# Multi-locus and ultra-long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer

Alfonso Benítez-Páez*, Yolanda Sanz

Microbial Ecology, Nutrition & Health Research Unit. Institute of Agrochemistry and Food Technology Institute (IATA-CSIC),Valencia, Spain.

*Corresponding author

C. Catedràtic Agustín Escardino Benlloch, 7.

46980 Paterna-Valencia

Spain

Tel: +34 963 900 022 ext. 2129.

Email: abenitez@iata.csic.es.

## Abstract

**Background:** The miniaturised and portable DNA sequencer MinION™ has demonstrated great potential in different analyses such as genome-wide sequencing, pathogen outbreak detection and surveillance, human genome variability, and microbial diversity. In this study, we tested the ability of the MinION™ platform to perform ultra-long amplicon sequencing in order to design new approaches to study microbial diversity using a multi-locus approach.

**Results:** Using R9 chemistry, we generated more than 17,000 reads in a single sequencing run of 40h. After compiling a robust database by parsing and extracting the *rrn* bacterial region from more than 67,000 complete or draft bacterial genomes, we were able to demonstrate that the data obtained during sequencing of the ultra-long amplicon in the MinION™ device was sufficient to study two mock microbial communities in a multiplex manner and to almost completely reconstruct the microbial diversity contained in the HM782D and D6305 mock communities.

**Conclusions:** Although nanopore-based sequencing produces reads with lower per-base accuracy compared with other platforms, we presented a novel approach consisting of multi-locus and ultra-long multiplex amplicon sequencing using the MinION™ MkIb DNA sequencer and R9 chemistry that helps to overcome the main disadvantage of this portable sequencing platform. Despite the technical issues impairing generation of 2D reads with more per-base accuracy, when processed with adequate alignment filters, the 1D reads obtained this way were sufficient to characterize the microbial species present in each mock community analysed. Improvements in nanopore chemistry, such as minimising base-calling errors and new library protocols able to produce rapid 1D libraries, will provide

more reliable information. Such data will be useful for more comprehensive and faster specific detection of microbial species and strains in complex ecosystems.

**Keywords:** MinION; Nanopore sequencer; Ribosomal operon; Ultra-long amplicon sequencing; Microbial diversity; Long-read sequencing

## Background

During the last two years, DNA sequencing based on single-molecule technology has completely changed the perception of genomics for scientists working in a wide range of scientific fields. This new perspective is not only supported by the technology itself but also by the affordability of these sequencing instruments. In fact, unprecedentedly, Oxford Nanopore Technologies (ONT) released the first miniaturised and portable DNA sequencer in early 2014, within the framework of the MinION™ Access Programme. Recently, the MARC consortium (MinION Analysis and Reference Consortium) has published results related to the study of the reproducibility and global performance of the MinION™ platform. These results indicate that this platform is susceptible of a large stochastic variation, essentially derived from the wet-lab and MinION™ operative methods, but also that variability has minimal impact on data quality [1].

The coordinated and collaborative work and mutual feedback between industry and the scientific community have enabled ONT to develop rapidly towards improving its portable platform for DNA sequencing, minimizing the stochastic variation during DNA library preparation. Consequently, in late Autumn 2015, ONT released MkIb, the latest version of MinION™, and in April 2016the fast mode chemistry (R9) was released, increasing the rate of sensing DNA strands from 30-70 to 280-500 bp/sec and reaching up to 95% of per-base accuracy in 2D reads (Clive G. Brown, CTO ONT, personal communication).

One of the most attractive capabilities of the MinION™ platform is the sequencing and assembly of complete bacterial genomes using exclusively nanopore reads [2] or through

hybrid approaches [3, 4]. Notwithstanding, the MinION™ platform has also been demonstrated useful in other relevant areas including: human genetic variant discovery [5, 6], detection of human pathogens [7, 8], detection of antibiotic resistance [9, 10], and microbial diversity [11, 12]. Regarding the latter, microbial diversity and taxonomic approaches are common and in high demand to analyse the microbiota associated to a wide variety of environment- and human-derived samples. However, these analyses are greatly limited by the short-read strategies commonly employed . Thanks to improvements in the chemistry of the most common, popular sequencing platforms in recent years, it is now possible to characterise microbial communities in detail, down to the family or even genus level, using genetic information derived from roughly 30% (~500nt) of the full 16S rRNA gene. Despite the massive coverage achieved with short-read methods, the limitation in terms of read length means taxonomic assignment at the species level is still unfeasible. For instance, taxonomy strategies based on short-reads from Illumina MiSeq platform offer a limited information that underestimates the microbial diversity of complex samples when compared with alternative approaches based on long DNA reads [13]. Consequently, implementation of long-read sequencing approaches to study larger fragments of marker genes will permit the design of new studies to provide evidence for the central role of precise bacterial species/strains in a great variety of microbial consortia. Recent studies at this regard have showed important advances in taxonomy analysis using long reads generated by single molecule technologies [11, 14, 15], indicating that the expansion or inclusion of more hypervariable regions in the analysis overcomes the disadvantage of working with error-prone DNA reads. With respect to the above, we have recently explored the performance of the MinION™ device. Our study demonstrates that data obtained from sequencing nearly full-length 16S rRNA gene amplicons is feasible to study microbial

communities through nanopore technology [11]. We wanted to move a step forward in this type of strategy, thus gaining more specificity when including several hypervariable markers in the analysis, at sequence and structural level, by designing a multi-locus and ultra-long amplicon sequencing method to study microbial diversity. At the same time, we also wanted to explore the affordability of the MinION™ technology to perform microbial diversity analyses by multiplexing several samples in one single MinION™ flowcell.. Accordingly, here we present a study of the 16S, 23S, and the internal transcribed spacer (ITS, that frequently encodes tRNA genes) simultaneous sequencing, using the MinION™ MkIb device and R9 chemistry, with prior generation of ~4.5kb DNA fragments by amplifying the nearly full-length operon encoding the two larger ribosomal RNA genes in bacteria, the *rrn* region (*rrn* hereinafter). We have studied the *rrn* of two mock microbial communities, composed of genomic DNA from 20 and 8 different bacterial species, obtained respectively from BEI Resources and ZYMO Research Corp., using the MinION™ sequencing platform in multiplex configuration.

## Data description

Raw data collected in this experiment were obtained as fast5 files using MinKNOW™ v0.51.3.40 (Oxford Nanopore Technologies) after conversion of electric signals into base calls via the Metrichor™ agent v2.40.17 and the Barcoding plus 2D Basecalling RNN for SQK-NSK007 workflow v1.107. Base-called data passing quality control and filtering were downloaded and data was converted to fasta format using the *poRe* package [16]. Fast5 raw data can be accessed at the European Nucleotide Archive (ENA) under the project ID PRJEB15264. Only one data set was generated after a sequencing run of MinION™ MkIb.

## Analysis

The complete or partial gene sequence of the RNA attached to the small subunit of the ribosome is classically used to perform taxonomy and diversity analysis in complex samples containing hundreds of microbial species. In the case of bacterial species, the 16S rRNA gene is the most widely used DNA marker for taxonomic identification of a particular species, given the relatively high number of hypervariable regions (V1 to V9) present across its sequence. Nowadays, it is possible to study the complete or almost full-length sequence of the 16S rRNA molecule thanks to single-molecule sequencing approaches [11, 14, 15, 17]. The identification of complex microbial communities at species-level with raw data obtained from MinION™ or PacBIO platforms is improving; however, uncertainty in taxonomic assignation is still noteworthy given the high proportion of errors in their reads. While future technical advances may improve the quality of DNA reads generated by third generation sequencing devices, new strategies can also be adopted to enhance the performance of these approaches. Consequently, we postulate that a good example of this is to study a common multi-locus region of the bacterial genome, which enables the simultaneous study of more variable regions and locus arrangements (sequence and structural variation), such as the operon encoding the ribosomal RNA. Using a complex sample where hundreds of microbial species are potentially present (DNA from human faeces) we carried out preliminary experiments to amplify the *rrn*. We observed that from the hypothetical configurations envisaged for the *rrn* (Figure 1A), we only obtained a clear amplification using the primer pairs S-D-Bact-0008-c-S-20 and 23S-2241R, indicating that the *rrn* preferentially seems to be transcriptionally arranged as follows: 16S-ITS-23S. A

detailed evaluation of the fragment size determined that main PCR products ranged from 4.3 to 5.4kbp (Figure 1B-D), being consistent with the expected size of PCR products amplifying the 16S, ITS, and 23S regions from several microbial species. The next step involved designing a multiplex sequencing approach to try to analyse more than one sample per sequencing run in one flowcell of MinION™; therefore, the primers were re-designed to include a distinctive barcode region at 5´ (Table 1). During PCR of the *rrn* we tagged the amplicon derived from the mock community HM782D with the barcode *bc01* in a dual manner, whereas the amplicons derived from sample D6305 were tagged with barcode *bc08* in similar way. Parallel experiments were conducted on HM782D and D6305 DNA, with comparative aims, using a conventional protocol of microbial diversity analysis and consisting of the V4-V5 16S amplicon sequencing by Illumina MiSeq paired-end approach (see methods).

One of the major handicaps when proposing this new *rrn* region to be used for taxonomy analysis is the need to compile a reference database to compare the reads produced by MinION™ device. Therefore, we proceeded to parse the genetic information of over 67,000 bacterial genomes whose sequences are publicly available in GenBank at NCBI. In this way, we retrieve and compile more than 47,000 *rrn* sequences that were subject of a clustering analysis to reduce the level of redundancy and to disclose the variability intrinsically associated to the *rrn* itself and to its individual components as well (Figure 2). We found that *rrn* region comprising the 16S, ITS, and 23S coding regions exhibits 2-fold more variation than that observed for the 16S molecule alone. Strikingly, the ITS regions showed the highest level of variability when over 3-fold more clusters were produced at 97, 98 and 99% sequence identity. When parsing the genetic information of over 67,000

bacterial genomes, we observed the ITS region frequently encodes one or several tRNA genes. We hypothesized this greater level of variation in the bacterial ITS is directly caused by the variety of tRNA genes encoded in this region, which would even enable different operons from the same bacterial chromosome to be distinguished, therefore, structural variation was additionally found at the ITS and in the *rrn* region by extension. Consequently, the variability observed in the *rrn* was the largest observed and thought to be meaningful for the aims of this study. We found that the smaller size of the ITS together with the high level of sequence and specially of structural variability explain the larger number of clusters recovered at 97%, 98% and 99% when compared with counterparts.

Once we could compile a reference database for comparison aims, we proceeded with the amplicon library construction and sequencing run obtaining raw data consisting of 17,038 reads and almost all were classified as 1D reads. For general knowledge, the DNA reads derived from the MinION™ device can be classified into three types: '1D template', '1D complement', and '2D' reads. The latter, 2D reads, are products of aligning and merging sequences from the template (read from leader adapter) and complement reads (a second adapter called hairpin or HP adapter must be generated), produced from the same DNA fragment. These contain a lower error rate, owing to strand comparison and mismatch correction. In addition to the technical issues indicative of a bad ligation of the HP adapter, we obtained 93% of reads (~15,900 reads) during the first 16h of run; thus, we obtained lower sequencing performance after re-loading with the second aliquot of the sequencing library and extended the run for another 24h (40h in sum). The fasta sequences were filtered by retaining those between 1500 and 7000 nt in length, obtaining at least enough sequence information to compare a DNA sequence equivalent to the 16S rRNA gene

length. After this filtering step, we retained 72% of sequences (12,278) and then we performed the respective barcode splitting. For this purpose, we modified the default parameters of the "split_barcodes.pl" perl script (Oxford Nanopore Technologies) by incorporating the information of the extended barcodes (Table 1), rather than the barcode information alone, and simultaneously increased the stringency parameter to 25 (14 by default). Afterwards the concatenation of reads were obtained from respective forward and reverse extended barcodes, then we retrieved a total of 2,019 (52% from forward and 48% from reverse barcodes) and 1,519 (53% from forward and 47% from reverse barcodes) 1D reads for HM782D and D6305 mock communities, respectively. Read-mapping was performed against the *rrn* database, compiling more than 22,000 *rrn* regions, retrieved from more than 67,000 genomes available in GenBank (see Availability of supporting data). The taxonomy associated to the best hit based on the alignment score followed by filtering steps (see methods) was used to determine the structure of each mock community. The MinION™ sequencing data produced the microbial structure presented in Figure 3 for the mock communities HM782D and D6305, respectively.

Figure 3 shows the bacterial species and their respective relative proportions retrieved from the analysis of the mock communities HM782D and D6305, respectively. With respect to the HM782D mock community, we were able to recover 20 representative species, accounting for 16 out of 20 species present in that artificial community (Figure 3A). However, the remaining four species that apparently are absent in this community have a close relationship to others detected correctly, namely *Bacillus subtilis, Bacillus thuringensis, Bacillus anthracis,* and *Propionibacterium sp.* Furthermore, we were unable to report the presence of just four species present in HM782D because proportions of

*Rhodobacter sphaeroides* and *Actinomyces odontolyticus* were below the predominance threshold (1%), being present in 0.25 and 0.12%, respectively. Similarly, other 40 different species but close to that present in the HM782D mock community (*Bacillus* spp., *Streptococcus* spp. *Clostriudium* spp., *Neisseria* spp., *Staphylococcus* spp, and *Listeria* spp.) had minor representation in data derived from *rrn* sequencing. With respect to *Rhodobacter sphaeroides* and *Actinomyces odontolyticus* lower proportions, we have previously demonstrated that the low levels of 16S reads are a consequence of amplification bias derived from the PCR reaction and not from sequencing itself [11]. In this case, the new primer pair used to generate the ultra-long amplicons would seem to work more efficiently than those previously used, but apparently they still present issues at bacterial coverage level. When we revised the whole taxonomy contained in our *rrn* database, the compiling of non *rrn* regions for *Deinococcus radiodurans* and *Helicobacter pylori* partially explained the lack of these species in HM782D analysed by the present approach. However, a new alignment process using individual 16S and 23S rRNA sequences obtained from GenBank and including those for *D. radiodurans* and *H. pylori*, respectively, demonstrated that at least *D. radiodurans* could be identified in a higher proportion than *A. odontolyticus* and *R. sphaeroides*, albeit in a lower proportion than our predominance threshold (data not shown). The Illumina MiSeq data obtained after sequencing the V4-V5 16S region permitted to characterize the genus distribution in the HM782D sample with the RDP Classifier. As a result, we compiled distribution of all 17 genus represented in the HM782D mock community (Figure 3C) and 4 additional genus with very low abundance (2 reads / 8,409 assigned). Moreover, when a OTU-picking approach was conducted we recovered 41 OTUs whose identity was evaluated in the SINA server (Supplementary Material 1). Globally, we recovered taxonomy identification of all

genus expected but only three species were well identified based on the Greengenes taxonomy (*H. pylori*, *P. acnes*, and *R. sphaeroides*) whereas one was wrongly identified (*Neisseria cinerea*).

Regarding the results obtained from the D6305 mock community, we found a total of 10 bacterial species present in this mixed DNA sample, eight of them matched the expected structure of the community, and additionally 18 close species had minor representation (*Bacillus* spp., *Enterococcus* spp., *Klebsiella* spp., *Lactobacillus* spp., *Streptococcus* spp., and *Staphylococcus* spp.). Similarly, we could recover the eight different component at genus level of this mock community by using Illumina MiSeq data (Figure 3C) plus seven additional and not related genus with very low abundance (< 7 reads / 8,046 assigned). At OTUs level, we retrieved a total of 14 sequences whose taxonomy identification is presented in the Supplementary Material 2. In this case, only *S. enterica* could be identified at species level. Using the MinION™ data we were able to recover 100% of the species present in the sample and the two additional members identified also have a close relationship within the *Bacillus* genus, as observed in the HM782D sample (Figure 3B). We have determined that coverage needed to retrieve all expected species in a non-even mock community with an abundance above 1% is ~13X in terms of the number of species of that community.

When compared to reference values and proportions theoretically expected for the species present in the two mock communities, we observed some deviations that were greater in certain species. Particularly, in the HM782D sample the lowest coverage biases were observed for *Actinomyces odontolyticus* (-5.36), *Rhodobacter sphaeroides* (-4.36), and

*Enterococcus faecalis* (-2.04). This indicates that such species, in addition to *D. radiodurans* and *H. pylori*, are more difficult to detect with the primers and PCR used here. By contrast, *Escherichia coli* (1.79) seems to be preferentially amplified, given that this species exhibited the highest positive coverage bias value (Figure 3C). The values obtained for D6305 were more homogeneous, and the lowest coverage bias was observed for *Lactobacillus fermentum* (-2.18) (Figure 3D). Additional analysis indicated that there was not significant correlation between coverage bias and GC content in *rrn*. Although the low coverage bias for some species can be solved by selecting another pair of primers, the ability to recover almost all of them, at least in a low proportion, in itself represents an important attribute of this approach for inter-sample comparisons. Interestingly, we observed a similar pattern of overrepresentation of *Bacillus* spp. sequences (>50%) in D6305 sample but not for *Escherichia* spp. sequences (~4%) in the HM782D mock community when Illumina MiSeq data was assessed (Figure 3C-D).

The high error rate of the 1D reads (ranging between 70 and 87% sequence identity, according to high quality alignments) makes barcoding de-multiplexing a difficult task in nanopore data. However, our results indicate that with the configuration and parameters presented here we could efficiently distinguish the reads generated from HM782D and D6305 amplicons. As a consequence, the performance of this ultra-long amplicon approach to properly assign microbial communities to samples was efficiently assisted by the parameters during the de-multiplexing process that were central to discern reads obtained from respective samples multiplexed in the MinION flowcell. For instance, the distribution of reads matching with close related species such as *Lactobacillus gasseri* and *Lactobacillus fermentum*, contained distinctively in HM782D and D6305 samples, was

indicative of the adequate execution of the de-multiplexing pipeline. The above was also exemplified for *Salmonella enterica* sequences that were determined only in D6305 despite its close relationship with *E. coli* at the 16S and 23S sequence level (close to 100%). Regarding the latter, the multiple sequence alignment built with *rrn* regions from both species was inspected directly distinguishing the ITS as the major source of variation between the two species (data not shown). Indeed, this was corroborated by the comparative analysis performed during the clustering step of the reference samples to create our *rrn* database.

## Discussion

The inventory of microbial species based on 16S rDNA sequencing is frequently used in biomedical research to determine microbial organisms inhabiting the human body and their relationship with disease. Recently, third-generation of DNA sequencing platforms have developed rapidly, facilitating the identification of microbial species and overcoming the read-length issues inherent to second-generation sequencing methods. These advances allow researchers to infer taxonomy and analyse diversity from the almost full-length bacterial 16S rRNA sequence [11, 14, 15, 17]. Particularly, the ONT platform deserves special attention given its portability and its fast development since the MinION™ became available in 2014. Notwithstanding, this technology is susceptible to a large stochastic variation, essentially derived from the wet-lab methods [1]. We corroborated this issue by obtaining a sequencing run where the raw data predominantly consisted of 1D reads as a consequence of the HP adapter ligation failure, despite following the manufacturer's instructions. However, we were able to develop an efficient analysis protocol where the

higher read quality offered by R9 chemistry and the updated Metrichor basecaller protocol proved pivotal to obtain 1D reads with a range of identity between 70 and 86%, with sufficient per-base accuracy to successfully perform the taxonomic analyses described herein.

Our preliminary results indicated that the *rrn* region in bacteria preferentially has a unique conformation (with the transcriptional arrangement of 16S-ITS-23S) and we could amplify this ~4.5Kbp region with the selected S-D-Bact-0008-c-S-20 and 23S-2241R primer pair. Once we were able to distinguish the feasibility to amplify the *rrn*, our approach comprised the study of two different mock communities in a multiplex manner, to be combined in one single MinION™ flowcell. By designing the respective forward and reverse primers tagged with specific barcodes recommended by ONT, we were able to retrieve extended barcode-associated reads, in spite of the large proportion of per-base errors contained in these types of reads. Using MinION™ data based on multilocus markers and ultra-long amplicon sequencing, we could reconstruct the structure of two commercially available mock communities. Although the expected proportions of some species in each community exhibited an important coverage bias, we were able to recover 80% (HM782D) and 100% (D6305) of bacterial species from the respective mock communities. Consequently, future analyses should be conducted to find an appropriate PCR approach using primers with a higher coverage for bacterial species.

We have analysed a great amount of genetic information with the aim of compiling a valuable database containing the genetic information for the *rrn* present in over 67,000 draft and complete bacterial genomes. The global length distributions in the region

indicated that the *rrn* was 4,993 ± 187 bp in length whereas the 16S, ITS, and 23S sub-regions were 1,612 ± 75, 488 ± 186, and 3,036 ± 160 bp in length, respectively. Using this genetic information of the *rrn* and clustered at 100% of sequence identity enabled us to establish a multi-locus marker able to discriminate the taxonomy of two mock communities containing very close species. The latter was possible given that simultaneous analysis of the 16S, ITS, and 23S molecules offered 2-fold more diversity that studying the 16S, ITS, or 23S sequences separately. Moreover, the ITS was distinguished individually as a highly variable genetic region in terms of sequence and length. Furthermore, it contributes notably to the higher variability observed in the *rrn* region, a fact evidenced in previous studies [18-21]. The accumulation of a larger number of variable sites in the *rrn* region, together with the particular structural variation of the ITS to potentially accommodate and encode tRNA genes, are thought to be central to discriminating bacterial species, despite the large proportion of per-base errors contained in MinION™ reads. Our data indicate that our MinION reads had a length average of 2,655 and 2,918 bases for HM782D and D6305, respectively. Consequently, the taxonomy assignment was predominantly based on the variability of at least two out of the three markers included in the *rrn*, no matter if reads were produced from the 16S or 23S edges of *rrn* amplicons. We expect this type of analysis will likely become more accurate over time as nanopore chemistry improves in near future, with the concomitant increase in throughput, which is pivotal to disclose the hundreds of species present in complex microbial communities for analysis in human or environmental studies. Therefore, the multi-locus, ultra-long and multiplex methods described here represent a promising analysis routine for microbial and pathogen identification, relying on the sequence variation accumulated in approximately 5kbp of DNA, roughly accounting for the assessment of 1.25% of an average bacterial genome (~4Mbp). Notwithstanding, we

cannot obviate that the current state of this approach presents some limitations in terms of the completeness of the *rrn* database created as well as the efficiency of the primers used to generate the ultra-long amplicons. At date, our database include *rrn* sequences from 2,479 different species grouped into 918 different genus. In consequence, urgent studies must be undertaken to generate a more complete database including the *rrn* genomic information from species inhibiting complex and real samples such as those derived from human body.

## Methods

*Bacterial DNA and rrn amplicons*

The complex DNA sample for preliminary studies of *rrn* region arrangement consisted of DNA isolated from faeces, kindly donated by a healthy volunteer upon informed consent. An aliquot of 200 mg of human faeces was used to isolate microbial DNA using the QIAamp DNA Stool Mini Kit (Qiagen) and following the manufacturer's instructions. Finally, DNA was eluted in 100 µL nuclease-free water and a DNA aliquot at 20 ng/µL was prepared for PCR reaction using the primer pairs S-D-Bact-0008-c-S-20 and 23S-2241R or 23S-129F and S-D-Bact-1391-a-A-17 for testing configurations shown in Figure 1A (Table 1). The band size was analysed using the Java-based GelAnalyzer tool (www.gelanalyzer.com). Genomic DNA for the reference mock microbial communities was kindly donated by BEI Resources (http://www.beiresources.org) and ZYMO Research Corp (http://www.zymoresearch.com). The composition of the mock communities was as follows: i) HM782D is a genomic DNA mixture of 20 bacterial species containing equimolar ribosomal RNA operon counts (100,000 copies per organism per µL), as indicated by the manufacturer; and ii) ZymoBIOMICS Cat No. D6305 (D6305 hereinafter)

is a genomic DNA mixture of eight bacterial species (and two fungal species) presented in equimolar amounts of DNA. According to manufacturers' instructions, 1 µL of DNA from each mock community was used to amplify all the genes contained in the *rrn*. DNA was amplified in triplicate by 27 PCR cycles at 95ºC for 30 s, 49ºC for 15 s, and 72ºC for 210 s. Phusion High-Fidelity Taq Polymerase (Thermo Scientific) and the primers S-D-Bact-0008-c-S-20 (mapping on 5' of 16S gene) and 23S-2241R (mapping on 3' of 23S gene), which target a wide range of bacterial 16S rRNA genes [22, 23]. For the Illumina MiSeq sequencing the V4-V5 hypervariable regions from bacterial 16S rRNA gene were amplified using 1 µL of DNA from each mock community and 25 PCR cycles at 95ºC for 20 s, 40ºC for 30 s, and 72ºC for 20 s. Phusion High-Fidelity Taq Polymerase (Thermo Scientific) and the 6-mer barcoded primers S-D-Bact-0563-a-S-15 (AYTGGGYDTAAAGNG) and S-D-Bact-0907-a-A-20 (CCGTCAATTYMTTTRAGTTT). As we wished to multiplex the sequencing of both mock communities into one single MinION™ flowcell, we designed a dual-barcode approach where respective primers were synthesized and fused with two different barcodes recommended by ONT (Table 1). Amplicons consisted of ~4.5kbp blunt-end fragments for MinION approach and ~380bp for Illumina MiSeq approach, and those were purified using the Illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare). Amplicon DNA was quantified using a Qubit 3.0 fluorometer (Life Technologies).

*Amplicon DNA library preparation*

The Genomic DNA Sequencing Kit SQK-MAP006 was ordered from ONT and used to prepare the amplicon library for loading into the MinION™. Approximately 0.9 µg of

amplicon DNA (0.3 per mock community plus 0.3 μg of an extra query sample consisting amplicons obtained from a genomic DNA mix of several uncharacterized microbial isolates) were processed for end repair using the NEBNextUltra II End Repair/dA-tailing Module (New England Biolabs), and followed by purification using Agencourt AMPure XP beads (Beckman Coulter) and washing twice with 1 volume of fresh 70% ethanol. Subsequently, and according to the manufacturer's suggestions, we used 0.2 pmol of the purified amplicon DNA (~594 ng, assuming fragments of 4.5kbp in length) to perform the adapter ligation step. Ten μL of adapter mix, 2 μL of HP adapter, and 50 μl of Blunt/TA ligase master mix (New England Biolabs) were added in that order to the 38 μl end-repaired amplicon DNA. The reaction was incubated at room temperature for 15 minutes, 1 μL HP Tether was added and incubated for an additional 10 minutes at room temperature. The adapter-ligated amplicon was recovered using MyOne C1-beads (Life Technologies) and rinsed with washing buffer provided with the Genomic DNA Sequencing Kit SQK-MAP006 (Oxford Nanopore Technologies). Finally, the sample was eluted from the MyOne C1-beads by adding 25 μL of elution buffer and incubating for 10 minutes at 37°C before pelleting in a magnetic rack. Samples for Illumina MiSeq approach were sent to the National Center for Genomic Analaysis (CNAG, Spain) for multiplex sequencing in one lane of MiSeq instrument with 2x300 paired-end configuration.

*Flowcell set-up*

A brand new, sealed R9 flowcell was acquired from ONT and stored at 4°C before use. The flowcell was fitted to the MinION™ MkIb prior to loading the sequencing mix, ensuring good thermal contact. The R9 flowcell was primed twice using 71 μL premixed nuclease-

free water, 75 µL 2x running buffer, and 4 µL fuel mix. At least 10 minutes were required to equilibrate the flowcell before each round of priming and before final DNA library loading.

*Amplicon DNA sequencing*

The sequencing mix was prepared with 59 µl nuclease-free water, 75 µl 2x running buffer, 12 µL DNA library, and 4 µL fuel mix. A standard 48-hour sequencing protocol was initiated using the MinKNOW™ v0.51.3.40. Base-calling was performed through data transference using the Metrichor™ agent v2.40.17 and the Barcoding plus 2D Basecalling RNN for SQK-NSK007 workflow v1.107. During the sequencing run, one additional freshly diluted aliquot of DNA library was loaded after 16 hours of initial input. The raw sequencing data derived from the two mock communities studied here was expected to account two-thirds of the data produced by the R9 flowcell used.

*The rrn database*

We built a database containing the genetic information for the 16S and 23S rRNA genes and the ITS sequence in all the complete and draft bacterial genomes available in the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria). A total of 67,199 genomes were analysed by downloading the "*fna*" files and parsing for rRNA genes into the respective "*gff*" annotation file. Chromosome coordinates for *rrn* regions were parsed and used to extract such a DNA sequences from complete chromosomes or DNA contigs assembled. The resulting *rrn* sequences were analysed and the length distribution was assessed. We retrieved a total of 47,698 *rrn* sequences with an average of 4,993 nt in

length. By selecting the size distribution equal to the 99th percentile (two-sided), we discarded potential incomplete or aberrant annotated *rrn* sequences and observed that *rrn* sequences can be found between 4,196 and 5,790nt; under these boundaries, our *rrn* database finally accounted for a total of 46,920 sequences. Equivalent databases were built by parsing the respective *rrn* sequences with the RNammer tool to discriminate the 16S, ITS, and 23S rRNA sequences [24]. To remove the level of redundancy of our *rrn* database and to maintain the potential discriminatory power at strain level, we performed clustering analysis using USEARCH v8 tool for sequence analysis and the option *-otu_radius_pct* equal 0 [25], thus obtaining a total of 22,350 reference sequences. For comparative aims, the *rrn* database and the 16S, ITS, and 23S databases were also analysed using the option *-otu_radius_pct* with values ranging from 1 to 3. For accessing to *rrn* database and the respective species annotation, see Availability of supporting data.

*MinION data analysis*

Read-mapping was performed using the LAST aligner v.189 [26] with parameters -q1 -b1 -Q0 -a1 -r1. Each 1D read was compared in a competitive way against the entire *rrn* database and the best hit was selected by obtaining the highest alignment score. Alignment length as well as alignment coordinates in target and query sequences were parsed from the LAST output and the sequence identity between matched regions was calculated using the python *Levenshtein* distance package. An iterative processing was used to determine thresholds for detection by evaluating the taxonomy distribution with reads subsampling and different levels of sequence identity in top scored alignments. High quality alignments were selected by filtering out those with identity values up to the 50th percentile of the distribution of identity values of all reads per sample (~69%). Therefore, taxonomy

assignment was based exclusively on alignments with ≥ 70% identity. Basic stats, distributions, filtering, and comparisons were performed in R v3.2.0 (https://cran.r-project.org). For relative quantification of species the singletons were removed and the microbial species considered to be predominantly present in the mock communities were those with a relative a proportion ≥ 1%, a value that demonstrated to be discriminative to always obtain the expected microbial diversity during the iterative processing of alignments. The coverage bias was calculated by obtaining fold-change ($Log_2$) of species-specific read counting against the expected (theoretical) average for the entire community according to information provided by the manufacturers.

*Illumina data analysis*

Fastq paired-end raw data was were assembled using *Flash* software [27]. The HM782D and D6305 reads were de-multiplexed and barcode and primers were removed using *Mothur* v1.36.1 [28]. The sequences were then processed for chimera removal using *Uchime* algorithm [29] and SILVA reference set of 16S sequences [30]. A normalized subset of 10,000 sequences per sample was created by random selection after shuffling (10,000X) of the original dataset. Taxonomy assessment was performed using the RDP classifier v2.7 [31]. The Operational Taxonomic Unit (OTU)-picking approach was performed with the normalized subset of 10,000 sequences and the *uclust* algorithm implemented in USEARCH v8.0.1623 and the options *-otu_radius_pct* equal 3 for clustering at 97% and *-min*size 2 for remove singletons [25]. SINA server was used for taxonomy identification of OTUs recovered from Illumina MiSeq data [32].

## Availability of supporting data

Accessions for the *rrn* database containing the reference sequences for alignments and taxonomic annotation is available at https://github.com/alfbenpa/rrn_db. The code source of the original *split_barcodes.pl* perl script is available at https://github.com/nanoporetech/barcoding/releases/tag/1.0.0 with ONT copyright.

## Abbreviations

EC , European Commission; ENA, European Nucleotide Archive; HDF, Hierarchical Data Format; ITS, internal transcribed spacer; NCBI, National Center for Biotechnology Information; ONT, Oxford Nanopore Technologies; PCR, Polymerase Chain Reaction; rDNA, DNA encoding for the Ribosomal RNA; rRNA, Ribosomal RNA;*rrn*, the DNA region containing the 16S and 23S bacterial rRNA genes and its respective ITS region; SINA, SILVA Incremental Aligner; USB, Universal Serial Bus.

## Competing interests

ABP is part of the MinION™ Access Programme (MAP).

## Authors' contributions

ABP and YS designed the study and managed the project. ABP performed the experiments, analysed and managed the data. ABP draft the manuscript. Both authors read and approved the final manuscript.

## Acknowledgements

This work and the contract to ABP is supported by the European Union's Seventh Framework Program under the grant agreement nº 613979 (MyNewGut).

## References


1. Ip CL, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, Leggett RM, Eccles DA, Zalunin V, Urban JM *et al*: **MinION Analysis and Reference Consortium: Phase 1 data release and analysis**. *F1000Res* 2015, **4**:1075.
2. Loman NJ, Quick J, Simpson JT: **A complete bacterial genome assembled de novo using only nanopore sequencing data**. *Nat Methods* 2015, **12**(8):733-735.
3. Karlsson E, Larkeryd A, Sjodin A, Forsman M, Stenberg P: **Scaffolding of a bacterial genome using MinION nanopore sequencing**. *Sci Rep* 2015, **5**:11996.
4. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, Watson M: **A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data**. *Gigascience* 2015, **4**:60.
5. Ammar R, Paton TA, Torti D, Shlien A, Bader GD: **Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes**. *F1000Res* 2015, **4**:17.
6. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W: **Nanopore sequencing detects structural variants in cancer**. *Cancer Biol Ther* 2016:1-8.
7. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM *et al*: **Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis**. *Genome Med* 2015, **7**(1):99.
8. Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR, Rosenzweig CN, Minot SS: **Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer**. *Gigascience* 2015, **4**:12.
9. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J: **MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island**. *Nat Biotechnol* 2015, **33**(3):296-300.
10. Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ: **Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes**. *J Antimicrob Chemother* 2015, **70**(10):2775-2778.
11. 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.
12. Li C, Cheng KR, Boey JHE, Ng HQA, Wilm A, Nagarajan N: **INC-Seq: Accurate single molecule reads using nanopore sequencing**. *bioRxiv* 2016:doi: http://dx.doi.org/10.1101/038042
13. Myer PR, Kim M, Freetly HC, Smith TP: **Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers**. *J Microbiol Methods* 2016, **127**:132-140.
14. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK: **Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system**. *PeerJ* 2016, **4**:e1869.

15. Shin J, Lee S, Go MJ, Lee SY, Kim SC, Lee CH, Cho BK: **Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing**. *Sci Rep* 2016, **6**:29681.
16. Watson M, Thomson M, Risse J, Talbot R, Santoyo-Lopez J, Gharbi K, Blaxter M: **poRe: an R package for the visualization and analysis of nanopore sequencing data**. *Bioinformatics* 2014, **31**(1):114-115.
17. Li C, Chng KR, Boey EJ, Ng AH, Wilm A, Nagarajan N: **INC-Seq: accurate single molecule reads using nanopore sequencing**. *Gigascience* 2016, **5**(1):34.
18. Fernandez J, Avendano-Herrera R: **Analysis of 16S-23S rRNA gene internal transcribed spacer of Vibrio anguillarum and Vibrio ordalii strains isolated from fish**. *FEMS Microbiol Lett* 2009, **299**(2):184-192.
19. Maslunka C, Gurtler V, Seviour R: **Unusual features of the sequences of copies of the 16S-23S rRNA internal transcribed spacer regions of Acinetobacter bereziniae, Acinetobacter guillouiae and Acinetobacter baylyi arise from horizontal gene transfer events**. *Microbiology* 2015, **161**(Pt 2):322-329.
20. Stewart FJ, Cavanaugh CM: **Intragenomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria**. *J Mol Evol* 2007, **65**(1):44-67.
21. Tambong JT, Xu R, Bromfield ES: **Intercistronic heterogeneity of the 16S-23S rRNA spacer region among Pseudomonas strains isolated from subterranean seeds of hog peanut (Amphicarpa bracteata)**. *Microbiology* 2009, **155**(Pt 8):2630-2640.
22. Hunt DE, Klepac-Ceraj V, Acinas SG, Gautier C, Bertilsson S, Polz MF: **Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity**. *Appl Environ Microbiol* 2006, **72**(3):2221-2225.
23. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO: **Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies**. *Nucleic Acids Res* 2012, **41**(1):e1.
24. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic Acids Res* 2007, **35**(9):3100-3108.
25. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**(19):2460-2461.
26. Frith MC, Hamada M, Horton P: **Parameters for accurate genome alignment**. *BMC Bioinformatics* 2010, **11**:80.
27. Magoc T, Salzberg SL: **FLASH: fast length adjustment of short reads to improve genome assemblies**. *Bioinformatics* 2011, **27**(21):2957-2963.
28. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ *et al*: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities**. *Appl Environ Microbiol* 2009, **75**(23):7537-7541.
29. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R: **UCHIME improves sensitivity and speed of chimera detection**. *Bioinformatics* 2011, **27**(16):2194-2200.

30. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools**. *Nucleic Acids Res* 2013, **41**(Database issue):D590-596.
31. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Appl Environ Microbiol* 2007, **73**(16):5261-5267.
32. Pruesse E, Peplies J, Glockner FO: **SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes**. *Bioinformatics* 2012, **28**(14):1823-1829.

## Figure legends

**Figure 1.**Organization of the *rrn* region in bacteria. A - hypothetical transcriptional arrangements expected for *rrn* and tested experimentally using two sets of primer pairs (see small arrows drawn in each configuration). B - Agarose gel electrophoresis of PCR reactions performed under the two hypothetical arrangements of *rrn*; lanes: 1) 1kb ruler (Fermentas), 2) PCR reaction from the top configuration in panel A, 3) PCR reaction from the bottom configuration in panel A. The GelAnalyser Java application was used to perform the band size analysis of the 1kb ruler standard (C) and the amplicons obtained from human faecal DNA (D).

**Figure 2.**Variability of the *rrn* region and its functional domains. The *rrn* database compiled after parsing more than 67,000 draft and complete bacterial genomes was assessed by clustering analysis at different levels of sequence identity: 97 (white bars), 98 (light grey bars), 99 (dark grey bars), and 100% (black bars). For comparative aims, the functional DNA sequences encoded into the *rrn* region were also individually studied. The fold numbers located at the top of respective ITS, 23S and *rrn* bars indicate the proportion of clusters found and normalized to those found in the 16S.

**Figure 3.**Microbial structure of the mock communities. A and B - microbial species and respective relative proportions determined to be present in the HM782D and D6305 mock communities, respectively, following the analysis of raw data obtained from *rrn* amplicon sequencing in the MinION$^{TM}$. C and D - Comparative analysis of the expected microbial species and proportions against the data obtained after mapping of reads generated by a 4.5kbp amplicon PCR and sequenced in MinION$^{TM}$ device, for HM782D and D6305 respectively. Data obtained from Illumina MiSeq sequencing of V4-V5 16S region from respective mock communities is also included. Given that taxonomy identification of MiSeq reads at species level only retrieved very few assignments, we arbitrary accommodated the distribution at genus level into the expected species. For those species belonging to the same genus (HM782D mock community), we distributed equally the genus abundance into the number of species.

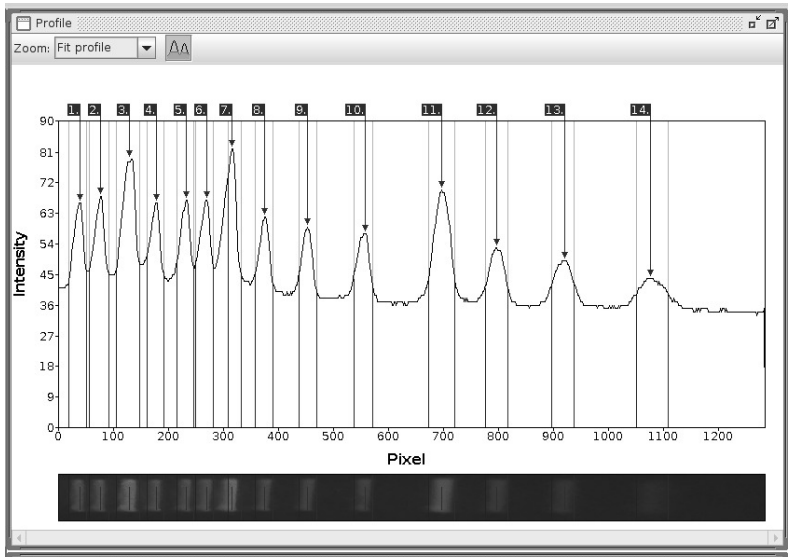
**Supplementary Material 1.** MS Excel file compiling the output information retrieved from SINA server (https://www.arb-silva.de/aligner/) for taxonomy identification of 41 OTUs derived from analysis of HM782D with Illumina MiSeq approach. Information regarding sequence quality, identity percentage, mapping coordinates against *E. coli* reference, length and taxonomy based on SILVA, Greengenes, and RDP databases is showed for all OTU.

**Supplementary Material 2.** MS Excel file compiling the output information retrieved from SINA server (https://www.arb-silva.de/aligner/) for taxonomy identification of 18 OTUs derived from analysis of D6305 with Illumina MiSeq approach. Information regarding sequence quality, identity percentage, mapping coordinates against *E. coli* reference, length and taxonomy based on SILVA, Greengenes, and RDP databases is showed for all OTU.

A

B

C

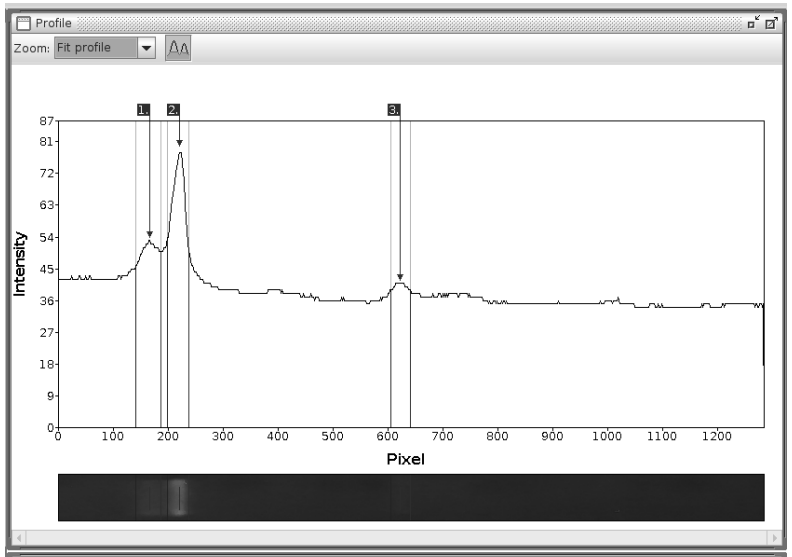



Analysis info

Show

| Lane # | Band # | Rf | Raw volume | Cal. volume | MW |
|---|---|---|---|---|---|
| 1. | 1. | 0.03 | 1895 | - | **10000** |
| 1. | 2. | 0.06 | 2041 | - | **8000** |
| 1. | 3. | 0.1 | 2802 | - | **6000** |
| 1. | 4. | 0.139 | 1775 | - | **5000** |
| 1. | 5. | 0.181 | 1794 | - | **4000** |
| 1. | 6. | 0.209 | 1906 | - | **3500** |
| 1. | 7. | 0.246 | 1739 | - | **3000** |
| 1. | 8. | 0.292 | 1768 | - | **2500** |
| 1. | 9. | 0.352 | 1698 | - | **2000** |
| 1. | 10. | 0.433 | 1803 | - | **1500** |
| 1. | 11. | 0.542 | 2840 | - | **1000** |
| 1. | 12. | 0.619 | 2030 | - | **750** |
| 1. | 13. | 0.716 | 1913 | - | **500** |
| 1. | 14. | 0.837 | 2489 | - | **250** |

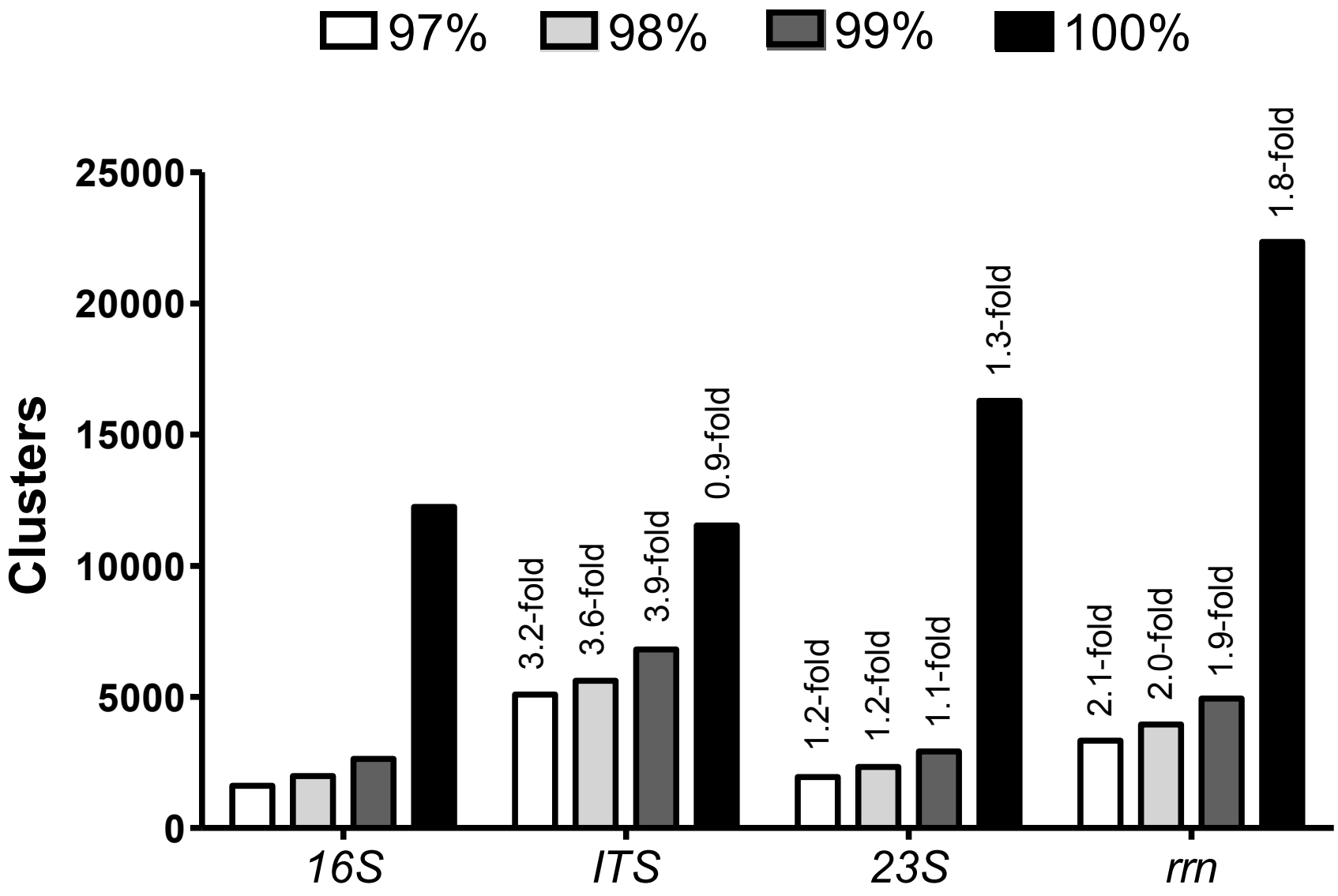
D




Analysis info

Show

| Lane # | Band # | Rf | Raw volume | Cal. volume | MW |
|---|---|---|---|---|---|
| 2. | 1. | 0.129 | 2315 | - | 5441 |
| 2. | 2. | 0.171 | 2602 | - | 4304 |
| 2. | 3. | 0.484 | 1488 | - | 1061 |


97%
98%
99%
100%
Clusters
25000
20000
15000
10000
5000
0
3.2-fold
3.6-fold
3.9-fold
0.9-fold
1.2-fold
1.2-fold
1.1-fold
1.3-fold
2.1-fold
2.0-fold
1.9-fold
1.8-fold
16S
ITS
23S
rrn

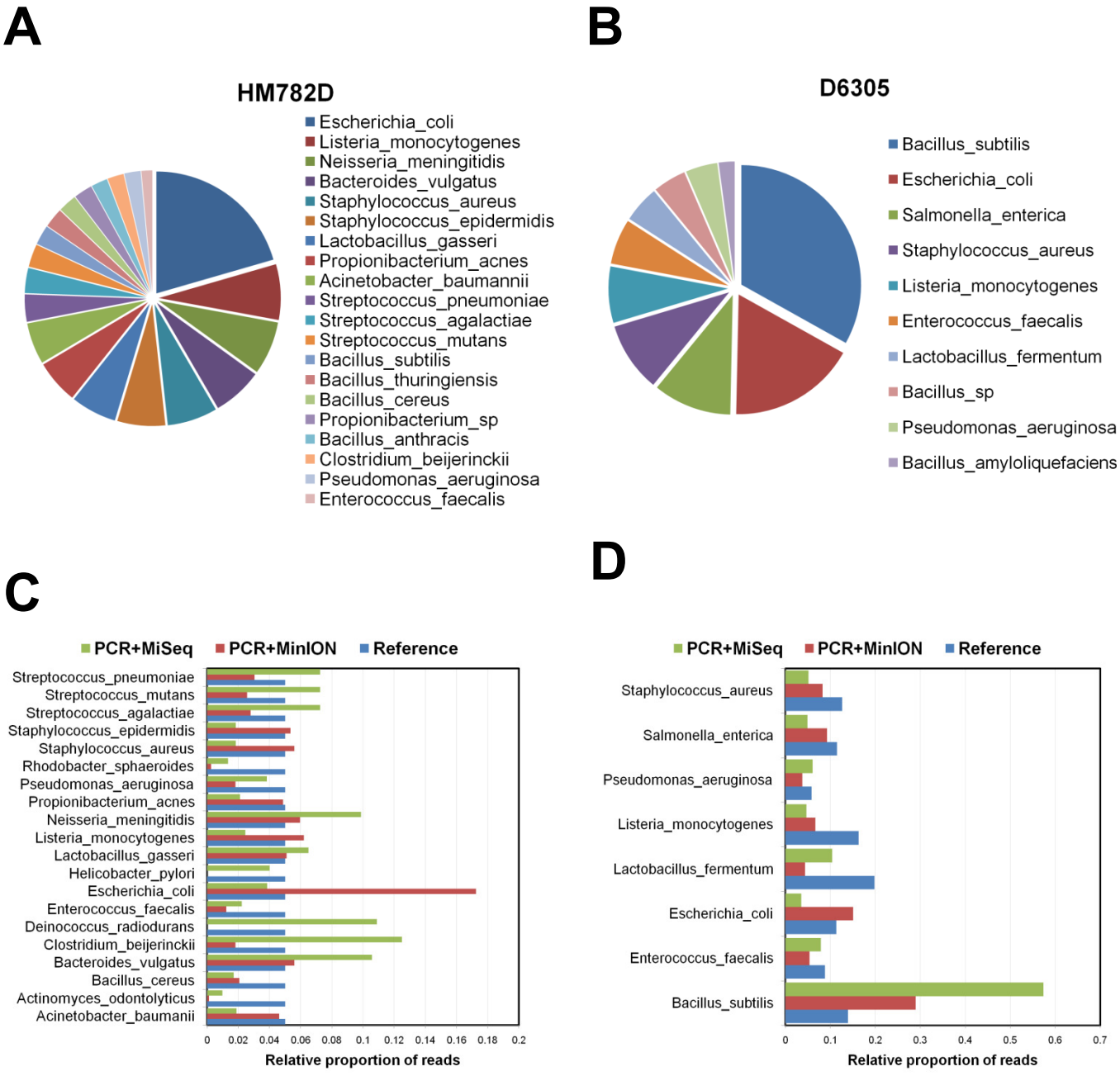

A
HM782D
Escherichia_coli
Listeria_monocytogenes
Neisseria_meningitidis
Bacteroides_vulgatus
Staphylococcus_aureus
Staphylococcus_epidermidis
Lactobacillus_gasseri
Propionibacterium_acnes
Acinetobacter_baumannii
Streptococcus_pneumoniae
Streptococcus_agalactiae
Streptococcus_mutans
Bacillus_subtilis
Bacillus_thuringiensis
Bacillus_cereus
Propionibacterium_sp
Bacillus_anthracis
Clostridium_beijerinckii
Pseudomonas_aeruginosa
Enterococcus_faecalis
B
D6305
Bacillus_subtilis
Escherichia_coli
Salmonella_enterica
Staphylococcus_aureus
Listeria_monocytogenes
Enterococcus_faecalis
Lactobacillus_fermentum
Bacillus_sp
Pseudomonas_aeruginosa
Bacillus_amyloliquefaciens
C
PCR+MiSeq
PCR+MinION
Reference
Streptococcus_pneumoniae
Streptococcus_mutans
Streptococcus_agalactiae
Staphylococcus_epidermidis
Staphylococcus_aureus
Rhodobacter_sphaeroides
Pseudomonas_aeruginosa
Propionibacterium_acnes
Neisseria_meningitidis
Listeria_monocytogenes
Lactobacillus_gasseri
Helicobacter_pylori
Escherichia_coli
Enterococcus_faecalis
Deinococcus_radiodurans
Clostridium_beijerinckii
Bacteroides_vulgatus
Bacillus_cereus
Actinomyces_odontolyticus
Acinetobacter_baumanii
0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2
Relative proportion of reads
D
PCR+MiSeq
PCR+MinION
Reference
Staphylococcus_aureus
Salmonella_enterica
Pseudomonas_aeruginosa
Listeria_monocytogenes
Lactobacillus_fermentum
Escherichia_coli
Enterococcus_faecalis
Bacillus_subtilis
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
Relative proportion of reads

Table 1. Barcodes and primers used to generate amplicon libraries.

| Sample | Barcode | Primer | Barcode extended[1] |
|---|---|---|---|
| **HM-782D** | (bc01)<br>GGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCT | (S-D-Bact-0008-c-S-20)<br>AGAGTTTGATCMTGGCTCAG | (bc01F)<br>GGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTAGAGTTTGATCMTGGCTCAG |
| | (bc01)<br>GGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCT | (23S-2241R)<br>ACCGCCCCAGTHAAACT | (bc01R)<br>GGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTACCGCCCCAGTHAAACT |
| **D6503** | (bc08)<br>GGTGCTGTTCAGGGAACAAACCAAGTTACGTTTAACCT | (S-D-Bact-0008-c-S-20)<br>AGAGTTTGATCMTGGCTCAG | (bc08F)<br>GGTGCTGTTCAGGGAACAAACCAAGTTACGTTTAACCTAGAGTTTGATCMTGGCTCAG |
| | (bc08)<br>GGTGCTGTTCAGGGAACAAACCAAGTTACGTTTAACCT | (23S-2241R)<br>ACCGCCCCAGTHAAACT | (bc08R)<br>GGTGCTGTTCAGGGAACAAACCAAGTTACGTTTAACCTACCGCCCCAGTHAAACT |
| **Other primers used** | | | |
| **Human fecal DNA** | | (S-D-Bact-1391-a-A-17)<br>GACGGGCGGTGWGTRCA | |
| | | (23S-129F)<br>CYGAATGGGRVAACC | |

1 Underlined sequences correspond with the barcode sequence