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.