nasal swab samples (2.4% vs. 7.8%), whereas traditional PCR/NGS showed a 4.7-fold increase in *Staphylococcus* abundance among the same two nasal swab samples (12.2% vs. 2.6%). Although the actual composition of these investigated nasal swab samples is unknown, the relative abundances obtained using micPCR/NGS likely represents a more accurate reflection of the true relative abundances as indicated using the SMC samples.

The reduction of PCR competition induced bias by micPCR/NGS would also improve the accurate characterisation of microbial communities that are investigated using other genetic markers, such as the internal transcribed spacer (ITS) region. For example, the ITS regions are often used to investigate fungal communities,[14] and the prokaryotic variant can be employed to improve the resolution of prokaryotic species identification.[15] However, these ITS fragments are known to be of uneven lengths, even among highly genetically related species, for which traditional PCR methods may promote preferential amplification of shorter ITS fragments compared to longer ITS fragments, resulting in biased microbiota profiles.[16] In contrast, the compartmentalization of template DNA molecules using micPCR reduces the type of PCR competition normally observed between fragments of different lengths,[17] and therefore would result in more accurate quantification of microbial relative abundances.

Standardization of microbiota profiling results. Current microbiota profiling studies use semi-quantitative 16S rRNA gene NGS methods, where the microbiota results are presented as proportional abundances rather than absolute abundances. This limitation lowers the reproducibility of 16S rRNA gene NGS results and complicates cross-study comparability.[18] For example, the interpretation of microbial community dynamics based on relative abundances can be misleading because fluctuations in the absolute abundance of one microorganism may cause an apparent change in the measured relative abundance of all other microorganisms.[19] To overcome this limitation, micPCR/NGS can be performed in combination with an internal calibrator (IC) to determine the composition and absolute quantity of microbial genera (**Chapter 4**). This IC consists of quantified genomic DNA from a bacterium that is selected for its absence in the natural microbial flora of the investigated samples and is added to each DNA extract prior to micPCR amplification. After NGS processing, the IC is used to calculate a correction factor that in turn is used to convert the obtained 16S rRNA gene NGS reads per OTU to 16S rRNA gene copies per OTU (equation 1). Because micPCR/NGS is less vulnerable to inevitable PCR amplification biases (chimera formation and PCR competition) compared to traditional PCR/NGS methods, micPCR allows for the utilization of just a single correction factor, obtained from a single IC, to convert all 16S rRNA gene NGS reads to

16S rRNA gene copies for each individual OTU detected within a polymicrobial sample. In contrast, alternative spike-in approaches that employ traditional PCR amplification methods, such as the SCML protocol,[20] or the use of artificial 16S rRNA gene spike-ins,[21] remain vulnerable to template-specific variations in PCR efficiencies and could easily result in erroneous quantitative microbiota profiles. Importantly, note that each sample is processed in triplicate using our 'calibrated' micPCR/NGS method, in order to average out any possible quantification bias generated due to differences in the distribution of micelle sizes between independent micPCR experiments.

$$\text{16S rRNA gene copies (OTU)} = \text{16S rRNA gene NGS reads (OTU)} \times \left(\frac{\text{16S rRNA gene copies (IC)}}{\text{16S rRNA gene NGS reads (IC)}} \right)$$

Equation 1. Calculation of 16S rRNA gene copies using micPCR/NGS in combination with an internal calibrator (IC).

In order to validate our calibrated micPCR/NGS method, we investigated the trueness and precision of the technique. Trueness is a term defined as the proximity of the measured result obtained compared to the actual 'true' reference value, whilst precision is a term defined as the closeness of agreement among a set of results. As shown in **Chapter 4**, both the calibrated micPCR/NGS and traditional PCR/NGS methods (used as comparator) generated similar average $\log_2$ fold-changes between the measured 16S rRNA gene copies compared to the expected number of 16S rRNA gene copies within a 10-fold dilution series of an SMC sample, indicating a similar and good trueness. However, the dispersal of replicate results obtained using the calibrated micPCR/NGS strategy was much smaller compared to the traditional PCR/NGS method, indicating the higher precision of the calibrated micPCR/NGS method. This is illustrated by SMC samples containing 2,500, 250, 25 and 2.5 16S rRNA gene copies per bacterial species, for which the traditional PCR/NGS generated 70, 3, 25 and 97-fold differences, respectively, between the actual measured and reference number of 16S rRNA gene copies. In contrast, the calibrated micPCR/NGS approach resulted in only 3, 3, 5 and 7-fold differences, respectively, within the same SMC samples. Importantly, the higher precision of the calibrated micPCR/NGS methodology lowers the number of random errors within the 16S rRNA gene NGS measurements and therefore increases the repeatability of quantitative microbiota profiling results.

The accuracy of the calibrated micPCR/NGS method for determining the number of 16S rRNA gene copies in samples with unknown composition was evaluated by comparing the results to direct measurements of the total 16S rRNA gene copies obtained using a 16S rRNA gene quantitative PCR (qPCR). These comparisons revealed an average of only a 1.4-fold difference (± 0.4) between both quantitative methods for 71 clinical and environmental samples that were included over multiple studies described in this thesis

(**Chapters 4**, **6**, **8**). Importantly, the 16S rRNA gene qPCR used for these comparisons utilizes a different universal 16S rRNA gene primer set,[22] targeting a different region of the 16S rRNA gene compared to the calibrated micPCR/NGS method. Therefore, the 16S rRNA gene qPCR is a complementary technique that enables the accurate validation of the calibrated micPCR/NGS method when determining the total numbers of 16S rRNA gene copies. In addition, an experimental comparison was made between a *Staphylococcus* OTU-specific biomass and a *Staphylococcus aureus* qPCR using an *S. aureus*-specific genetic marker,[23] which revealed an average of only a 1.3-fold (± 0.4) difference between the two methodologies when using 13 *S. aureus* culture-positive samples (**Chapter 6**). This result demonstrates (again) the intrinsic accuracy of the calibrated micPCR/NGS method, but also highlights the ability to incorporate quantitative results obtained from additional (species-specific) qPCRs into the calibrated micPCR/NGS results. This unique feature is particularly useful if the researcher wants to obtain quantitative species-level data, noting that the sequencing of partial 16S rRNA genes (currently the most commonly used method for 16S rRNA gene NGS) often lacks the discriminatory power to differentiate prokaryotes at the species taxonomic level.[24]

Removal of contaminating DNA. DNA contamination derived from the (laboratory) environment and consumables used in the experimental set-up can significantly influence the results of 16S rRNA gene NGS methods. Potential sources of DNA contamination include DNA extraction kits,[25,26] PCR reagents,[27,28] and possibly contaminating human (skin, oral, and respiratory) microbiota from the researchers themselves. Importantly, the inevitable introduction of DNA contamination within 16S rRNA gene NGS experiments is a particular challenge for researchers working with samples containing a low microbial biomass, for example anterior nasal swabs. In these samples, the microbiota DNA levels present on swabs may not be high enough to generate a significant 'signal' above the 'signal' obtained from background contaminating DNA, resulting in unreliable microbiota profiles.[29] Although several methods have been published that are designed to eliminate and/or reduce background impurities from DNA reagents (including irradiation with UV light, enzymatic degradation, treatment with ethidium monoazide (EMA), etc.), these methods tend to suffer from poor reproducibility or impact negatively on PCR sensitivity.[30-32] In this respect, Biesbroek et al. suggested the use of a 'lower bacterial density threshold' of 10^6 bacteria per mL and 1 pg/µL of template DNA, when working with low biomass samples.[33] However, this suggestion results in the automatic exclusion of many low biomass clinical samples from 16S rRNA gene NGS investigation, including such microbiologically interesting low biomass samples such as joint fluids, cerebrospinal fluids, blood samples or other samples derived from normally 'sterile' body sites. To overcome this limitation and to facilitate the investigation of low biomass (clinical) samples using 16S rRNA gene NGS methods, we developed and validated an alterna-

tive strategy that enables the mathematical removal of contaminating DNA from low biomass (clinical) samples (**Chapter 4**).

The mathematical removal of contaminating DNA using the calibrated micPCR/NGS protocol comprises two steps. First, randomly occurring DNA contamination (derived from the sample-processing environment) are eliminated via the mathematical removal of OTUs that cannot be reproducibly measured in triplicate measurements of the test sample. Secondly, the intrinsic DNA contamination obtained from DNA extraction kits and PCR reagents/consumables is removed via the subtraction of the number of 16S rRNA gene copies that have been amplified, in triplicate, from negative extraction controls (NECs). As shown in **Chapter 4**, correcting for both types of DNA contamination resulted in the complete removal of contaminating 16S rRNA gene copies from SMC samples using the calibrated micPCR/NGS method. In contrast, using the same two-step strategy to remove contaminating DNA, but in combination with a traditional PCR/NGS method, still resulted in contaminating 16S rRNA gene copies being reported for SMC samples, even after mathematical correction. This finding can be explained by the higher accuracy of the calibrated micPCR/NGS method to quantify contaminating DNA from NEC samples and illustrates the requirement of the clonal-based micPCR amplification strategy for the accurate subtraction of contaminating DNA from actual samples. Consequently, the limit of detection (LOD) of the calibrated micPCR/NGS method was determined at only 25 16S rRNA gene copies per OTU in SMC samples, which is lower than the LOD of the traditional PCR/NGS method that was estimated as 250 16S rRNA gene copies per OTU using the same SMC samples. Importantly however, traditional PCR/NGS, in contrast to calibrated micPCR/NGS, was not able to generate any 16S rRNA gene amplicons from actual low biomass clinical samples included in our study as this method only generated non-specific, low molecular weight amplicons (presumed to be of human DNA origin). This result indicates that the LOD for the traditional PCR/NGS method is even higher for actual clinical samples than was estimated using SMC samples. Thus, the high accuracy and low LOD of the calibrated micPCR/NGS, makes this method the preferential method to determine accurate and quantitative microbiota profiles for low biomass samples that are hampered by contaminating prokaryotic DNA.

Simplification of bioinformatics analysis. 16S rRNA gene NGS analysis is provided by an extensive array of sophisticated bioinformatics programs, such as mothur and QIIME,[34,35] with an overview outlined in a recent review by Nilakanta et al.[36] Whilst some of these bioinformatics programs have a graphical user interface (GUI) to provide access to these technologies for the research or clinical scientist, their use remains complex for non-bioinformatics educated users. To address this challenge, together with colleagues, we have integrated the full set of 125+ mothur tools into Galaxy, which is a project dedicated to simplify the use of complex command-line bioinformatics tools

using a 'user-friendly' web interface (**Chapter 5**).[37-40] These tools are collectively called the 'Galaxy mothur Toolset (GmT)', which provide access to all of the individual mothur components as separate tools, whilst retaining the full flexibility of mothur by creating custom bioinformatics pipelines. In addition, GmT supports the integration of third party visualization tools, including KRONA and Phinch,[41,42] as well as reporting tools such as iReport,[43] allowing for the easy interpretation, sharing and storage of the results obtained from 16S rRNA gene NGS experiments. The GmT is freely accessible for all users via the 'Galaxy's Tool Shed' at https://toolshed.g2.bx.psu.edu.

In order to process 16S rRNA gene NGS data that is generated using the calibrated micPCR/NGS method, we adjusted a GmT-based bioinformatics pipeline to our specific use-case (**Chapter 6**). This dedicated bioinformatics pipeline performs all of the 'standard' steps (e.g. quality filtering, OTU clustering, sequence classification, etc.) involved with 16S rRNA gene NGS analysis, but also performs 'calibrated micPCR/NGS-specific' steps that include: i) averaging over multiple technical replicates, ii) converting the number of obtained 16S rRNA gene NGS reads per OTU to 16S rRNA gene copies per OTU via the use of an IC and iii) correcting for contaminating DNA using the two-step DNA contamination removal strategy previously described above. All these processes are started and executed via a single push of the button, with the results being presented to the user via an interactive web report in Galaxy using the iReport tool. This standardized report visualizes the resultant microbiota profiles and summarizes the results of three diversity estimators (Chao1, Shannon and Simpson indices). In addition, an extensive overview of the quality control measurements taken during this automatically performed 16S rRNA gene NGS analysis is also provided, allowing the user to manually evaluate the results using the quality measurements pre-installed within the bioinformatics pipeline. The newly developed GmT bioinformatics pipeline has been combined with the calibrated micPCR/NGS methodology to create an 'end-to-end' microbiota profiling platform that is referred to as 'MYcrobiota'.

Evaluation of MYcrobiota for routine diagnostic use

Microbiota analysis has promising applications in the field of routine (clinical) microbiological diagnostics. However, 16S rRNA gene NGS methods have not yet made the transition from research into routine clinical/environmental diagnostic practice due to the lack of a validated protocol and the requirement for expert bioinformaticians to analyse the NGS data obtained.[18,44] As described in the previous part of this chapter, the MYcrobiota platform overcomes both of these limitations and therefore, we evaluated the utility of MYcrobiota for use in the routine clinical and environmental microbiological diagnostic laboratories.

Clinical microbiological diagnostics. The application of MYcrobiota for routine clinical microbiological diagnostics use was evaluated by investigating a total of 63 clinical samples, including 40 (polymicrobial) clinical samples obtained from patients presenting a variety of damaged skin conditions (**Chapter 6**) and 23 low biomass clinical samples obtained from patients who were suspected to have bacterial septic arthritis (**Chapter 7**). The results obtained with MYcrobiota were compared to the results obtained with culture-based methods, which are the current 'gold standard' methods for pathogen detection in the routine clinical microbiological diagnostic laboratory. As shown in **Chapter 6**, 36 of the 38 aerobic bacteria identified within the damaged skin samples using routing culturing methods were also identified using MYcrobiota, although the majority of the 447 bacterial taxa identified using MYcrobiota were presumed to belong to the commensal flora or were not cultured at all. These results indicate that the resolution power of MYcrobiota was superior compared to the culture-based methods commonly used in routine clinical microbiological diagnostic laboratories. The vast majority of the additional bacteria identified using MYcrobiota represented bacteria that are obligate anaerobes, which are difficult to culture within a laboratory environment.[45] Importantly, anaerobic bacteria are a common cause of endogenous bacterial infections and their culture-free detection by MYcrobiota would provide clinicians with very useful information about the aetiologies of such infections that cannot be (easily) provided using routine culturing methods. It should be noted however that partial 16S rRNA genes, which are currently targeted using MYcrobiota, lack the discriminative power to differentiate prokaryotes to the species taxonomic level.[24] This species-level determination is often seen as essential for clinical diagnostics, as only specific species within a genus may be pathogenic. Importantly however, the identification of bacterial species using species-specific qPCRs is one way to circumvent this limitation (as described above), while other strategies would require relatively simple adjustments to be made to the MYcrobiota platform that enables the micelle-based amplification, sequencing and analysis of multiple hypervariable 16S rRNA gene regions,[46] or other genetic markers, such as *rpoB*,[47] *gyrB*,[48] the ITS region,[15] and many other candidates that enables taxonomic discrimination at the species-level.[49]

The availability of relative abundance results in combination with absolute abundance results using MYcrobiota generates a highly accurate and comprehensive overview of the microbial composition of polymicrobial clinical samples. This is in contrast to routine culturing methods, which are unlikely to provide a complete understanding of the microbial composition of a sample containing mixed microorganisms. In fact, routine culturing methods have primary been developed to select and identify only the 'established pathogens' (for which clinicians have associative clinical experience) and often disregard the abundance of microorganisms that are not classified as pathogens. For example, the only two discrepant (culture-positive and MYcrobiota-negative) microorganisms

measured in the damaged skin samples described in **Chapter 6** were identified as *Pseudomonas aeruginosa* and *Staphylococcus aureus*, with estimated relative abundances of 10% and 40% respectively, using routine culturing methods. Although both bacterial taxa were initially not reported by MYcrobiota, both correlating OTUs were detected using the calibrated micPCR/NGS method but below the method's LOD – the relative abundance for the *Pseudomonas* OTU was 0.14% and for the *Staphylococcus* OTU 0.07%. It is very likely that the abundance of these bacterial species are overestimated using routine culturing methods, as both established pathogens are fast-growing bacteria with a clearly distinguishable colony phenotype that are more easily selected compared to other microorganisms present within the same clinical sample.[50] This finding is further supported by a recent report describing the consistent overestimation of *S. aureus* abundance in cystic fibrosis sputa samples by culture-based methods compared to a 16S rRNA gene NGS method.[51] Importantly however, the question remains if these low abundant established pathogens – that are also frequently found in polymicrobial samples obtained from healthy individuals – are the true causative agents of an ongoing infection. Indeed, recent studies suggest that low relative abundant bacteria may contribute to pathogenesis by altering the microenvironment to facilitate colonization or virulence gene expression of neighbouring microorganisms, following the 'keystone pathogen' hypothesis, but more studies are needed to identify potential keystone pathogens in different kinds of polymicrobial clinical samples in order to determine the actual prevalence of these so-called 'keystone pathogen-induced inflammations' compared to 'dominant pathogen-induced inflammations'.[52] These studies would depend heavily on obtaining accurate and comprehensive microbiota profiles, such as generated by MYcrobiota, that may lay the groundwork to further establish keystone pathogen-induced inflammations using animal disease models and to explore the mechanisms by which these potential keystone pathogens mediate disease. Importantly, the identification of keystone pathogens, and the ability to detect them using MYcrobiota as a routine clinical-diagnostic test, would have significant benefits as it could facilitate accurate and more targeted prescription of antimicrobial treatments for polymicrobial or complex dysbiotic diseases.

The ability of MYcrobiota to remove contaminating DNA allows the accurate detection of potentially pathogenic microorganisms at very low abundances, or alternatively, the confirmation of culture-negative results. To illustrate this potential, we investigated 23 joint fluid samples obtained from 19 patients with suspected bacterial septic arthritis using routine culturing methods and MYcrobiota. As shown in **Chapter 7**, all joint fluid samples resulted in culture-negative results, whereas MYcrobiota detected the presence of bacterial DNA in 10 out of 23 joint fluid samples, whilst confirming the culture-negative results in the other 13 joint fluid samples. Most of the additionally bacteria detected using MYcrobiota have been previously described as non-conventional pathogens in

bacterial septic arthritis cases and included not only 'difficult-to-culture' bacteria, such as anaerobic bacteria (e.g. *Parvimonas*, *Prevotella*, and *Ruminococcus*) and fastidious bacteria (e.g. *Kingella* and *Ureaplasma*), but also unexpected and 'easy-culturable' bacteria, such as *Enterococcus* and *Turicella*. These findings indicate that MYcrobiota is a very useful platform that enables the culture-free detection (i.e. identification and quantification) of anaerobic, fastidious, and (unexpected) culturable microorganisms, which will greatly improve the identification of the bacteriological aetiology of infections such as bacterial septic arthritis. However, extensive clinical validation studies will be needed in order to validate the routine introduction of MYcrobiota into clinical diagnostic laboratories.

Environmental microbiological diagnostics. The universal applicability of MYcrobiota was evaluated by employing the method to assess drinking water quality. As shown in **Chapter 8**, a total of 30 drinking water samples were collected at consecutive locations along an operational drinking water distribution system (DWDS) during a 5-month period. Using MYcrobiota, we observed spatial and temporal microbial variations within the drinking water samples obtained. These variations were not detected using routine culturing methods. The ability to detect such microbial variations with high sensitivity will aid the evaluation of current and future treatment strategies, such as oxidation and filtration processes, that are generally applied by drinking water utilities to ensure the delivery of safe and high-quality drinking water.[53] Interestingly, MYcrobiota performed equally as well as flow cytometry (FCM) with regard to the measurements of bacterial dynamic trends over the water trajectory, with FCM being one of the techniques that is currently being proposed as a replacement for bacterial culture for the routine microbiological assessment of drinking water quality.[54] However, although FCM is certainly useful for counting the total and viable number of bacterial cells over the water trajectory,[55,56] the method does not provide taxonomic information about the prokaryotes detected within the DWDS. The identification and quantification of prokaryotic taxa is however required in order to adequately evaluate the complex nature of microbial communities within DWDS samples. For example, although only a small subset of seven bacterial genera were shown to dominate the specific DWDS investigated, the absolute abundances of these bacteria shifted across the DWDS, illustrating large differences in microbial community compositions between 'treated' and 'distributed' drinking water samples – due to the application of drinking water treatment processes at the beginning of the DWDS. In summary, this study demonstrates that the use of MYcrobiota enables the culture-independent monitoring of bacteria that resides within DWDSs. Further, MYcrobiota allows drinking water utility companies to obtain accurate measurements of spatial and temporal microbial dynamics within their DWDS, facilitating the continual assessment of the desired 'biological stability' of drinking water over the whole drinking water trajectory, thereby helping maintain drinking water quality standards.

CONCLUSIONS

1. MYcrobiota utilizes a novel micelle PCR/NGS methodology that limits both the formation of chimera sequences and PCR competition induced bias, thereby improving the accurate characterization of microbial communities.
2. MYcrobiota provides the relative abundances and the absolute abundances for each individual operational taxonomic unit (OTU) present within a sample, which enables the subtraction of any non-sample associated contaminating OTUs via the processing of negative extraction controls.
3. MYcrobiota uses a dedicated and easy-to-use bioinformatics pipeline that allows for a fully automated sequence interpretation of 16S rRNA gene NGS data that is obtained using the micelle PCR methodology without the requirement for advanced bioinformatics skills.
4. MYcrobiota generates a highly accurate and comprehensive overview of the microbial composition of clinical samples or, alternatively, confirms the absence of 16S rRNA gene copies in culture-negative clinical samples.
5. Although MYcrobiota was initially developed for use in the field of routine clinical microbiological diagnostics, the microbiota analysis platform is much more widely applicable as demonstrated in the analysis of microbial dynamics in an operational drinking water distribution system.

FUTURE PERSPECTIVES

MYcrobiota enables the accurate quantification of prokaryotic taxa within (polymicrobial) samples through the reduction of PCR amplification artefacts (chimera formation and PCR competition) and the removal of contaminating 16S rRNA gene molecules derived from the experimental set-up. However, the possible effects of sample handling,[57] DNA extraction,[58] and primer specificity,[59] are still factors affecting the complete accuracy of 16S rRNA gene NGS results, even with the development of MYcrobiota. In fact, these factors should preferably be optimized for each type of test sample the researcher is investigating in order to ensure the generation of truly unbiased microbiota profiles. In addition, the micelles used for micPCR are generated by mixing water in oil using a standardized commercially available kit in combination with a vortex. This (uncontrolled) process could lead to the generation of unevenly shaped or broken micelles that potentially introduce quantification bias as larger/intact micelles will result in more 16S rRNA gene amplicons compared to smaller/broken micelles. Although this limitation can be overcome by processing each sample in triplicate to average out any possible quantification bias using our current validated micPCR protocol, the robustness and

quantitative accuracy of micPCR would benefit from generating micelles (or equivalents) with a much higher precision and repeatability. Interestingly, generating highly reproducible micro-sized droplets using dedicated droplet generators are already available and additional studies are needed to investigate the utility of these droplets for 16S rRNA gene NGS microbiota profiling.[60,61]

Like all short-read 16S rRNA gene NGS methods, MYcrobiota currently lacks accurate prokaryotic identification at the species-level due to the lack of the discriminative power of the partial 16S rRNA gene used. This limitation can be overcome by incorporation of quantitative results obtained from additional species-specific qPCRs or through the implementation of specific genetic markers, as discussed previously. In addition, an alternative strategy to obtain prokaryotic species identification is by employing shotgun metagenomics approaches that have the potential to detect all genomic contents derived from all of the microorganisms (including bacteria, archaea, fungi, protists, and viruses) present in a test sample.[62] These methods not only allow for the (species-level) characterisation of microbial communities across all domains of life, but also provides knowledge on the population gene composition of microbial communities, such as the prevalence and complexity of antibiotic resistance genes within faecal samples.[63,64] This type of research application may eventually be important in clinical diagnostics, where for example, the rapid detection of antibiotic resistance genes could potentially improve the clinical decision-making process. To this end, multiple online platforms and bioinformatics tools are available that accept sequence data as queries and return predictions of their antibiotic resistance gene content, often with confidence-related statistics.[65,66]

In addition, shotgun metagenomics approaches provide information about the abundances of genes involved in functional pathways that, for example, can be explored for associations with diseases. These metagenome-wide association studies (MWAS) would provide signatures of health and disease in the microbiome that may be clearer at the functional level than at the taxonomic level.[67] For example, Qin et al. identified approximately 60,000 microbial markers that are associated with type 2 diabetes and demonstrated that a selection of 50 of these markers are able to distinguish between samples from healthy subjects and subjects with type 2 diabetes.[68] Importantly however, the desired progression from identifying microbiome-disease associations to identifying the functions of the microbiome in disease is currently hindered by a basic lack of functional characterization of the vast majority of microbial genes that are detected via shotgun metagenomics approaches. Therefore, future microbiome studies should aim to provide more insights into gene functions within human/environmental microbiomes in order to improve the power of MWAS. One elegant example of research addressing this issue has been recently published by Cohen et al. who 'mined' the human microbiota for genes encoding metabolites that mimic human signalling molecules.[69] Interestingly, the bacterial metabolites identified in this study were involved in host-microbial interactions

that potentially regulate human physiology in healthy and disease states and represent a possible resource for the discovery of small-molecule therapeutics. However, such shotgun metagenomics approaches remain expensive, computationally challenging when using short-read sequences and, for clinical samples, possess a low sensitivity due to the relative excess of human DNA compared to prokaryotic DNA.[70] These disadvantages mean that the introduction of shotgun metagenomics into the routine (clinical) microbiological diagnostic laboratory is not yet feasible.

Recently, a third generation of sequencing platforms, including PacBio (Pacific Biosystems) and MinION (Oxford Nanopore Techniques) have become commercially available. These sequencing instruments can overcome the limitations of short-read NGS-platforms, because they produce sequence reads between 5,000 and 15,000 nucleotides in length, which is much longer compared to those obtained with current NGS-platforms that produce sequence reads that span only a few hundred nucleotides.[71] Longer sequence reads simplify genome assemblies and also allows for a more reliable assignment of DNA sequences to *in silico* stored reference genomes compared to smaller sequence fragments.[72] In addition, the MinION platform collects and analyses sequence data in real-time, which can significantly shorten the time-to-result compared to other NGS-platforms.[73] Nonetheless, the applicability of shotgun metagenomics, or targeted amplicon sequencing approaches using third generation sequencing platforms is still far from certain, due to high costs per sample, low throughput and relatively high base-calling error rates.[74] Until then, MYcrobiota can conveniently fill the gap between traditional 'gold standard' microbiological methods (culture and PCR) and the as yet unfulfilled power of third generation sequencing-based metagenomics.

REFERENCES

1. Faith JJ, Guruge JK, Charbonneau M, et al. The long-term stability of the human gut microbiota. *Science* 2013; **341**: 1237439.
2. 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.
3. 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.
4. 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.
5. Smyth RP, Schlub TE, Grimm A, et al. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene* 2010; **469**: 45-51.
6. Edgar RC, Haas BJ, Clemente JC, et al. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; **27**: 2194-2200.
7. 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.
8. 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.
9. 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.
10. Cline J, Braman JC, Hogrefe HH. PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res* 1996; **24**: 3546-3551.
11. 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.
12. Sim K, Cox MJ, Wopereis H, et al. Improved detection of bifidobacteria with optimised 16S rRNA-gene based pyrosequencing. *PLoS One* 2012; **7**: e32543.
13. Kalle E, Kubista M, Rensing C. Multi-template polymerase chain reaction. *Biomol Detect Quantif* 2014; **2**: 11-29.
14. 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.
15. Sabat AJ, van Zanten E, Akkerboom V, et al. Targeted next-generation sequencing of the 16S-23S rRNA region for culture-independent bacterial identification - increased discrimination of closely related species. *Sci Rep* 2017; **7**: 3434.
16. 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.
17. Williams R, Peisajovich SG, Miller OJ, et al. Amplification of complex gene libraries by emulsion PCR. *Nat Methods* 2006; **3**: 545-550.
18. 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.

19. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol* 2016; **26**: 330-335.

20. Stammler F, Glasner J, Hiergeist A, et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 2016; **4**: 28.

21. Tourlousse DM, Yoshiike S, Ohashi A, et al. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res* 2017; **45**: e23.

22. Yang S, Lin S, Kelen GD, et al. Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. *J Clin Microbiol* 2002; **40**: 3449-3454.

23. Martineau F, Picard FJ, Roy PH, et al. Species-specific and ubiquitous-DNA-based assays for rapid identification of *Staphylococcus aureus*. *J Clin Microbiol* 1998; **36**: 618-623.

24. Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 2007; **10**: 504-509.

25. 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.

26. Mohammadi T, Reesink HW, Vandenbroucke-Grauls CMJE, et al. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *J Microbiol Methods* 2005; **61**: 285-288.

27. Rand KH, Houck H. *Taq* polymerase contains bacterial DNA of unknown origin. *Mol Cell Probes* 1990; **4**: 445-450.

28. Tanner MA, Goebel BM, Dojka MA, et al. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl Environ Microbiol* 1998; **64**: 3110-3113.

29. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12**: 87.

30. Corless CE, Borrow R, Edwards-Jones V, et al. Contamination and sensitivity issues with a real-time universal 16S rRNA PCR. *J Clin Microbiol* 2000; **38**: 1747-1752.

31. Hein I, Schneeweiss W, Stanek C, et al. Ethidium monoazide and propidium monoazide for elimination of unspecific DNA background in quantitative universal real-time PCR. *J Microbiol Methods* 2007; **71**: 336-339.

32. Humphrey B, McLeod N, Turner C, et al. Removal of contaminant DNA by combined UV-EMA treatment allows low copy number detection of clinically relevant bacteria using pan-bacterial real-time PCR. *PLoS One* 2015; **10**: e0132954.

33. Biesbroek G, Sanders EA, Roeselers G, et al. Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PloS one* 2012; **7**: e32942.

34. 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.

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

36. Nilakanta H, Drews KL, Firrell S, et al. A review of software for analyzing molecular sequences. *BMC Res Notes* 2014; **7**: 830.

37. Afgan E, Baker D, Van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016; **44**: W3–W10.

38. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005; **15**: 1451-1455.

39. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* 2010; Chapter 19: Unit 19.10.1-21.

40. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010; **11**: R86.

41. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 2011; **12**: 385.

42. Bik HM, Pitch Interactive. Phinch: an interactive, exploratory data visualization framework for -omic datasets. *bioRxiv* 2014: 009944.

43. Hiltemann S, Hoogstrate Y, van der Spek P, et al. iReport: a generalised Galaxy solution for integrated experimental reporting. *Gigascience* 2014; **3**: 19.

44. Deurenberg RH, Bathoorn E, Chlebowicz MA, et al. Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnol* 2017; **243**: 16-24.

45. Brook I. Clinical review: bacteremia caused by anaerobic bacteria in children. *Crit Care* 2002; **6**: 205-211.

46. 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.

47. Adekambi T, Drancourt M, Raoult D. The *rpoB* gene as a tool for clinical microbiologists. *Trends Microbiol* 2009; **17**: 37-45.

48. Yamamoto S, Harayama S. PCR amplification and direct sequencing of *gyrB* genes with universal primers and their application to the detection and taxonomic analysis of *Pseudomonas putida* strains. *Appl Environ Microbiol* 1995; **61**: 3768.

49. 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.

50. Dowd SE, Wolcott RD, Sun Y, et al. Polymicrobial nature of chronic diabetic foot ulcer biofilm infections determined using bacterial tag encoded FLX amplicon pyrosequencing (bTEFAP). *PLoS One* 2008; **3**: e3326.

51. 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.

52. Hajishengallis G, Darveau RP, Curtis MA. The keystone-pathogen hypothesis. *Nat Rev Microbiol* 2012; **10**: 717-725.

53. Prest EI, Hammes F., van Loosdrecht MCM, et al. Biological stability of drinking water: controlling factors, methods, and challenges. *Front Microbiol* 2016; **7**: 45.

54. Van Nevel S, Koetzsch S, Proctor CR, et al. Flow cytometric bacterial cell counts challenge conventional heterotrophic plate counts for routine microbiological drinking water monitoring. *Water Res* 2017; **113**: 191-206.

55. Prest EI, Hammes F, Kötzsch S, et al. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. *Water Res* 2013; **47**: 7131-7142.

56. Prest EI, El-Chakhtoura J, Hammes F, et al. Combining flow cytometry and 16S rRNA gene pyrosequencing: a promising approach for drinking water monitoring and characterization. *Water Res* 2014; **63**: 179-189.

57. Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol* 2012; **12**: 158.

58. 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.

59. Mao DP, Zhou Q, Chen CY, et al. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol* 2012; **12**: 66.

60. Hindson CM, Chevillet JR, Briggs HA, et al. Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods* 2013; **10**: 1003-1005.

61. Hindson BJ, Ness KD, Masquelier DA, et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* 2011; **83**: 8604-8610.

62. Miller RR, Montoya V, Gardy JL, et al. Metagenomics for pathogen detection in public health. *Genome Med* 2013; **5**: 81.

63. Andersen H, Connolly N, Bangar H, et al. Use of shotgun metagenome sequencing to detect fecal colonization with multidrug-resistant bacteria in children. *J Clin Microbiol* 2016; **54**: 1804-1813.

64. Hu Y, Yang X, Qin J, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun* 2013; **4**: 2151.

65. Xavier BB, Das AJ, Cochrane G, et al. Consolidating and exploring antibiotic resistance gene data resources. *J Clin Microbiol* 2016; **54**: 851-859.

66. Bradley P, Gordon NC, Walker TM, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* 2015; **6**: 10063.

67. Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. *Nat Rev Microbiol* 2016; **14**: 508-522.

68. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012; **490**: 55-60.

69. Cohen LJ, Esterhazy D, Kim SH, et al. Commensal bacteria make GPCR ligands that mimic human signaling molecules. *Nature* 2017; **549**: 48-53.

70. Hilton SK, Castro-Nallal E, Pérez-Losada M, et al. Metataxonomic and metagenomic approaches vs. culture-based techniques for clinical pathology. *Front Microbiol* 2016; **7**: 484.

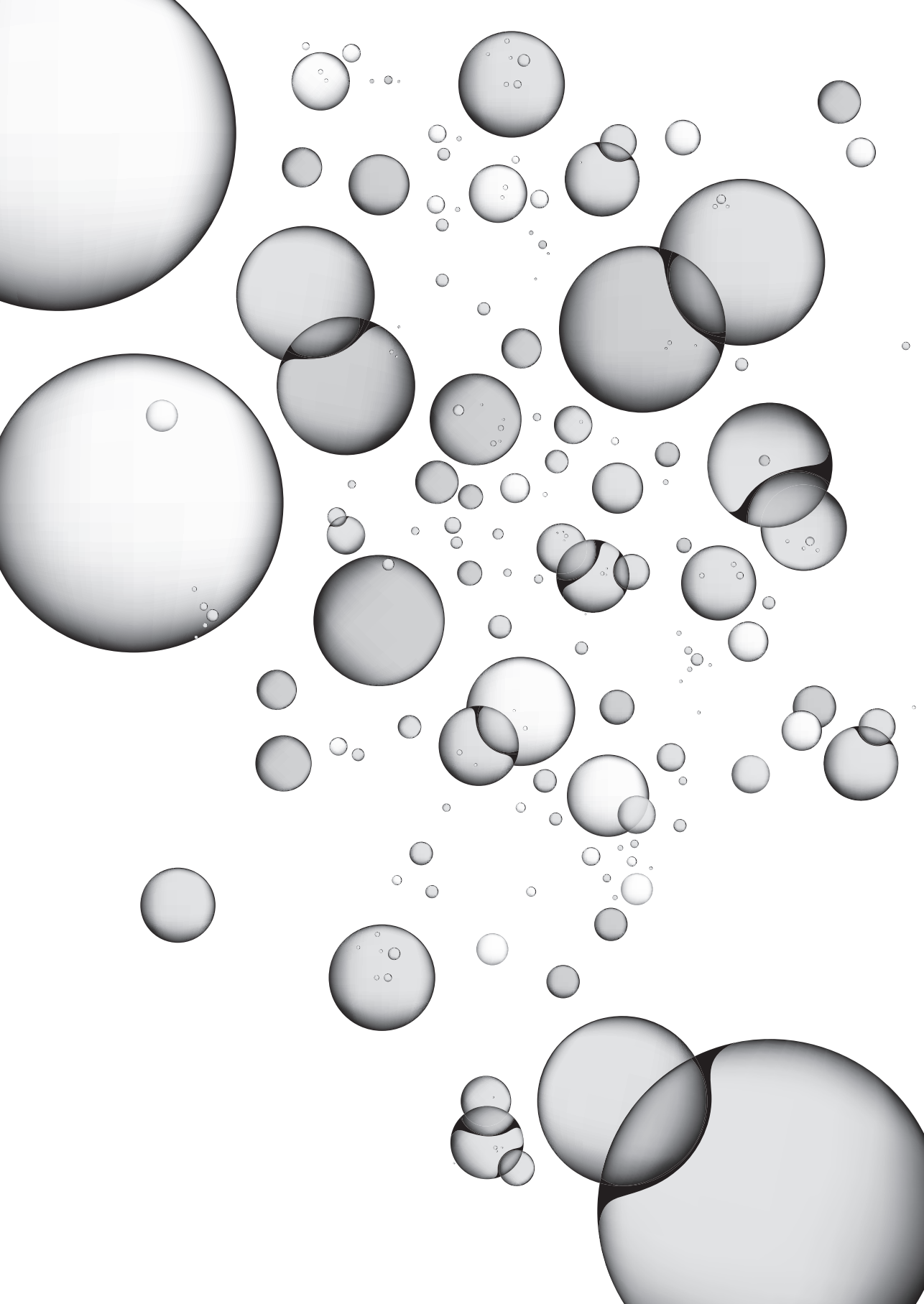
71. Eisenstein M. Startups use short-read data to expand long-read sequencing market. *Nat Biotechnol* 2015; **33**: 433-435.

72. 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.

73. Quick J, Ashton P, Calus S, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* 2015; **16**: p. 114.

74. Brown BL, Watson M, Minot SS, et al. MinION nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience* 2017; **6**: 1-10.

75. 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.



Chapter 10

Nederlandse samenvatting

Micro-organismen, zoals virussen, bacteriën, archaea en andere eencelligen, vormen complexe leefgemeenschappen die op alle denkbare plaatsen op aarde aanwezig zijn, van de diepzee tot onze darm en van de bosbodem tot onze huid. Ze zijn de oudste en meest uitbundige levensvorm op aarde, zowel in hoeveelheid als diversiteit, en vervullen vele belangrijke functies. Zo nemen micro-organismen deel aan biochemische processen die het leven op onze planeet mogelijk maken. Ook is het inmiddels algemeen bekend dat de microbiële gemeenschappen die in en op ons lichaam voorkomen biochemisch van groot belang zijn en cruciaal zijn voor onze gezondheid. Om deze reden wordt de microbiële samenstelling, die ook wel de 'microbiota' wordt genoemd, intensief bestudeerd. Onderzocht wordt wat de samenstelling van de microbiota van verschillende lichaamslocaties is en men probeert de invloed van de microbiota op onze gezondheid vast te stellen. De bacteriële samenstelling van de microbiota wordt veelal onderzocht door het 16S-rRNA-gen te onderzoeken. Het 16S-rRNA-gen heeft de unieke eigenschap dat het aanwezig is in alle bacteriën (en archaea). Tevens bevat dit gen naast geconserveerde DNA-gebieden ook variabele DNA-gebieden die uniek zijn per bacteriesoort en dus gebruikt kunnen worden voor bacteriële identificatie. De variabele DNA-gebieden worden veelal in kaart gebracht door ze eerst te vermeerderen (i.e. amplificeren) met behulp van PCR-technieken om vervolgens de volgorde van de bouwstenen van het DNA te bepalen middels next-generation sequencing (NGS) technieken. Deze werkwijze stelt men in staat om de microbiota in een monster te onderzoeken zonder de afzonderlijke bacteriën te kweken. Dit laatste is een belangrijk gegeven aangezien de meeste bacteriën niet kweekbaar zijn in een laboratorium. Echter, deze kweek-onafhankelijke technieken brengen weer andere beperkingen en valkuilen met zich mee die de rapportage van betrouwbare microbiota resultaten in de weg kan staan (**hoofdstukken 1** en **2**). Eén van de grootste uitdagingen is de standaardisatie van de 16S-rRNA-gen NGS-methoden in de verschillende laboratoria. Standaardisatie is essentieel, want alleen met gestandaardiseerde werkwijzen kunnen de data van verschillende laboratoria vergeleken worden waardoor er een overkoepelende kwaliteitsbewaking kan worden ingevoerd. Deze vorm van kwaliteitsbewaking is belangrijk voor research doeleinden, maar vooral ook voor de toepassing van microbiota onderzoek in routine (medisch) microbiologisch diagnostisch gebruik. Het doel van het onderzoek beschreven in dit proefschrift is dan ook om een gestandaardiseerde 16S-rRNA-gen NGS-methode te ontwikkelen en de toepasbaarheid van deze methode te onderzoeken in de routine (medisch) microbiologische diagnostiek.

De ontwikkeling van een gestandaardiseerde 16S-rRNA-gen NGS-methode vereist een verbetering van de nauwkeurigheid en reproduceerbaarheid van huidige 16S-rRNA-gen NGS-methoden. De veelgebruikte 16S-rRNA-gen NGS-methoden zijn afhankelijk van de PCR-techniek, waarbij (verschillende) 16S-rRNA-genen gezamenlijk, in één enkel reactievaatje, geamplificeerd worden. Deze methode kent twee belangrijke nadelen: 1) het

ontstaan van chimeren en 2) PCR-competitie tussen verschillende 16S-rRNA-genen van verschillende bacteriën. Chimeren zijn samengesteld uit 16S-rRNA-genen, afkomstig van verschillende bacteriën, die leiden tot de foutieve detectie van niet-aanwezige of zelfs van niet-bestaande bacteriën. De PCR-competitie, waarbij het ene 16S-rRNA-gen sneller amplificeert dan het andere 16S-rRNA-gen, heeft een over- of onderschatting van bepaalde bacteriën als gevolg. Om deze PCR-artefacten (chimeren en PCR-competitie) te voorkomen is in **hoofdstuk 3** een nieuwe amplificatiemethode beschreven waarbij de 16S-rRNA-genen verdeeld worden over een groot aantal fysiek verschillende 'reactievaatjes' die gecreëerd worden door middel van water-in-olie emulsies. Tijdens deze 'micelle PCR' (micPCR) wordt vervolgens elk afzonderlijk DNA-molecuul met een 16S-rRNA-gen individueel geamplificeerd, elk molecuul in één druppel van de emulsie. In **hoofdstuk 3** wordt aangetoond dat deze benadering de kans op PCR-artefacten drastisch verminderd en de nauwkeurigheid van microbiota bepalingen sterk verbetert.

In **hoofdstuk 4** werd het micPCR/NGS-protocol verder uitgebreid met de toevoeging van een interne kalibrator die het mogelijk maakt om de hoeveelheid van de gedetecteerde bacteriën weer te geven als het aantal 16S-rRNA-genkopieën. Hiermee kan de micPCR/NGS-methode een weerbarstig probleem van microbiota studies oplossen, namelijk het probleem van het achtergrondsignaal van het al aanwezige bacterieel DNA uit de gebruikte reagentia. Het achtergrond signaal is met name belangrijk voor de analyse van monsters met een lage hoeveelheid bacteriën. De hoge nauwkeurigheid waarmee de gekalibreerde micPCR/NGS-methode de bacteriën kwantificeert maakt het nu mogelijk om het achtergrondsignaal te identificeren en te kwantificeren, waarmee de microbiota resultaten van het monster kunnen worden gecorrigeerd. De validatie-experimenten beschreven in **hoofdstuk 4** tonen aan dat deze correctie leidt tot de volledige verwijdering van het achtergrondsignaal van de microbiota resultaten van monsters met een bekende microbiële samenstelling. Deze strategie maakt het mogelijk om de microbiota van monsters met een erg lage biomassa nauwkeurig te kunnen bepalen. Bovendien zullen monsters die geen bacterieel DNA bevatten ook daadwerkelijk een negatief microbiota resultaat opleveren.

De in **hoofdstukken 3** en **4** beschreven micPCR/NGS-methode vraagt het nodige rekenwerk voordat het eindresultaat verkregen wordt. Om de microbiota bepalingen toegankelijk te maken voor routine gebruik is in **hoofdstuk 5** de ontwikkeling van de 'Galaxy mothur Toolset (GmT)' beschreven. Deze GmT bevat meer dan 125 verschillende algoritmen, elk met vooraf ingestelde 'standaard' instellingen, waarmee de gebruiker zelf zijn eigen 16S-rRNA-gen NGS-data 'analyse-workflow' kan samenstellen middels een gebruiksvriendelijke webinterface. Daarnaast biedt GmT ondersteuning voor diverse visualisatie- en rapportagehulpmiddelen zodat de microbiota resultaten eenvoudig kunnen worden geïnterpreteerd, gedeeld en opgeslagen. Voor het verwerken van de gekalibreerde micPCR/NGS-data is er een aangepaste GmT analyse-workflow

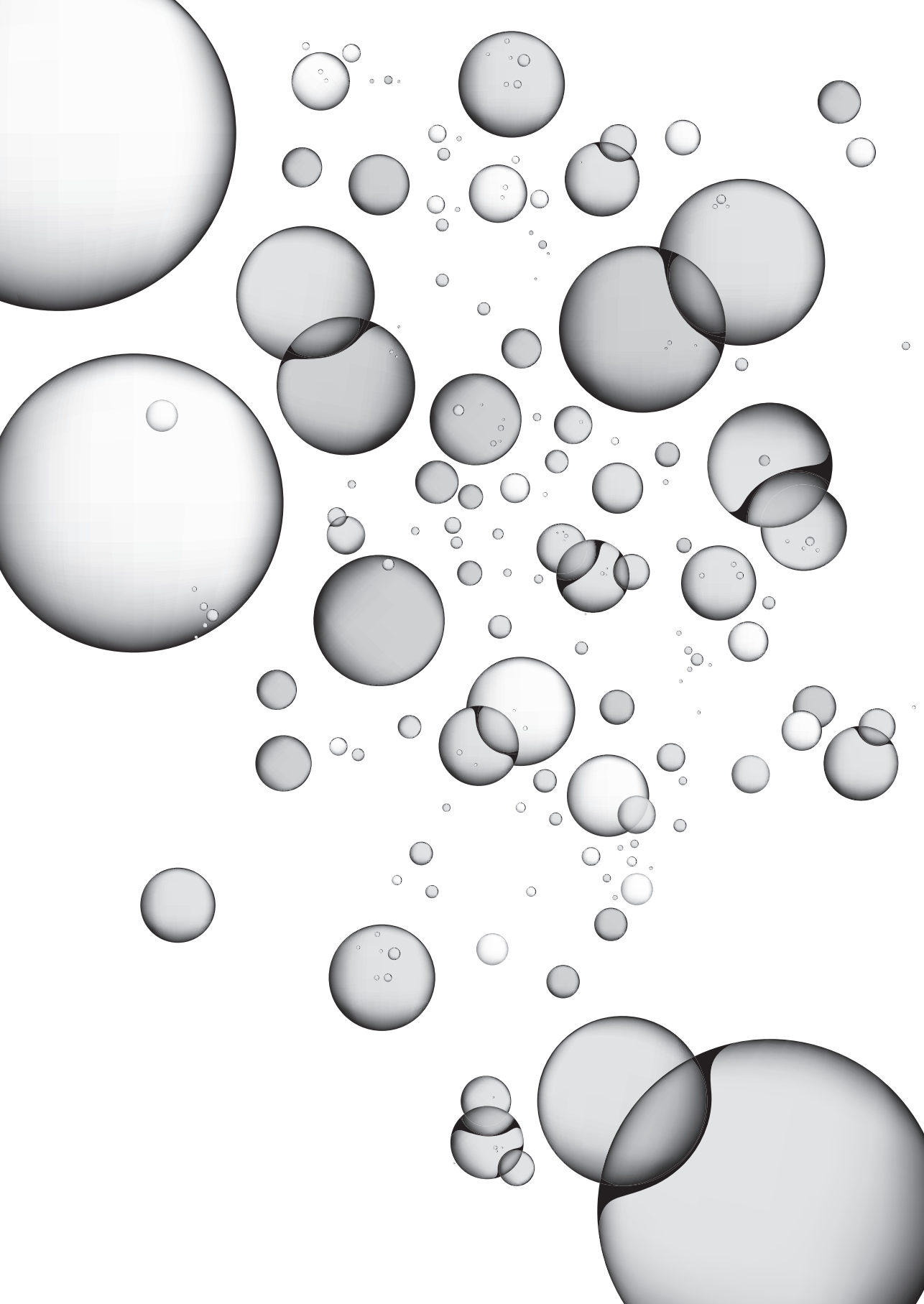
ontwikkeld waarbij de ruwe NGS-data met één druk op de knop kan worden verwerkt tot kwantitatieve microbiota resultaten (**hoofdstuk 6**). Deze op maat gemaakte GmT analyse-workflow maakt samen met de gekalibreerde micPCR/NGS-methode deel uit van het gestandaardiseerde microbiota detectieplatform: 'MYcrobiota'.

In de **hoofdstukken 6** en **7** werd het gebruik van MYcrobiota in de routine medisch microbiologische diagnostiek geëvalueerd waarbij de MYcrobiota resultaten vergeleken werden met de resultaten van routine kweekmethoden. In **hoofdstuk 6** werden pus- en wondmonsters onderzocht en in **hoofdstuk 7** onderzochten we gewrichtsvochten van ontstoken heup-, schouder- en kniegewrichten. Deze studies toonden aan dat de meeste gekweekte pathogene bacteriën ook werden geïdentificeerd met MYcrobiota. Tevens kan met MYcrobiota de normale (niet-pathogene) flora en de niet- of moeilijk-kweekbare bacteriën in kaart gebracht worden. Onder de niet- of moeilijk-kweekbare bacteriën, werden vooral zuurstof intolerante (obligaat anaerobe) bacteriën aangetroffen. Aangezien MYcrobiota niet gevoelig is voor bacterie-specifieke kweekcondities kan dit platform dan ook een belangrijke rol spelen bij het aantonen van potentieel pathogene bacteriën in anaerobe infecties, of andere infecties veroorzaakt door niet- of moeilijk-kweekbare bacteriën, die niet worden gedetecteerd met routine kweekmethoden. Zo bleken 10 van de 23 gewrichtsvochten, waaruit geen bacteriën kon worden gekweekt, toch positief voor bacterieel DNA met MYcrobiota die het vermoeden van de arts op bacteriële septische artritis in deze patiënten bevestigde (**hoofdstuk 7**). Het moet echter wel opgemerkt worden dat de gedeeltelijke 16S-rRNA-genen, die momenteel gebruikt worden voor de microbiële identificatie middels MYcrobiota, niet voldoen om bacteriën te classificeren op het species niveau wat zeer wenselijk is voor de medische toepassing. Het gebruik van andere genen, zoals *rpoB* of *gyrB* die wel dit taxonomisch onderscheidt kunnen maken, zou uitkomst kunnen bieden, maar behoeft wel enkele (relatief eenvoudige) aanpassingen aan het MYcrobiota-platform.

Het MYcrobiota-platform is ontwikkeld voor toepassing in de medische microbiologie. In **hoofdstuk 8** wordt het universeel toepasbare karakter van het MYcrobiota-platform aangetoond door het bepalen van de bacteriële samenstelling van drinkwater. In dit onderzoek werden drinkwatermonsters geanalyseerd van zes opeenvolgende locaties binnen een operationeel drinkwater-distributiesysteem (DWDS). In tegenstelling tot de gangbare kweekmethode, was MYcrobiota in staat om (kleine) bacteriële variaties te detecteren tussen de verschillende locaties binnen het DWDS. Deze variaties konden worden verklaard als gevolg van de drinkwater-behandelingsstrategieën die aan het begin van het DWDS werden toegepast om de levering van veilig en hoogwaardig drinkwater te waarborgen. Het vermogen om dergelijke variaties te detecteren met behulp van MYcrobiota kan daarom van groot belang zijn voor de continue monitoring van de effecten van de toegepaste drinkwater-behandelingsstrategieën en biedt drink-

waterbedrijven de mogelijkheid om in te grijpen als de beoogde 'biologische stabiliteit' binnen het DWDS verstoord wordt.

Samenvattend, de ontwikkeling van het MYcrobiota-platform beschreven in dit proefschrift realiseert een gestandaardiseerde en gevalideerde bepaling van de microbiota. MYcrobiota is ontwikkeld voor de medische microbiologie, maar is veel breder toepasbaar zoals aangetoond in het drinkwateronderzoek. Verwacht wordt dat de implementatie van MYcrobiota een bijdrage kan leveren aan onze kennis van de complexe microbiële gemeenschappen die overal om ons heen te vinden zijn.



Appendices

Dankwoord

Curriculum Vitae

List of publications

PhD portfolio

DANKWOORD

Na een intensieve periode van 4 jaar is het zover. Met het schrijven van dit dankwoord leg ik de laatste hand aan mijn proefschrift. Graag wil ik alle collega's van zowel het Streeklaboratorium Haarlem als het Erasmus MC, familie en vrienden bedanken die mij de afgelopen jaren – direct of indirect – hebben geholpen gedurende mijn promotietraject. Ook wil ik de medeauteurs van de verschillende onderzoeken bedanken voor hun bijdrage aan dit proefschrift, op welke manier dan ook. De volgende personen wil graag in het bijzonder bedanken:

Dr. Jansen, beste Ruud. Jij hebt me door de jaren heen opgeleid van jonge analist naar de onderzoeker die ik nu ben, met als resultaat dit proefschrift. We werken al meer dan 12 jaar samen en gedurende deze tijd heb je mij geleerd om met een kritische houding onderzoek te doen. Ik heb het altijd erg prettig gevonden om met jou te brainstormen over nieuwe ideeën, of om gewoon een kletspraatje te maken, die bij voorkeur plaatsvonden onder het genot van een glaasje wijn en een stukje kaas. Het was fijn dat we vaak op dezelfde golflengte zaten. Ik ben je erg dankbaar voor jouw begeleiding en immer kritische blik, niet alleen op het onderzoek, maar eigenlijk op alles.

Dr. Hays, beste John. Ik ben je enorm dankbaar dat je mij de kans hebt gegeven om een promotietraject te starten binnen het Erasmus MC. Je was altijd 'kritisch-enthousiast' als ik nieuwe resultaten kwam laten zien en toonde daarnaast ook persoonlijke belangstelling. Bedankt voor al de vrijheid die jij me hebt gegeven tijdens het promotietraject en het vertrouwen dat je in me had. Misschien een beetje cliché, maar de snelheid waarmee jij de manuscripten nakeek is ongekend! Ik bewonder je passie voor het onderzoek en de gedrevenheid en ambitie waarmee jij jouw vak uitvoert. Mogen er nog vele samenwerkingen volgen.

Prof.dr. Mouton, beste Johan. Graag wil ik jou bedanken voor jouw rol tijdens de laatste fase van mijn promotietraject, het corrigeren van de laatste stukken van dit proefschrift en het in goede banen leiden van de papierwinkel rondom de promotie.

Leden van de promotiecommissie. Prof.dr. Boucher, Prof.dr. Savelkoul en Dr. Melchers wil ik bedanken voor het plaatsnemen in de kleine commissie en het beoordelen van mijn proefschrift. Ook de overige leden van de commissie wil ik uiteraard bedanken voor hun tijd, aandacht en oppositie.

Collega's van het Streeklaboratorium Haarlem. Na mijn vertrek in 2013 kwam ik snel en onverwacht terug op het 'oude nest' om het praktische werk dat beschreven is in dit proefschrift uit te voeren. Ik wil jullie bedanken voor de goede werksfeer en collegialiteit gedurende deze jaren. Het was een erg gezellige tijd, zowel op de werkvloer als tijdens de vrijdagmiddagborrels, barbecues, feestjes en etentjes. Dankjewel Cock, Chau, Debby, Elly†, José, Katja, Linda, Kenny, Mike, Marjolein, Nicole, Paul, Rosanna en Sandra. Naast de analisten van de afdeling moleculaire biologie wil ik ook graag de wetenschappelijke staf, artsen-microbioloog en managementleden bedanken: Bjorn, Dick, Ed, Jan, Jayant, Paul, Rob, Sinwen, Theo en Willem. Jullie hebben het mede mogelijk gemaakt dat ik mijn praktische werk met veel plezier kon uitvoeren binnen het Streeklaboratorium Haarlem en ik heb tevens veel geleerd van de vakinhoudelijke discussies die ik met jullie heb mogen voeren. Sjoerd en Wil, ik wil jullie bedanken voor de statistische ondersteuning die jullie mij tijdens mijn promotietraject hebben gegeven. Door jullie enthousiasme en nieuwsgierigheid hebben jullie ook zeker bijgedragen aan de inhoud.

Collega's van het Erasmus MC. Beste (oud) kamergenoten van Na-902 en onderzoekers en technisch personeel uit 'Nb/Nc', ik wil jullie allemaal bedanken voor de gezelligheid en de gesprekken over van alles en nog wat als ik weer eens langs kwam buurten in Rotterdam. Veel succes in de toekomst met het (afronden van promotie) onderzoek! Andrew and Saskia, I want to thank both of you for the important role you have played in the research that is described in this thesis. Although our discussions could last for (many) hours, you always kept it interesting with an emphasis on humour. I must thank you for both your patience with me and my constant changing wishes regarding the bioinformatics analysis of our data. I'm happy that I have worked with a great and dedicated team as your Bioinformatics department and I hope that our collaboration will continue in the future.

Vrienden en familie. Ik ben blij dat ik mij gelukkig mag prijzen met een aantal hele goede vrienden. Erik, Frank, Martijn, Ruud en Anne, Tom en Nelli, Debbie, Jacqueline, Marjolijn en Marleen: bedankt voor de interesse in mijn promotieonderzoek, maar ook zeker voor de jaarlijkse skivakanties en andere gezellige en minder sportieve momenten, waarop ik even helemaal niet bezig hoefde te zijn met mijn promotietraject. Uiteraard wil ik ook mijn ouders, zussen en schoonfamilie bedanken voor alle support en jullie blijvende interesse in mijn onderzoek! Sarita en Ruud, erg fijn dat jullie mij willen bijstaan als paranimfen tijdens de promotieplechtigheid.

Dayenne. Ondanks dat mijn promotieonderzoek een totaal ander vakgebied is dan je eigen, wist je mij altijd te helpen wanneer ik weer eens mijn (chaotische) gedachten moest vertalen naar een goedlopend manuscript. Op de juiste momenten wist jij me aan te moedigen en ik weet zeker dat dit proefschrift zonder jouw liefde, hulp en interesse nooit zo mooi was geworden. Bedankt voor al je support en ik ben blij dat ik dit onderzoek heb kunnen doen met jou aan mijn zijde.

CURRICULUM VITAE

Stefan Alexander Boers was born on the 7th of May 1984 in Oldenzaal, the Netherlands. After finishing his secondary education (HAVO) at the Twents Carmel College De Thij in Oldenzaal in 2001, he started studying for a BAS degree in Biology and Medical Laboratory Research at the Saxion Hogeschool Enschede. During this time, he specialized in molecular biology by performing research on the lactic acid bacterium *Lactobacillus plantarum* at NIZO Food Research and the opportunistic pathogen *Legionella pneumophila* at the National Institute for Public Health and the Environment (RIVM). He graduated in 2005 and in the same year he started working as a microbiological technician at the Regional Laboratory of Public Health Kennemerland in Haarlem (SLH). There, he was responsible for the culture-based detection of *Legionella* spp. and the molecular typing of patient isolates and environmental strains of *Legionella* as part of the National Legionella Outbreak Detection Program. As well as being responsible for the *Legionella*-work, he took part in the Employees Council and performed internal audits on a regular basis. To further develop his knowledge and professional skills, he combined his work at the SLH with obtaining a MSc degree in Biomolecular Sciences at the VU University Amsterdam, which started in 2009. As part of this master study, he performed two internships at the SLH under the supervision of Dr. Ruud Jansen. During the first internship, he successfully implemented a *bcr/abl* RT-PCR for the molecular diagnosis of the Philadelphia chromosome translocation in the routine practice of the clinical diagnostic laboratory and for his second internship he developed a high-throughput multilocus sequence typing method ('HiMLST') by adapting current next-generation sequencing (NGS) technology. To attain his MSc degree, he also wrote a Master thesis on CRISPR-Cas entitled 'CRISPR-Cas: An RNA-directed Adaptive Immunity System in Prokaryotes'. Following his graduation (*cum laude*) in 2012, he was appointed as senior molecular technician at the SLH and was responsible for the development and implementation of new approaches to aid the routine molecular diagnostics of infectious diseases. In 2013, he accepted a PhD position and began working in the Department of Medical Microbiology and Infectious Diseases at the Erasmus University Medical Centre (Erasmus MC), Rotterdam, performing research under the supervision of Dr. John Hays (Erasmus MC) and Dr. Ruud Jansen (SLH). During his PhD studies, he was part of a consortium of researchers participating in the European Union FP7-funded project 'TAILORED-Treatment', whose main goal was to establish a broad-based strategy that could be used to better target antibiotic prescribing to patients in cases of respiratory tract infection and sepsis. As a lead researcher in the TAILORED-Treatment consortium, Stefan was responsible for the development of a complete microbiota profiling workflow using NGS techniques and managed the quality of NGS results used in the project. In 2015, he took a leave of absence to volunteer for a period of 6 weeks as team leader at the Dutch

Mobile Ebola Lab Koidu, Sierra Leone, to diagnose Ebola infection during the 'West African Ebola virus epidemic (2013-2016)'. Here, he gained experience and management skills in the diagnostic testing and management of Advisory Committee on Dangerous Pathogens (ACDP) Category 4 infectious material. Stefan currently works at the Leiden University Medical Centre (LUMC), Leiden, under the supervision of Dr. Eric Claas, where he has started a 2-year training programme to become a qualified Medical Molecular Microbiologist (MMM).

LIST OF PUBLICATIONS

- Lück PC, Hahn F, Senger M, **Boers SA**, Brandsema P. European network cooperation to identity hotel as source for pneumonia caused by *Legionella pneumophila* serogroup 2. *Euro Surveill* 2008; 13: 18903.

- **Boers SA**, van Ess I, Euser SM, Jansen R, Tempelman FR, Diederen BM. An outbreak of a multiresistant methicillin-susceptible *Staphylococcus aureus* (MR-MSSA) strain in a burn centre: the importance of routine molecular typing. *Burns* 2011; 37: 808-813.

- **Boers SA**, van der Reijden WA, Jansen R. High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS One* 2012; 7: e39630.

- Euser SM, Bruin JP, Brandsema P, Reijnen L, **Boers SA**, Den Boer JW. Legionella prevention in the Netherlands: an evaluation using genotype distribution. *Eur J Clin Microbiol Infect Dis* 2013; 32: 1017-1022.

- Bruin JP, Kostrzewa M, van der Ende A, Badoux P, Jansen R, **Boers SA**, Diederen BMW. Identification of *Haemophilus influenza* and *Haemophilus haemolyticus* by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Eur J Clin Microbiol Infect Dis* 2014; 33: 279-284.

- Haenen OLM, van Zanten E, Jansen R, Roozenburg I, Engelsma MY, Dijkstra A, **Boers SA,** Voorbergen-Laarman M, Moller AVM. *Vibrio vulnificus* outbreaks in Dutch eel farms since 1996: strain diversity and impact. *Dis Aquat Organ* 2014; 108: 201-209.

- Souverein D, **Boers SA**, Veenendaal D, Euser SM, Kluytmans J, Den Boer JW. Polyclonal spread and outbreaks with ESBL positive gentamicin resistant *Klebsiella* spp. in the region Kennemerland, The Netherlands. *PLoS One* 2014; 9: e101212.

- **Boers SA**, Burggrave R, van Westreenen M, Goessens WHF, Hays JP. Whole-genome mapping for high-resolution genotyping of *Pseudomonas aeruginosa*. *J Microbiol Methods* 2014; 106: 19-22.

- **Boers SA**, Hays JP, Jansen R. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Sci Rep* 2015; 5: 14181.

- **Boers SA**, Jansen R, Hays JP. Suddenly everyone is a microbiota specialist! *Clin Microbiol Infect* 2016; 22: 581-582.

- Voor in 't Holt AF, Wattel AA, **Boers SA**, Jansen R, Hays JP, Goessens WHF, Vos MC. Detection of healthcare-related extended-spectrum beta-lactamase-producing *Escherichia coli* transmission events using combined genetic and phenotypic epidemiology. *PLoS One* 2016; 11: e0160156.

- **Boers SA**, Hays JP, Jansen R. Novel micelle PCR-based method for accurate, sensitive and quantitative microbiota profiling. *Sci Rep* 2017; 7: 45536.

- **Boers SA**, de Zeeuw M, Jansen R, van der Schroeff MP, van Rossum AMC, Hays JP, Verhaegh SJC. Characterization of the nasopharyngeal and middle ear microbiota in gastroesophageal reflux-prone versus gastroesophageal reflux non-prone children. *Eur J Clin Microbiol Infect Dis* 2018; 37: 851-857.

- **Boers SA***, Hiltemann SD*, Stubbs AP, Jansen R, Hays JP. Development and evaluation of a culture-free microbiota profiling platform (MYcrobiota) for clinical diagnostics. *Eur J Clin Microbiol Infect Dis* 2018; 37: 1081-1089.

- van Houten C, Oved K, Eden E, Cohen A, Engelhad D, **Boers SA**, Kraaij R, Karlsson R, Fernandez D, Gonzalez E, Li Y, Stubbs A, Moore E, Hays J, Bont L. Observational multi-centre, prospective study to characterize novel pathogen- and host-related factors in hospitalized patients with lower respiratory tract infections and/or sepsis – The "TAILORED-Treatment" Study. *BMC Infect Dis* 2018; 18: 377.

- Croughs PD, Klaassen CHW, van Rosmalen J, Maghdid DM, **Boers SA**, Hays JP, Goessens WHF, on behalf of the Dutch Antibiotic Resistance Surveillance Group. Unexpected mechanisms of resistance in Dutch *Pseudomonas aeruginosa* isolates collected during fourteen years of surveillance. *Int J Antimicrob Agents* 2018; in press.

- **Boers SA**, Reijnen L, Herpers BL, Hays JP, Jansen R. Detection of bacterial DNA in septic arthritis samples using the MYcrobiota platform. *J Clin Rheumatol* 2018; in press.

- **Boers SA**, Prest EI, Taučer-Kapteijn M, Knezev A, Schaap PG, Hays JP, Jansen R. Monitoring of microbial dynamics in a drinking water distribution system using the culture-free, user-friendly, MYcrobiota platform. *Sci Rep* 2018; in press.

- Hiltemann SD*, **Boers SA***, van der Spek PJ, Jansen R, Hays JP, Stubbs AP. Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur. *Submitted for publication*.

- van der Weide H, ten Kate MT, Vermeulen-de Jongh D, van der Meijden A, Wijma R, **Boers SA**, Hays JP, Goessens WHF, Bakker-Woudenberg IAJM. Successful monotherapy of tigecycline compared to meropenem in ESBL- and KPC-positive *Klebsiella pneumoniae* pneumonia models in rats. *Submitted for publication*.

- Heikema AP, Horst-Kreft D, **Boers SA**, Jansen R, de Ridder MAJ, van Houten CB, Bont LJ, Stubbs AP, Hays JP. Nanopore 16S rRNA gene sequencing of the human nasal microbiota indicates the (bioinformatics) loss of established genera and shows high bacterial species diversity. *Submitted for publication*.

- **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*.

* These authors contributed equally to the work.

PHD PORTFOLIO

Name: Stefan A. Boers
Institute: Erasmus University Medical Centre
Department: Medical Microbiology and Infectious Diseases
PhD period: 2013 – 2017
Promotor: Prof.dr. J.W. Mouton
Co-promotor: Dr. J.P. Hays
Dr. R. Jansen

In-depth courses	**Year**
- Partek Course on Microarray and NGS, MolMed, Rotterdam, the Netherlands	2014
- Metagenomics Approaches and Data Analysis, NBIC, Nijmegen, the Netherlands	2014
- Course on Whole Genome Sequencing, MolMed, Rotterdam, the Netherlands	2014
- A broad spectrum of NGS Applications in Molecular Medicine, MolMed, Rotterdam, the Netherlands	2014
- Research Integrity, Erasmus MC, Rotterdam, the Netherlands	2017

National and International conferences	**Year**
- Introduction to Microbiota Research, Glasgow, Scotland (oral presentation)	2014
- 19th Molecular medicine day, Rotterdam, the Netherlands (poster presentation)	2015
- Scientific spring meeting KNVM & NVMM, Papendal, the Netherlands (oral presentation)	2015
- 9th European Meeting on Molecular Diagnostics (EMMD), Noordwijk aan Zee, the Netherlands (oral presentation)	2015
- Microbiology: What can NGS do for us, Utrecht, the Netherlands	2015
- Applied Bioinformatics in Life Sciences, Leuven, Belgium (poster presentation)	2016
- Scientific spring meeting KNVM & NVMM, Papendal, the Netherlands (poster presentation)	2016
- 26th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID), Amsterdam, the Netherlands (eposter mini oral presentation)	2016
- 2nd Annual Microbiology & Immunology Virtual Conference (poster presentation)	2016
- 4th Microbiome R&D and Business Collaboration Forum: Europe, Amsterdam, the Netherlands (poster presentation)	2017
- Scientific spring meeting KNVM & NVMM, Papendal, the Netherlands (poster presentation)	2017
- 7th Congress of European Microbiologist (FEMS), Valencia, Spain (oral and poster presentation)	2017

Scientific meetings	**Year**
- Departmental Journal Clubs (oral presentations)	2013-2017
- Departmental Research meetings (oral presentations)	2013-2017
- TAILORED-Treatment consortium meetings (oral presentations)	2013-2017

Teaching	**Year**
- Supervision of 2nd year medical students "VO Infectieziekten"	2013-2017
- Lecturer in the summer course of the research master "Infection and Immunity"	2015-2016
- Lecturer in the MolMed course "Galaxy for NGS"	2017