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

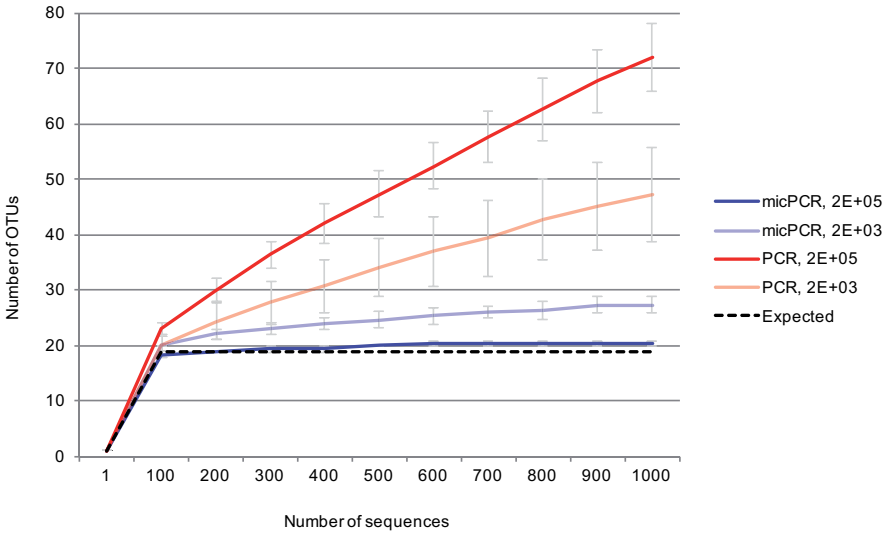

Figure 1. Comparison of rarefaction analyses between micPCR/NGS and traditional PCR/NGS using an equimolar, synthetic microbial community. The number of observed OTUs in the synthetic microbial community is shown as the function of the number of sequences obtained using micPCR/NGS reactions containing 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.