RESEARCH ARTICLE

REVISED Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the *rrn* operon [version 2; peer review: 2 approved, 3 approved with reservations]

Anna Cuscó [1], Carlotta Catozzi [2,3], Joaquim Viñes[1,3], Armand Sanchez[3], Olga Francino[3]

[1]Vetgenomics, SL, Bellaterra (Cerdanyola del Vallès), Barcelona, 08193, Spain
[2]Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, Milano, Italy
[3]Molecular Genetics Veterinary Service (SVGM), Universitat Autonoma of Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, 08193, Spain

v2 **First published:** 06 Nov 2018, **7**:1755 (https://doi.org/10.12688/f1000research.16817.1)
Latest published: 01 Aug 2019, **7**:1755 (https://doi.org/10.12688/f1000research.16817.2)

Abstract

Background: Profiling the microbiome of low-biomass samples is challenging for metagenomics since these samples are prone to contain DNA from other sources (e.g. host or environment). The usual approach is sequencing short regions of the 16S rRNA gene, which fails to assign taxonomy to genus and species level. To achieve an increased taxonomic resolution, we aim to develop long-amplicon PCR-based approaches using Nanopore sequencing. We assessed two different genetic markers: the full-length 16S rRNA (~1,500 bp) and the 16S-ITS-23S region from the *rrn* operon (4,300 bp).
Methods: We sequenced a clinical isolate of *Staphylococcus pseudintermedius*, two mock communities and two pools of low-biomass samples (dog skin). Nanopore sequencing was performed on MinION™ using the 1D PCR barcoding kit. Sequences were pre-processed, and data were analyzed using EPI2ME or Minimap2 with *rrn* database. Consensus sequences of the 16S-ITS-23S genetic marker were obtained using canu.
Results: The full-length 16S rRNA and the 16S-ITS-23S region of the *rrn* operon were used to retrieve the microbiota composition of the samples at the genus and species level. For the *Staphylococcus pseudintermedius* isolate, the amplicons were assigned to the correct bacterial species in ~98% of the cases with the16S-ITS-23S genetic marker, and in ~68%, with the 16S rRNA gene when using EPI2ME. Using mock communities, we found that the full-length 16S rRNA gene represented better the abundances of a microbial community; whereas, 16S-ITS-23S obtained better resolution at the species level. Finally, we characterized low-biomass skin microbiota samples and detected species with an environmental origin.
Conclusions: Both full-length 16S rRNA and the 16S-ITS-23S of the *rrn* operon retrieved the microbiota composition of simple and complex microbial communities, even from the low-biomass samples such as dog

Open Peer Review

Reviewer Status ✔ ? ✔ ? ?

	Invited Reviewers				
	1	**2**	**3**	**4**	**5**
REVISED **version 2** published 01 Aug 2019					? report
version 1 published 06 Nov 2018	✔ report	? report	✔ report	? report	? report

1 **Alfonso Benítez-Páez**, Institute of Agrochemistry and Food Technology-Spanish National Research Council (IATA-CSIC), Valencia, Spain

2 **Rasmus H. Kirkegaard**, Aalborg University (AAU), Aalborg, Denmark

3 **Amanda Warr**, University of Edinburgh, Edinburgh, UK

4 **Kon Chu**, Seoul National University, Seoul, South Korea

5 **Lee J. Kerkhof**, Rutgers University, New Brunswick, USA

skin. For an increased resolution at the species level, targeting the 16S-ITS-23S of the *rrn* operon would be the best choice.

Keywords
microbiome, microbiota, 16S, rrn operon, nanopore, canine, low-biomass, skin, dog

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Nanopore Analysis gateway.

Corresponding author: Anna Cuscó (anna.cusco@vetgenomics.com)

Author roles: Cuscó A: Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Catozzi C**: Investigation, Validation, Visualization, Writing – Review & Editing; **Viñes J**: Investigation, Visualization, Writing – Review & Editing; **Sanchez A**: Conceptualization, Funding Acquisition; **Francino O**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by two grants awarded by Generalitat de Catalunya (Industrial Doctorate program, 2013 DI 011 and 2017 DI 037).

How to cite this article: Cuscó A, Catozzi C, Viñes J *et al.* **Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the *rrn* operon [version 2; peer review: 2 approved, 3 approved with reservations]** F1000Research 2019, **7**:1755 (https://doi.org/10.12688/f1000research.16817.2)

First published: 06 Nov 2018, **7**:1755 (https://doi.org/10.12688/f1000research.16817.1)

REVISED **Amendments from Version 1**

We hereby present a revised version of our manuscript, based on the comments made by the referees and more results since version 1 was published.

The main changes of the manuscript when compared to version 1 are those stated here:

- We have assessed the performance of the mapping approach with the uncorrected long-amplicons strategy presented here.
- We have expanded the results adding a section of the *de novo* assembly of the 16S-ITS-23S genetic marker.
- We have provided more detail on the methodology used from the lab bench to the bioinformatics part, with supplementary file 1 providing the commands used for the mapping approach.
- The mapping results have been updated since in version 1 chimeras were computed differently for each amplicon and led to confusing results (more chimera for 16S rRNA than for 16S-ITS-23S). Now for both amplicons, we have performed base-level alignment using Minimap2 as the first step before yacrd (detailed in Supplementary File 1).
- The figures and tables have been simplified, re-ordered and completed to be more informative and clear. We got rid of the graphic bars and kept the heatmaps.

Finally, we want to thank all the referees for their time to evaluate our work and their invaluable feedback.

See referee reports

Introduction

The microbiota profile of low-biomass samples such as skin is challenging for metagenomics. These samples are prone to contain DNA contamination from the host or exogenous sources, which can overcome the DNA of interest[1,2]. Thus, the usual approach is amplifying and sequencing certain genetic markers that are ubiquitously found within the studied kingdom rather than performing metagenomics. Ribosomal marker genes are a common choice: 16S rRNA and 23S rRNA genes to taxonomically classify bacteria[3,4]; and ITS1 and ITS2 regions for fungi[5,6].

Until now, most studies of microbiota rely on massive parallel sequencing, and target a short fragment of the 16S rRNA gene, which presents nine hypervariable regions (V1-V9) that are used to infer taxonomy[7,8]. The most common choices for host-associated microbiota are V4 or V1-V2 regions, which present different taxonomic coverage and resolution depending on the taxa[9,10].

Apart from the biases derived from the primer choice, short fragment strategies usually fail to assign taxonomy reliably at the genus and species level. This taxonomic resolution is particularly useful when associating microbiota to clinics such as in characterizing disease status or when developing microbiota-based products, such as pre- or pro-biotics[11]. For example, in human atopic dermatitis (AD) the signature for AD-prone skin when compared to healthy skin was enriched for *Streptococcus* and *Gemella*, but depleted in *Dermacoccus*. Moreover, nine different bacterial species were identified to have significant AD-associated microbiome differences[12]. In canine atopic dermatitis, *Staphylococcus pseudintermedius* has been classically associated with the disease. Microbiota studies of canine atopic dermatitis presented an overrepresentation of *Staphylococcus* genus[13,14], but the species was only confirmed when complementing the studies using directed qPCRs for the species of interest[13] or using a *Staphylococcus*-specific database and V1-V3 region amplification[14].

With the launching of single-molecule technology sequencers (e.g. PacBio or Oxford Nanopore Technologies), these short-length associated issues can be overcome by sequencing the full-length 16S rRNA gene (~1,500 bp) or even the nearly-complete *rrn* operon (~4,300 bp), which includes the 16S rRNA gene, ITS region, and 23S rRNA gene.

Several studies assessing the full-length 16S rRNA gene have already been performed using Nanopore sequencing to: i) characterize artificial bacterial communities (mock community)[15–17]; ii) complex microbiota samples, from the mouse gut[18], wastewater[19], microalgae[20] and dog skin[21]; and iii) the pathogenic agent in a clinical sample[22–24]. Some studies have been performed using the nearly-complete *rrn* operon to characterize mock communities[25] and complex natural communities[26].

Here, we aim to assess these two long-amplicon approaches using MinION™ (Oxford Nanopore Technologies), a single-molecule sequencer that is portable, affordable with a small budget and offers long-read output. Its main limitation is a higher error rate than massive sequencing. We will test our approaches by sequencing several samples with different degrees of complexity: i) a clinical isolate of *Staphylococcus pseudintermedius*, ii) two bacterial mock communities; and iii) two complex skin microbiota samples.

Methods

Samples and DNA extraction

We first sequenced a pure bacterial isolate of *S. pseudintermedius* obtained from the ear of a dog affected by otitis.

Then, we used two DNA mock communities as simple and well-defined microbiota samples:

- HM-783D, kindly donated by BEI resources, containing genomic DNA from 20 bacterial strains with staggered ribosomal RNA operon counts (between 10^3 and 10^6 copies per organism per µl).
- ZymoBIOMICS™ Microbial Community DNA standard that contained a mixture of genomic DNA extracted from pure cultures of eight bacterial strains.

As a complex microbial community, we used two DNA sample pools from the skin microbiota of healthy dogs targeting two different skin sites: i) dorsal back (DNA from two dorsal samples from Beagle dogs); and ii) chin (DNA from five chin samples from Golden Retriever/Labrador crossed dogs). Skin microbiota samples were collected using Sterile Catch-All™ Sample Collection Swabs (Epicentre Biotechnologies) soaked in sterile SCF-1 solution (50 mM Tris buffer (pH 8), 1 mM

EDTA, and 0.5% Tween-20). DNA was extracted from the swabs using the PowerSoil™ DNA isolation kit (MO BIO) and blank samples were processed simultaneously (for further details on sample collection and DNA extraction see 27).

PCR amplification of ribosomal markers

We evaluated two ribosomal markers in this study: the full-length 16S rRNA gene (~1,500 bp) and the 16S-ITS-23S region of the ribosomal operon (*rrn)* (~4,300 bp). Before sequencing, bacterial DNA was amplified using a nested PCR, with a first PCR to add the specific primer sets tagged with the Oxford Nanopore universal tag and a second PCR to add the barcodes from the PCR barcoding kit (EXP-PBC001) (Supplementary Table 1). Each PCR reaction included a no-template control sample to assess possible reagent contamination.

For the first PCR, we targeted: i) the full-length 16S rRNA gene using 16S-27F[28] and 16S-1492R[29] primer set and ii) the 16S-ITS-23S of the *rrn* operon using 16S-27F and 23S-2241R[28] primer set (Supplementary Table 1). All the three primers contained the Oxford Nanopore tag, which is an overhang that allows barcoding the samples during the second PCR.

PCR mixture for the full-length 16S rRNA gene (25 µl total volume) contained 5 ng of DNA template (or 2.5 µl of unquantifiable initial DNA), 1X Phusion® High Fidelity Buffer, 0.2 mM of dNTPs, 0.4 µM of 16S-27F, 0.8 µM of 16S-1492R and 0.5 U of Phusion® Hot Start II Taq Polymerase (Thermo Scientific, Vilnius, Lithuania). The PCR thermal profile consisted of an initial denaturation of 30 s at 98°C, followed by 25 cycles of 15 s at 98°C, 15 s at 51°C, 45 s at 72°C, and a final step of 7 min at 72°C.

PCR mixture for the 16S-ITS-23S of the *rrn* operon (50 µl total volume) contained 5 ng of DNA template (or 2.5 µl of unquantifiable initial DNA), 1X Phusion® High Fidelity Buffer, 0.2 mM µl dNTPs 1 µM each primer and 1 U Phusion® Hot Start II Taq Polymerase. The PCR thermal profile consisted of an initial denaturation of 30 s at 98°C, followed by 25 cycles of 7 s at 98°C, 30 s at 59°C, 150 s at 72°C, and a final step of 10 min at 72°C.

The amplicons were cleaned-up with the AMPure XP beads (Beckman Coulter) using a 0.5X and 0.45X ratio for the 16S rRNA gene and the 16-ITS-23S of the *rrn* operon, respectively. Then, they were quantified using Qubit™ fluorometer (Life Technologies, Carlsbad, CA) and the volume was adjusted to begin the second round of PCR with 0.5 nM of the first PCR product or the complete volume when not reaching the required DNA mass (mostly in the samples that amplified with the 16S-ITS-23S genetic marker).

PCR mixture for the barcoding PCR (100 µl total volume) contained 0.5 nM of the first PCR product (50 ng for the 16S rRNA gene and 142 ng for the 16S-ITS-23S), 1X Phusion® High Fidelity Buffer, 0.2 mM µl dNTPs, and 2 U Phusion® Hot Start II Taq Polymerase. Each PCR tube contained the DNA, the PCR mixture and 2 µl of the specific barcode. The PCR thermal profile consisted of an initial denaturation of 30 s at 98°C, followed by 15 cycles of 7 s at 98°C, 15 s at 62°C, 45 s (for the 16S rRNA gene) or 150 s (for *rrn* operon) at 72°C, and a final step of 10 min at 72°C.

Again, the amplicons were cleaned-up with the AMPure XP beads (Beckman Coulter) using a 0.5X and 0.45X ratio for the 16S rRNA gene and the whole *rrn* operon, respectively. For each sample, quality and quantity were assessed using Nanodrop and Qubit™ fluorometer (Life Technologies, Carlsbad, CA), respectively. The samples with higher DNA concentrations were checked by agarose gel to see the size profile of the PCR products (Supplementary Figure 1).

The different barcoded samples were pooled in equimolar ratio to obtain a final pool (1,000–1,500 ng in 45 µl) to do the sequencing library. In few cases, 16S-ITS-23S amplicons did not reach the initial amount of required DNA and we proceeded with lower input material.

Nanopore sequencing library preparation

The Ligation Sequencing Kit 1D (SQK-LSK108; Oxford Nanopore Technologies) was used to prepare the amplicon library to load into the MinION™ (Oxford Nanopore Technologies), following the manufacturer's protocol. Input DNA samples were composed of 1–1.5 µg of the barcoded DNA pool in a volume of 45 µl and 5 µl of DNA CS (DNA from lambda phage, used as a positive control in the sequencing). The DNA was processed for end repair and dA-tailing using the NEBNext End Repair/dA-tailing Module (New England Biolabs). A purification step using 1X Agencourt AMPure XP beads (Beckman Coulter) was performed.

For the adapter ligation step, a total of 0.2 pmol of the end-prepped DNA were added in a mix containing 50 µl of Blunt/TA ligase master mix (New England Biolabs) and 20 µl of adapter mix and then incubated at room temperature for 10 min. We performed a purification step using Adapter Bead Binding buffer (provided in the SQK-LSK108 kit) and 0.5X Agencourt AMPure XP beads (Beckman Coulter) to finally obtain the DNA library.

We prepared the pre-sequencing mix (14 µl of DNA library) to be loaded by mixing it with Library Loading beads (25.5 µl) and Running Buffer with fuel mix (35.5 µl). We used two SpotON Flow Cells Mk I (R9.4.1) (FLO-MIN106). After the quality control, we primed the flowcell with a mixture of Running Buffer with fuel mix (RBF from SQK-LSK108) and Nuclease-free water (575 µl + 625 µl). Immediately after priming, the nanopore sequencing library was loaded in a dropwise fashion using the SpotON port.

Once the library was loaded, we initiated a standard 48 h sequencing protocol using the MinKNOW™ software v1.15.

Data analysis workflow

The samples were run using the MinKNOW software. After the run, fast5 files were base-called and de-multiplexed using Albacore v2.3.1. A second de-multiplexing round was performed

with Porechop v0.2.3[30], where only the barcodes that agreed with Albacore were kept. Porechop was also used to trim the barcodes and the adapters from the sequences, as well as 45 extra base pairs from each end that correspond to the length of the universal tags and custom primers (See Supplementary Figure 2 for a schematic overview of the process and Supplementary File 1 for the bioinformatics workflow of the mapping approach).

After the trimming, reads were selected by size: 1,200 bp to 1,800 bp for 16S rRNA gene; and 3,500 to 5,000 bp for the 16S-ITS-23S of the *rrn* operon. Afterwards, we removed chimeras with the following approach: i) we mapped each mock community to its mock database and the complex samples to the complete *rrn* database using Minimap2 v2.16 (with base-level alignment and z-score set to 70)[31]; ii) chimeras were detected and removed using yacrd v0.5[32].

To assign taxonomy to the trimmed and filtered reads we used to strategies: 1) a mapping-based strategy using Minimap2 v2.16[31] (with base-level alignment and z-score set to 70); or 2) a taxonomic classifier using What's in my Pot (WIMP)[33], a workflow from EPI2ME in the Oxford Nanopore Technologies cloud (based on Centrifuge software[34]).

For the mapping-based strategy, we performed Minimap2 again with the non-chimeric sequences. We applied extra filtering steps to retain the final results: we kept only those reads that aligned to the reference with a block equal or larger than 1,000 bp (for 16S rRNA gene) and 3,000 bp (for the 16S-ITS-23S of the *rrn* operon). For reads that hit two or more references, only the alignments with the highest Smith-Waterman alignment score (AS score) were kept.

The reference databases used in this study were:

- Mock DB: a collection of the complete genomes that were included in each mock community, as described by the manufacturer. The HM-783D database was retrieved from NCBI using the reference accession numbers, while Zymobiomics mock community has already its database online on the Amazon AWS server.

- *rrn* DB: sequences from the whole ribosomal operon from 22,351 different bacterial species retrieved from Genbank by Benitez-Paez *et al.*[25]. We have manually added a sequence of the *rrn* operon from *S. pseudintermedius*.

For assessing the mapping-based strategy, we have made a subset with the *rrn* DB to exclude all the operons that were representatives of Gammaproteobacteria class as an example to see how the alignment-based approach performs when missing main references within the database. These operons were identified by introducing a list of all the genera of the *rrn* DB as a batch in NCBI Taxonomy browser. The sequences belonging to Gammaproteobacteria (code 1236) were removed from the *rrn* DB.

For the taxonomic classification using the WIMP workflow, which uses the NCBI database, only those hits with a classification score >300 were kept[34].

Ampvis2 package in R was used to plot the heatmaps[35] and the Phyloseq package, to plot the alpha rarefaction curves[36].

An earlier version of this article can be found on bioRxiv (doi: https://doi.org/10.1101/450734)

Results

We have assessed the performance of the full-length 16S rRNA and the 16S-ITS-23S rRNA region of the ribosomal (*rrn*) operon to profile the microbial composition of several samples: a bacterial isolate, two mock communities and two complex skin samples (chin and dorsal back).

The samples amplified using the full-length 16S rRNA gene recovered a higher percentage of reads after the quality control when compared to 16S-ITS-23S of the *rrn* operon: 73–95% vs. 30–79%. For the 16S-ITS-23S of the *rrn* operon, the largest percentage of reads was lost during the length trimming step since some of the reads presented lengths that were shorter than expected (Supplementary Table 2).

Bacterial isolate analysis

We first sequenced an isolate of *S. pseudintermedius* obtained from a canine otitis. When using WIMP approach with the 16S-ITS-23S of the *rrn* operon, 97.5% of the sequences were correctly assigned at the species level as *S. pseudintermedius*. However, with the full-length 16S rRNA gene, 68% of the sequences were correctly assigned at the species level as *S. pseudintermedius,* while 13% at the genus one and ~20% were wrongly assigned, either by not reaching the species level or by giving an incorrect species (Table 1).

When using the mapping approach with the *rrn* DB, we obtained no hit to *S. pseudintermedius*. Instead, they were hitting mostly to *Staphylococcus schleiferi*, which is a closely related species; there were also few hits to *Staphylococcus hyicus* and *Staphylococcus agnetis*. This result was due to the *rrn* DB did not contain any representative of *S. pseudintermedius*. When including *S. pseudintermedius* sequence to the *rrn* DB (*rrn* + *S. pseudintermedius*) both markers retrieved the correct result with more than 97% of the assignments hitting the correct reference.

When comparing the alignment results obtained with Minimap2 and *rrn* DB vs *rrn* DB + *S. pseudinteremedius*, we found that the Smith-Waterman alignment score (AS) presented higher values in the correct alignments, especially when using 16S-ITS-23S marker gene (Figure 1). Thus, the AS score could

Table 1. Taxonomy assignments of *S. pseudintermedius* isolate. Taxonomic assignments were obtained i) using WIMP workflow with NCBI RefSeq database; ii) Minimap2 with *rrn* DB; and iii) Minimap2 with *rrn* DB including *S. pseudintermedius*.

	WIMP (NCBI RefSeq DB)		Minimap2 (*rrn* DB)		Minimap2 (*rrn* DB + *S. pseudintermedius*)	
Taxonomy	**16S**	**16S-ITS-23S**	**16S**	**16S-ITS-23S**	**16S**	**16S-ITS-23S**
Staphylococcus pseudintermedius	68.1%	97.6%	-	-	97.9%	97.5%
Staphylococcus sp	13.1%	0.3%	-	-	-	-
*Staphylococcus schleiferi**	2.2%	0.3%	94.9%	82.1%	1.94%	2.2%
*Staphylococcus aureus**	3.2%	0.2%	0.0%	0.0%	0.0%	0.0%
*Staphylococcus lutrae**	2.7%	0.1%	-	-	-	-
*Staphylococcus hyicus**	0.3%	0.1%	3.2%	8.2%	0.0%	0.2%
*Staphylococcus agnetis**	0.2%	0.0%	1.6%	9.7%	0.0%	0.1%
Other *Staphylococcus**	3.7%	1.4%	0.3%	0.0%	0.2%	-
Other species*	6.5%	0.1%	-	-	-	-

*Incorrect bacterial species assignment

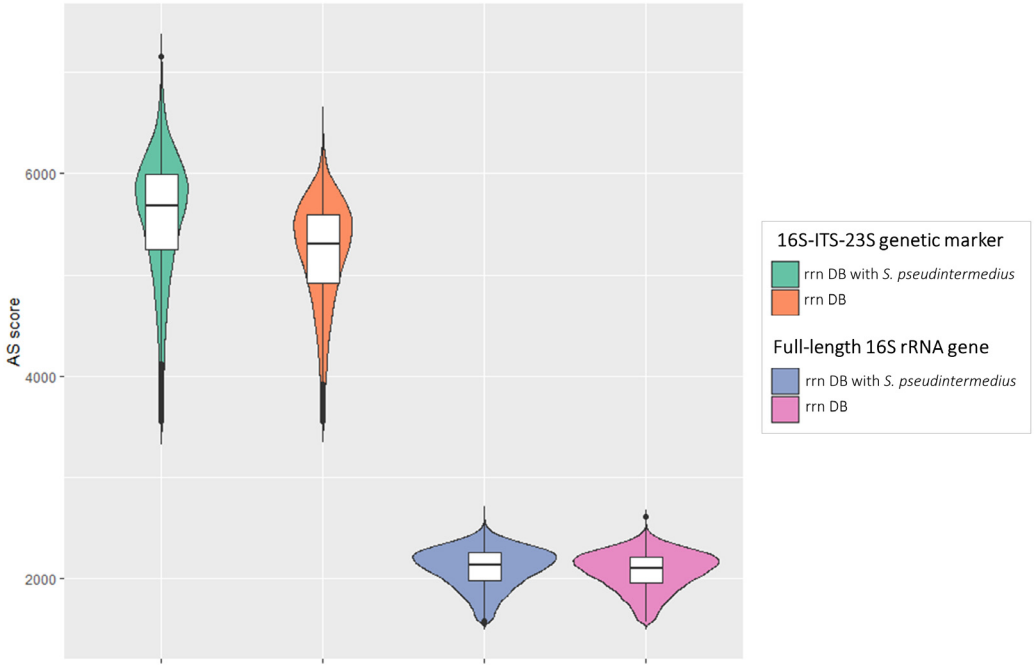

Figure 1. Violin plot representing the distribution of AS score for *S. pseudintermedius* isolate. Alignment scores for each genetic marker were obtained using Minimap2 and either *rrn* DB or *rrn* DB with a reference added for *S. pseudintermedius*. The first two plots are for 16S-ITS-23S, whereas the second ones are for 16S rRNA.

be a filter to identify a wrong taxonomic assignment due to the lack of a reference in the database.

Assessment of the mapping-based strategy

As we have already seen for *S. pseudintermedius*, in the mapping strategy a lack of a reference in the database led to an incorrect taxonomic assignment. Since both marker genes chosen for the microbiota profiling are highly conserved among bacteria, the mapping strategy (through Minimap2) will always align to some reference.

To check the behavior of the mapping approach when using an incomplete database, we have performed an example test using the mock communities. We have mapped the mock communities

both against the complete *rrn* DB and against a subset of the *rrn* DB without any representative of the Gammaproteobacteria class. The *rrn* DB[25] contains 22,351 different bacterial species, including representatives of the species in both mock communities. We have chosen Gammaproteobacteria because each mock community contains three Gammaproteobacteria species, representing around 24% of the total microbial composition.

We checked the alignment score values and the alignment block length to detect any differences on the alignment performance when using complete or an incomplete database. We plotted two histograms: i) read counts distributed by the alignment block length; and ii) read counts distributed by the alignment score (AS) (Figure 2). For the 16S-ITS-23S genetic marker, we detected a clear pattern: when aligning to the *rrn* DB without Gammaproteobacteria, both histograms changed from a left-skewed distribution to a bimodal distribution with two peaks (Figure 2). A new peak appeared at the lower values that included the wrong taxonomic assignments, which are species not present in the mock community or the non-concordant hits when compared to the complete *rrn* DB results. Thus, for the 16S-ITS-23S marker, the initial filtering step by alignment block length will get rid of most of the incorrect taxonomic assignments. However, this pattern was not observed with the full-length 16S rRNA gene (Figure 2) or when closely-related references were present in the database, as seen above for the *S. pseudintermedius* isolate. So, to further confirm taxonomic results (especially for the 16S rRNA gene), we assigned taxonomy using two different bioinformatics approaches that work with different databases.

Mock community analyses

We analyzed two microbial mock communities to validate the ability of the presented approach: i) to quantify what is expected and detect biases of the technique; and ii) to reach a reliable taxonomic assignment at the species level.

For the first aim, we used the HM-783D mock community that contained genomic DNA from 20 bacterial strains with

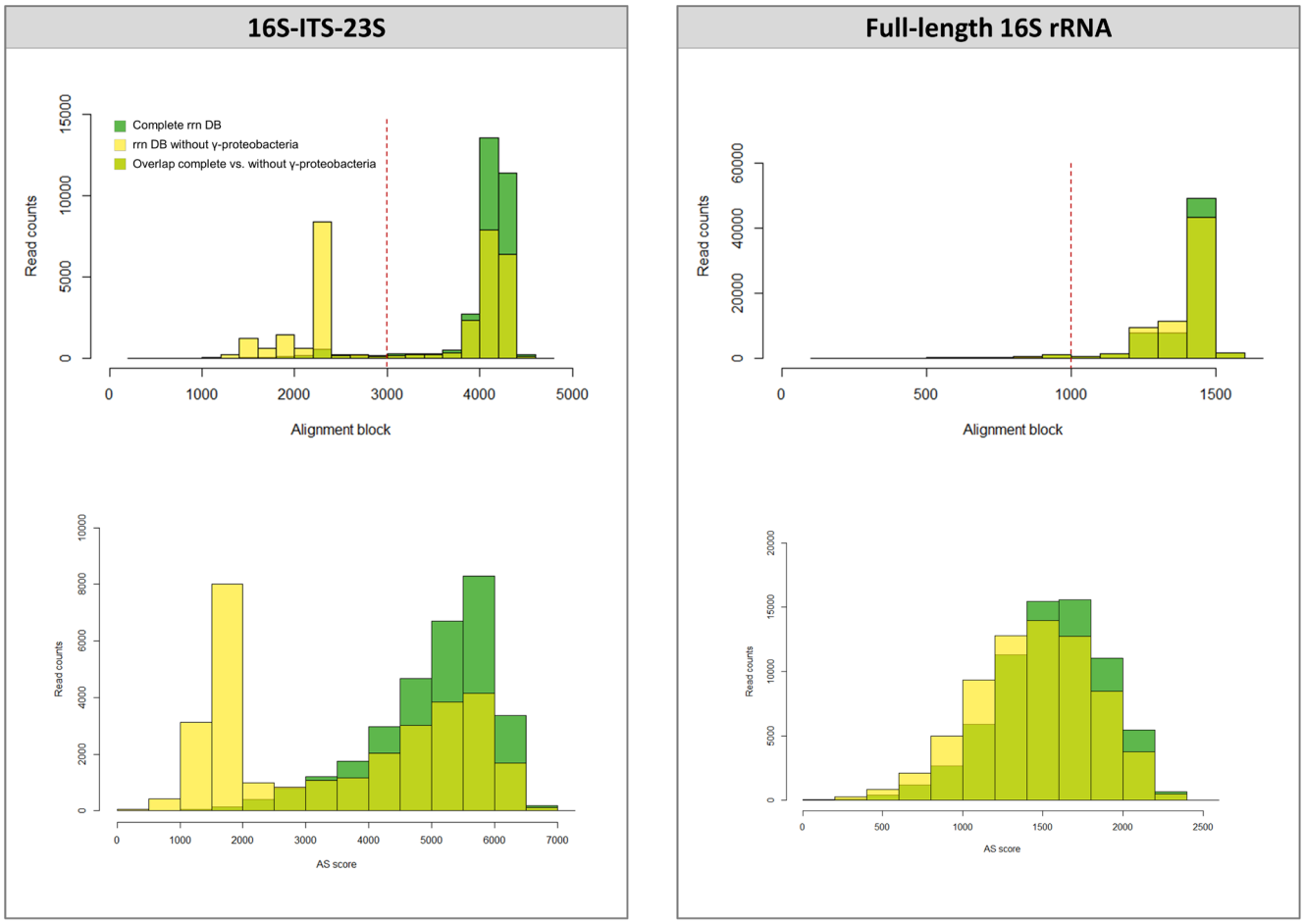

Figure 2. Assessment of the mapping-based strategy using a sample of Zymobiomics mock community. Top part, histograms of the read counts distributed by alignment block depending on the database used (complete *rrn* DB vs *rrn* DB without Gammaproteobacteria). Alignment threshold at 3,000 bp (for 16S-ITS-23S) and 1,000 (for 16S rRNA gene) are marked with a red dashed line. Below part, histograms of the read counts distributed by AS scores depending on the database used (complete *rrn* DB vs *rrn* DB without Gammaproteobacteria).

staggered ribosomal RNA operon counts (from 10^3 to 10^6 copies per organism per µl). This mock community would allow us determining if our approach reliably represents the actual bacterial composition of the community, especially considering the low-abundant species. We assigned taxonomy with Minimap2 and a database containing only the 20 representative bacterial genomes in the HM-783D mock community (Minimap2-mock DB).

On the one hand, using the full-length 16S rRNA gene we detected all the bacterial species present in the mock community, even the low-abundant ones. On the other hand, using the 16S-ITS-23S of the *rrn* operon we detected only the most abundant species (at least 10^4 operon copies) (Figure 3A). This could be due to the lower sequencing depth obtained with 16S-ITS-23S when compared with 16S rRNA (Supplementary Table 2). Moreover, the relative abundances of 16S-ITS-23S sequences were more biased than those obtained from 16S rRNA gene sequencing, which confirmed that the primers for 16S-ITS-23S of the *rrn* operon need to be improved for a better representation of the actual abundances (Figure 3B and 3C).

To assess if the technique and the analyses would give a reliable taxonomy at the species level we used Zymobiomics mock community, which contains equal quantities of 8 bacterial species. The expected 16S rRNA gene content for each representative is also known, so we were able to determine if the different analysis approaches reliably represented the actual bacterial composition of the community. We sequenced the Zymobiomics mock community twice per marker gene and found that the replicates presented equivalent results.

We assigned taxonomy with three different approaches: i) Minimap2 and a database containing only the 8 bacterial species of the correspondent mock community (mock DB); ii) Minimap2 and a database containing sequences for the *rrn* operon of 22,351 different bacterial species (*rrn* DB[25]) and iii) WIMP from EPI2ME and NCBI RefSeq database.

Similarly to what we have seen for the HM-783D mock community, when using the mapping strategy with Minimap2 and the mock database, we detected that the full-length 16S rRNA gene retrieved better the actual abundances of the mock community. The 16S-ITS-23S genetic marker over- and under-represented most of the bacterial species in the mock community. When using larger databases such as *rrn* DB and NCBI RefSeq, both the full-length 16S rRNA gene and the 16S-ITS-23S were able to detect 8 out of 8 bacterial species of the Zymobiomics mock community (Figure 4A). However, we also detected other taxa that included mostly higher taxonomic rank taxa (sequences not assigned to species level), but also not expected taxa (wrongly-assigned species), especially with WIMP and NCBI RefSeq database (see Supplementary Table 3 for complete taxonomic assignments).

On one hand, the mapping approach using the *rrn* DB provided highly similar results to the reference, despite the larger size of this database (Figure 4A), especially with 16S-ITS-23S marker (99% of the reads were correctly assigned at the species level with 16S-ITS-23S; near 90% for 16S). On the other hand, with the WIMP workflow and NCBI RefSeq database, a larger number of sequences are classified as "Other taxa". Again, this is especially remarkable when using the full-length 16S rRNA gene, with > 50% of the taxonomic assignments not hitting the expected bacterial species. The results were also confirmed by alpha diversity analyses: WIMP strategy overestimated the actual bacterial diversity, when compared to *rrn* DB and the reference (Figure 4B).

Complex microbial community analyses

We profiled two complex and uncharacterized microbial communities from dog skin (chin and dorsal). We used both long-amplicon markers and the two bioinformatics approaches –Minimap2 and *rrn* DB and WIMP with NCBI RefSeq database– to corroborate the results.

For chin samples of healthy dogs, we found a high abundance of *Pseudomonas* species (>40% of total relative abundance using 16S rRNA and >60% using 16S-ITS-23S) followed by other genus with lower abundances such as *Erwinia* and *Pantoea*. Focusing on *Pseudomonas*, at the species level we were able to detect that the most abundant species was *Pseudomonas koreensis*, followed by *Pseudomonas putida* and *Pseudomonas fluorescens* (Figure 5A and Supplementary Table 3). On the other hand, dorsal skin samples were dominated by bacteria from the genera *Stenotrophomonas*, *Sanguibacter*, and *Bacillus*. We reached species level for *Stenotrophomonas rhizophila* and *Sanguibacter keddieii*. It should be noted that *Glutamicibacter arilaitensis* is the same species as *Arthrobacter arilaitensis*, with newer nomenclature (Figure 5B and Supplementary Table 3). For both skin sample replicates, the results of the most abundant species converged using the two different methods and allowed for characterizing this complex low-biomass microbial community at the species level.

Finally, analyzing the dorsal skin samples, we also detected the presence of contamination from the previous nanopore run (Supplementary Table 2). We sequenced dorsal skin samples twice: one with a barcode previously used for sequencing the HM-783D mock community and another one with a new barcode. We were able to detect mock community representatives within the re-used barcode (Figure 5B). Some of them were found only in the sample that was using the re-used barcode (Sample_1); others were also present in the skin sample, such as *Bacillus cereus* or *Staphylococcus aureus*. In total, this contamination from the previous run was representing ~6% of the sample composition.

De novo assembly of the 16S-ITS-23S genetic marker

To further confirm the results obtained with the complex samples using directly the raw reads, we performed the assembly and consensus of the 16S-ITS-23S genetic marker using canu and we assigned taxonomy of the consensus sequences using BLAST.

For the HM-783D mock community, we were able to retrieve some of the most abundant bacterial species blasting with >99% of identity to their reference (*Escherichia coli*, *Staphylococcus*

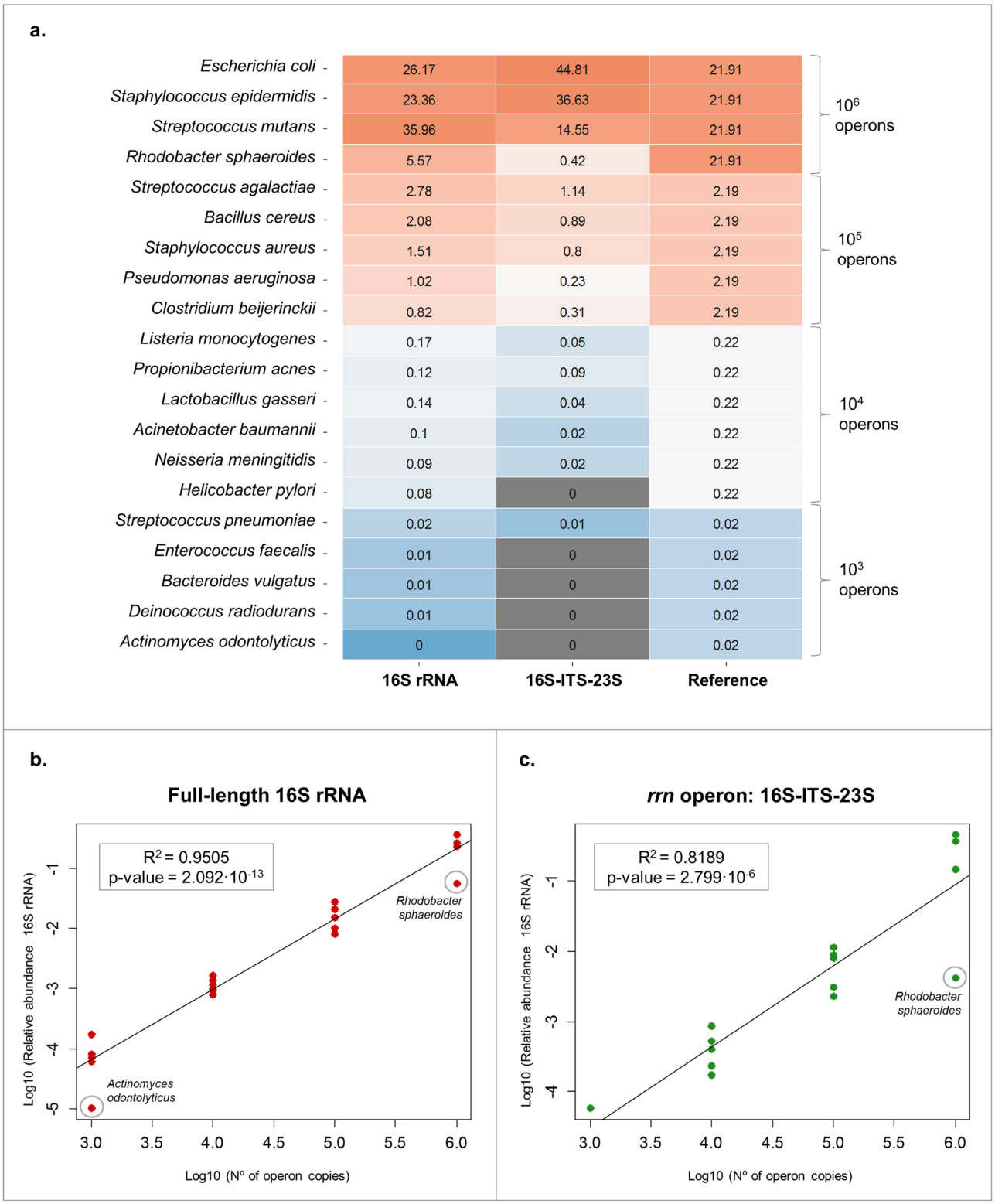

Figure 3. HM-783D mock community analysis. (**A**) Heat map representing the HM-783D mock community composition when mapped to its mock database. Grey colour represents the bacteria that were not detected (<10^4 copies with *rrn* operon). (**B**) Linear regression analysis of relative read proportions obtained using full-length 16S rRNA gene for all bacterial species present in HM-783D mock community and the actual operon copies (in log scale). (**C**) Linear regression analysis of relative read proportions obtained using the 16S-ITS-23S genetic marker for all bacterial species present in HM-783D mock community and the actual operon copies (in log scale).

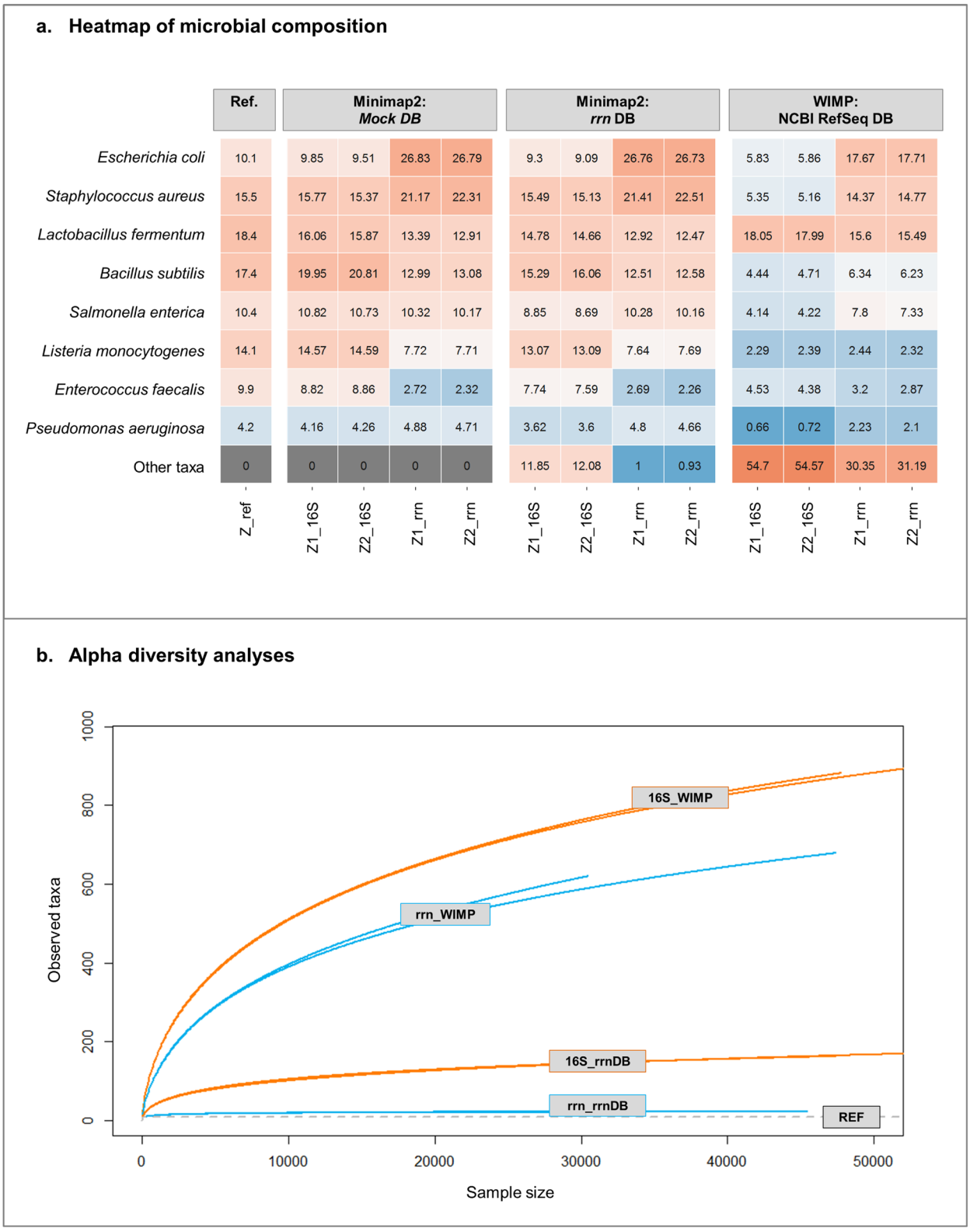

Figure 4. Zymobiomics mock community taxonomic analysis and diversity. (**A**) Heat map representing the relative abundance of the Zymobiomics mock community. "REF" column represents the theoretical composition of the mock community regarding the 16S rRNA gene content of each bacterium. (**B**) Alpha diversity rarefaction plot using observed taxa metrics.

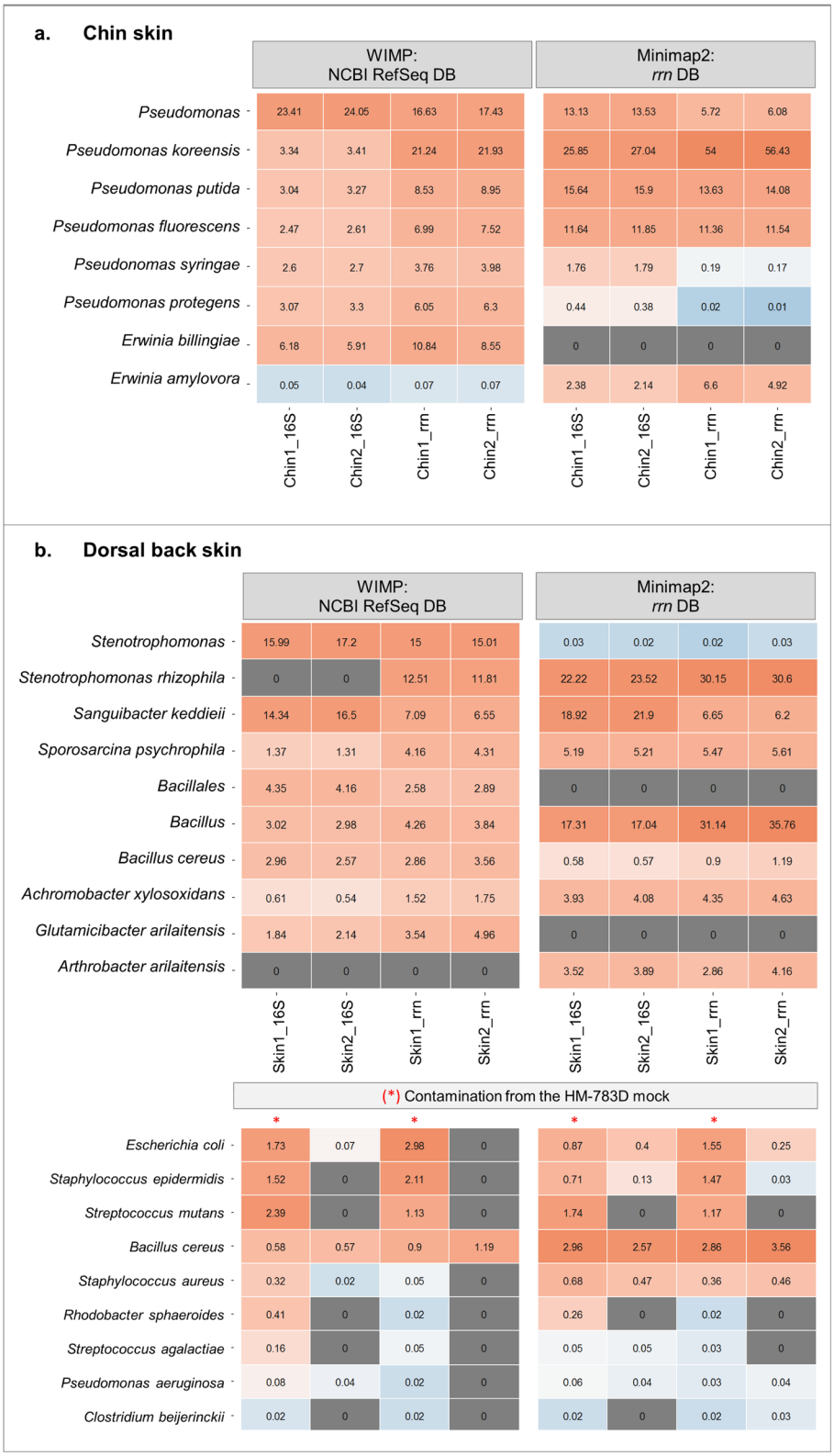

Figure 5. Microbiota composition of complex communities: skin samples of healthy dogs. (**A**) Chin samples: heat map representing the relative abundance of the main bacterial species in chin samples using WIMP and Minimap2. (**B**) Dorsal skin samples: heat map representing the relative abundance of the main bacterial species using WIMP and Minimap2. The lower heat map represents the remaining contamination from the previous run using HM-783D mock community within the same flowcell. Samples marked with a red line shared barcode with the mock community.

epidermidis, *Streptococcus mutans* and *Clostridium beijerinnkii*). For the Zymobiomics mock community, we found a consensus sequence for all the bacterial species with >99% of identity. The only two exceptions were *Salmonella enterica* that presented a consensus sequence with an identity of 98.8% and *Escherichia coli* that presented no consensus sequence (Table 2 and Supplementary Table 4).

For the complex microbial communities, the *de novo* consensus sequences usually presented lower identity than those of the mock community. For the dorsal skin, the ones with higher consensus accuracy were *Stenotrophomonas rhizophila* (99.3 %), *Streptococcus mutans* (99.1 %) and *Arthrobacter arilaitensis* (98.6 %), thus confirming the results retrieved using directly the raw reads. For the chin, the three contigs with higher consensus accuracy hit *Pseudomonas fluorescens* (>98.5%). We detected also other contigs with lower consensus accuracy values, with previously not seen taxonomic assignments (Supplementary Table 4).

Discussion

The full-length 16S rRNA and the 16S-ITS-23S of the *rrn* operon identified the bacterial isolate and revealed the microbiota composition of the mock communities and the complex skin samples, even at the genus and the species level. So, we present this long-amplicon approach as a method to profile the microbiota of low-biomass samples at deeper taxonomic levels.

The long amplicons were analyzed as raw reads (uncorrected), using both a mapping-based approach through Minimap2 and the "What's in My Pot" workflow by Oxford Nanopore to confirm results using a double approach. Although Nanopore sequencing has a high error rate (average accuracy for the *S. pseudintermedius* isolate: 89%), we compensated this low accuracy with longer fragments to assess the taxonomy of several bacterial communities. In general, the longer the marker, the higher the taxonomical resolution with both analyses performed. In this case, the longer 16S-ITS-23S marker remarkably improved the taxonomy assignment at the species level.

Moreover, we also performed de novo assembly of the 16S-ITS-23S amplicons obtaining consensus sequences that allowed us to validate some of the taxonomy retrieved with the long-amplicon raw reads. A de novo approach allowed retrieving consensus sequences with high accuracy (>99% of identity) for simple microbial communities. However, when working with more complex communities this consensus accuracy was generally lower. These lower accuracies found in complex microbial communities could be due to the lower sequencing depth, an uneven distribution of the bacterial species, and a mix of some closely similar species within the same contig.

Mock communities' analyses allowed us assessing the performance and the biases of the methodology, from the lab bench to the bioinformatics analyses and final results. In general, we found that the full-length 16S rRNA gene represents better the abundances of a microbial community; whereas, 16S-ITS-23S obtains better resolution at the species level.

So, do the long-amplicon approaches represent the actual bacterial composition? On one hand, we detected biases of our primer sets for both genetic markers, since some of the species of the mock communities were over- and under-represented. For example, *Actinomyces odontolyticus* and *Rhodobacter sphaeroides* seem to not amplify properly, neither with 16S rRNA gene, nor the 16S-ITS-23S of the *rrn* operon. Previous studies also detected the same pattern for these specific bacteria even when using or comparing different primer sets[16,21]. Overall, the 16S rRNA primer set seemed less biased than the 16S-ITS-23S of the *rrn* operon, which over- and underrepresented most of the bacterial species, suggesting that the 16S-ITS-23S primers should be improved for unbiased representation of the community.

On the other hand, with the HM-783D staggered mock community –with some low-abundant species– we aimed to assess the sensitivity of both approaches. With the 16S rRNA marker gene, we detected all bacterial members of both mock communities. However, when using the 16S-ITS-23S of the

Table 2. Zymobiomics contigs for the 16S-ITS-23S genetic marker and their taxonomic assignment through obtained through blasting.

Contig	Length	n° of reads	covStat	NCBI name	NCBI Accession	% Query coverage	e-value	% identity
tig00000001	4,346	10,900	13,568.84	*Salmonella enterica*	CP012344.2	99%	0.0	98.8%
tig00000003	3,972	2,237	15,554.25	*Bacillus subtilis*	CP002183.1	99%	0.0	99.4%
tig00000004	4,253	7,519	17,430.61	*Staphylococcus aureus*	CP029663.1	99%	0.0	99.1%
tig00000007	4,188	1,588	19,724.96	*Pseudomonas aeruginosa*	CP032257.1	100%	0.0	99.5%
tig00000009	4,061	645	18,811.23	*Enterococcus faecalis*	CP025021.1	100%	0.0	99.1%
tig00000010	4,064	1,107	17,360.31	*Listeria monocytogenes*	CP035187.1	99%	0.0	99.2%
tig00000012	4,019	1,347	16,386.90	*Lactobacillus fermentum*	CP034099.1	100%	0.0	99.4%

rrn operon, some of the low-abundant species in the HM-783D mock community were not detected. This was probably due to the fact that we obtained a lower number of reads –up to one magnitude less than with the 16S rRNA gene. Since we combined 16S rRNA and 16S-ITS-23S amplicons in the same run, this led to an underrepresentation of the 16S-ITS-23S amplicons and consequently a lower sequencing depth. This was probably due to the combination of various issues: i) not enough DNA mass to begin with the indicated number of molecules; ii) reads with shorter size than expected (~1,500 bp); iii) shorter fragments tend to be sequenced preferentially with Nanopore sequencing. Thus, for future studies our recommendation would be multiplexing samples with the same amplicon size to avoid underrepresentation of the longest one and improving PCR parameters or adding more PCR cycles to the longer amplicons to get more input DNA mass.

In the bioinformatics analyses, our aim was to confirm the results with two independent workflows and different databases rather than comparing them. We saw that the most abundant species were usually concordant with both strategies at a qualitative level. Some exceptions were due to the lack of that species in the *rrn* database, such as that seen for *S. pseudintermedius*. With the WIMP workflow and the 16S rRNA gene, many sequences did not reach species level. Previous studies analyzing the microbiota obtained with Nanopore reads have compared the performance of several databases using the 16S-ITS-23S[26] and software for the 16S rRNA gene[22]. They also found similar results as reported here: some false positives associated to specific software[22], as well as a high impact on the unclassified reads depending on the size of the database used[26].

When using EPI2ME (WIMP with NCBI Ref database), the amplicons from the *S. pseudintermedius* isolate were assigned to the correct bacterial species in ~98% and ~68% of the cases, using the 16S-ITS-23S of the *rrn* operon and 16S rRNA gene, respectively. In a previous study, Moon and collaborators used the full-length 16S rRNA gene for characterizing an isolate of *Campylobacter fetus* and the marker correctly assigned the species for ~89% of the sequences using EPI2ME[23]. The ratio of success on the correct assignment at the species level depends on the species itself and its degree of sequence similarity in the selected genetic marker. Within the *Staphylococcus* genus, the 16S rRNA gene presents the highest similarity (around ~97%) when compared to other genetic markers[37].

On the other hand, we observed that the mapping strategy (through Minimap2) could lead to a wrong assigned species if the interrogated bacterium is not represented on the chosen database. Minimap2 provides faster results than EPI2ME, but it needs an accurate comprehensive and representative database. Extra filtering steps using the alignment block length or Smith-Waterman alignment score could potentially be used to discard a wrong taxonomic assignment.

Switching to complex microbial communities, we found that dog chin was colonized by different *Pseudomonas* species. Recently, Meason-Smith and collaborators found *Pseudomonas* species associated with malodor in bloodhound dogs[38]. However, these were not the main bacteria found within the skin site tested, but were in low abundance, differing from what we have found here. On the other hand, Riggio and collaborators detected *Pseudomonas* as one of the main genera in canine oral microbiota in the normal, gingivitis and periodontitis groups[39]. However, the *Pseudomonas* species were not the same that we have detected here. It should be noted that we had characterized these chin samples with 16S V1-V2 amplicons in a previous study[27], where we found some mutual exclusion patterns for *Pseudomonadaceae* family. This taxon showed an apparent "invasive pattern", which could be mainly explained for the recent contact of the dog with an environmental source that contained larger bacterial loads before sampling[27]. Thus, our main hypothesis is that the *Pseudomonas* species detected on dog chin came from the environment, since they have been previously isolated from environments such as soil or water sources[40,41].

None of the most abundant species in dog dorsal skin had previously been associated with healthy skin microbiota either in human or in dogs. Some of them have an environmental origin, such as *Stenotrophomonas rhizophila*, which is mainly associated with plants[42]; or *Sporosarcina psychrophila,* which is widely distributed in terrestrial and aquatic environments[43]. The *Bacillus cereus* main reservoir is also the soil, although it can be a commensal of root plants and guts of insects, and can also be a pathogen for insects and mammals[44]. Overall, environmental-associated bacteria have already been associated with dog skin microbiota and are to be expected, since dogs constantly interact with the environment[27].

Regarding *Stenotrophomonas* in human microbiota studies, Flores *et al*. found that this genus was enriched in atopic dermatitis patients that were responders to emollient treatment[45]. However, previous studies on this skin disease found *Stenotrophomonas maltophila* associated to the disease rather than *Stenotrophomonas rhizophila*[46]. *Achromobacter xylosoxidans* has been mainly associated with different kind of infections, as well as skin and soft tissue infections in humans[47]. However, both dogs included in this pool were healthy and with representatives of both genus/species, a fact that reinforces the need to study the healthy skin microbiome at the species level before considering some species pathogenic. The other abundant bacteria detected on dog skin have been isolated in very different scenarios: *Sanguibacter keddieii* from cow milk and blood[48,49]; and *Glutamicibacter arilaitensis* (formerly *Arthrobacter arilaitensis*) is commonly isolated in cheese surfaces[50,51].

In general, we obtained taxonomy assignment down to species level with both the full-length 16S rRNA gene and the 16S-ITS-23S of the *rrn* operon, although it was not always feasible due to: i) high similarity of the marker chosen within some genera, especially for the 16S rRNA gene; ii) an incomplete database; and iii) sequencing errors. In the light of these results, for an increased resolution at the species level, the 16S-ITS-23S of the *rrn* operon would be the best choice. At the expenses of an increased taxonomic resolution, we could have missed few bacterial species due to unlinked *rrn* genes. While in host-associated environments (e.g. gut) bacterial species with unlinked

rrn genes are rare, if profiling natural environments (e.g. soil) this approach may be missing a significant proportion of the diversity[52]. So far, this genetic marker does not have as many complete and curated databases as 16S rRNA gene. If choosing to use 16S-ITS-23S genetic marker, we could add some filtering steps to filter out the "wrongly assigned taxonomy" and have more reliable taxonomic results, both using the alignment block length and the AS score using Minimap2.

Other gene-marker strategies have been further described to profile the microbiota with Nanopore sequencing. For example, sequencing 16S rRNA genes by Intramolecular-ligated Nanopore Consensus (INC-seq)[15,53] that allowed retrieving corrected full-length 16S rRNA genes. Another approach would be sequencing the cDNA from size selected small subunit rRNAs that allows retrieving many 16S rRNA genes using a primer-free approach[54]. However, these alternative strategies have been applied to the 16S rRNA gene that has a limited taxonomic resolution within some genera. Recently an approach using unique molecular identifiers (UMIs) for obtaining corrected full-length *rrn* operon has been applied to characterize a mock community[55]. The characterization of full-length ribosomal operons by the UMI approach has the potential to expand databases to make them more comprehensive with higher taxonomic resolution.

Further studies should be aiming to obtain reads with higher accuracy, either using consensus methods or applying new developments (new techniques, new basecallers or new R10 pores, etc.). Studies comparing marker-based strategies with metagenomics will determine the most accurate marker for microbiota studies in low-biomass samples.

Data availability

Underlying data

The datasets analyzed during the current study are available in the NCBI Sequence Read Archive, under the Bioproject accession number PRJNA495486.

Extended data

All the supplementary data has been added in an OSF repository (doi: http://doi.org/10.17605/OSF.IO/8MYKV)[22]. We provide here a complete list:

- Supplementary Table 1. Primer sequences for amplifying the full-length 16S rRNA gene and 16S-ITS-23S of the *rrn* operon.

- Supplementary Table 2. Samples included in the study, run summary and quality control results. *For mock communities, the mock DB was used. For complex communities, the *rrn* DB.

- Supplementary Table 3. Taxonomic assignments table of each sample with the different approaches.

- Supplementary Table 4. De novo results obtained with canu and their taxonomic assignment using BLAST.

- Supplementary Figure 1. Photo of the agarose gel electrophoresis of some of the samples.

- Supplementary Figure 2. Main workflow overview. Detailed bioinformatics workflow can be found in Supplementary File 1.

- Supplementary File 1. Bioinformatics workflow used for the mapping approach.

Grant information

This work was supported by two grants awarded by Generalitat de Catalunya (Industrial Doctorate program, 2013 DI 011 and 2017 DI 037).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Salter SJ, Cox MJ, Turek EM, *et al.*: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.** *BMC Biol.* 2014; **12**(1): 87.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Kong HH, Andersson B, Clavel T, *et al.*: **Performing Skin Microbiome Research: A Method to the Madness.** *J Invest Dermatol.* 2017; **137**(3): 561–568.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Ludwig W, Schleifer KH: **Bacterial phylogeny based on 16S and 23S rRNA sequence analysis.** *FEMS Microbiol Rev.* 1994; **15**(2–3): 155–173.
PubMed Abstract | Publisher Full Text

4. Yarza P, Ludwig W, Euzéby J, *et al.*: **Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses.** *Syst Appl Microbiol.* 2010; **33**(6): 291–299.
PubMed Abstract | Publisher Full Text

5. Iwen PC, Hinrichs SH, Ruppy ME: **Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal pathogens.** *Med Mycol.* 2002; **40**(1): 87–109.
PubMed Abstract | Publisher Full Text

6. Hibbett DS, Ohman A, Glotzer D, *et al.*: **Progress in molecular and morphological taxon discovery in *Fungi* and options for formal classification of environmental sequences.** *Fungal Biol Rev.* 2011; **25**(1): 38–47.
Publisher Full Text

7. Clarridge JE 3rd: **Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.** *Clin Microbiol Rev.* 2004; **17**(4): 840–862.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Janda JM, Abbott SL: **16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls.** *J Clin Microbiol.* 2007; **45**(9): 2761–2764.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Walters WA, Caporaso JG, Lauber CL, *et al.*: **PrimerProspector: *de novo* design and taxonomic analysis of barcoded polymerase chain reaction primers.** *Bioinformatics.* 2011; **27**(8): 1159–1161.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Kuczynski J, Lauber CL, Walters WA, *et al.*: **Experimental and analytical tools for studying the human microbiome.** *Nat Rev Genet.* 2012; **13**(1): 47–58.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Grice EA: **The skin microbiome: potential for novel diagnostic and therapeutic approaches to cutaneous disease.** *Semin Cutan Med Surg.* 2014; **33**(2): 98–103.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Chng KR, Tay AS, Li C, *et al.*: **Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare.** *Nat Microbiol.* 2016; **1**(9): 16106.
PubMed Abstract | Publisher Full Text

13. Pierezan F, Olivry T, Paps JS, *et al.*: **The skin microbiome in allergen-induced canine atopic dermatitis.** *Vet dermatol.* 2016; **27**(5): 332–e82.
PubMed Abstract | Publisher Full Text

14. Bradley CW, Morris DO, Rankin SC, *et al.*: **Longitudinal Evaluation of the Skin Microbiome and Association with Microenvironment and Treatment in Canine Atopic Dermatitis.** *J Invest Dermatol.* 2016; **136**(6): 1182–90.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Li C, Chng KR, Boey EJ, *et al.*: **INC-Seq: accurate single molecule reads using nanopore sequencing.** *GigaScience.* 2016; **5**(1): 34.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Benítez-Páez A, Portune KJ, Sanz Y: **Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer.** *GigaScience.* 2016; **5**: 4.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Brown BL, Watson M, Minot SS, *et al.*: **MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach.** *GigaScience.* 2017; **6**(3): 1–10.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Shin J, Lee S, Go MJ, *et al.*: **Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing.** *Sci Rep.* 2016; **6**: 29681.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Ma X, Stachler E, Bibby K: **Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization.** *bioRxiv.* 2017.
Publisher Full Text

20. Shin H, Lee E, Shin J, *et al.*: **Elucidation of the bacterial communities associated with the harmful microalgae *Alexandrium tamarense and Cochlodinium polykrikoides* using nanopore sequencing.** *Sci Rep.* 2018; **8**(1): 5323.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Cusco A, Vines J, D'Andreano S, *et al.*: **Using MinION to characterize dog skin microbiota through full-length 16S rRNA gene sequencing approach.** *bioRxiv.* 2017.
Publisher Full Text

22. Mitsuhashi S, Kryukov K, Nakagawa S, *et al.*: **A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer.** *Sci Rep.* 2017; **7**(1): 5657.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Moon J, Kim N, Lee HS, *et al.*: ***Campylobacter fetus* meningitis confirmed by a 16S rRNA gene analysis using the MinION nanopore sequencer, South Korea, 2016.** *Emerg Microbes Infect.* 2017; **6**(11): e94.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Moon J, Jang Y, Kim N, *et al.*: **Diagnosis of *Haemophilus influenzae* Pneumonia by Nanopore 16S Amplicon Sequencing of Sputum.** *Emerg Infect Dis.* 2018; **24**(10): 1944–1946.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Benítez-Páez A, Sanz Y: **Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer.** *GigaScience.* 2017; **6**(7): 1–12.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Kerkhof LJ, Dillon KP, Häggblom MM, *et al.*: **Profiling bacterial communities by MinION sequencing of ribosomal operons.** *Microbiome.* 2017; **5**(1): 116.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Cuscó A, Belanger JM, Gershony L, *et al.*: **Individual signatures and environmental factors shape skin microbiota in healthy dogs.** *Microbiome.* 2017; **5**(1): 139.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Zeng YH, Koblížek M, Li YX, *et al.*: **Long PCR-RFLP of 16S-ITS-23S rRNA genes: a high-resolution molecular tool for bacterial genotyping.** *J Appl Microbiol.* 2013; **114**(2): 433–447.
PubMed Abstract | Publisher Full Text

29. Klindworth A, Pruesse E, Schweer T, *et al.*: **Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.** *Nucleic Acids Res.* 2013; **41**(1): e1.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Wick R: **Porechop**.
Reference Source

31. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Marijon P: **yacrd: Yet Another Chimeric Read Detector for long reads**.
Reference Source

33. Juul S, Izquierdo F, Hurst A, *et al.*: **What's in my pot? Real-time species identification on the MinION.** *bioRxiv.* 2015.
Publisher Full Text

34. Kim D, Song L, Breitwieser FP, *et al.*: **Centrifuge: rapid and sensitive classification of metagenomic sequences.** *Genome Res.* 2016; **26**(12): 1721–1729.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Andersen KS, Kirkegaard RH, Karst SM, *et al.*: **ampvis2: an R package to analyse and visualise 16S rRNA amplicon data.** *bioRxiv.* 2018.
Publisher Full Text

36. McMurdie PJ, Holmes S: **phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.** *PLoS One.* 2013; **8**(4): e61217.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Ghebremedhin B, Layer F, König W, *et al.*: **Genetic classification and distinguishing of *Staphylococcus* species based on different partial *gap*, 16S rRNA, *hsp60*, *rpoB*, *sodA*, and *tuf* gene sequences.** *J Clin Microbiol.* 2008; **46**(3): 1019–1025.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Meason-Smith C, Older CE, Ocana R, *et al.*: **Novel association of *Psychrobacter* and *Pseudomonas* with malodour in bloodhound dogs, and the effects of a topical product composed of essential oils and plant-derived essential fatty acids in a randomized, blinded, placebo-controlled study.** *Vet Dermatol.* 2018.
PubMed Abstract | Publisher Full Text

39. Riggio MP, Lennon A, Taylor DJ, *et al.*: **Molecular identification of bacteria associated with canine periodontal disease.** *Vet Microbiol.* 2011; **150**(3–4): 394–400.
PubMed Abstract | Publisher Full Text

40. Peix A, Ramírez-Bahena MH, Velázquez E: **Historical evolution and current status of the taxonomy of genus *Pseudomonas*.** *Infect Genet Evol.* 2009; **9**(6): 1132–1147.
PubMed Abstract | Publisher Full Text

41. Mehri I, Turki Y, Chair M, *et al.*: **Genetic and functional heterogeneities among fluorescent *Pseudomonas* isolated from environmental samples.** *J Gen Appl Microbiol.* 2011; **57**(2): 101–14.
PubMed Abstract | Publisher Full Text

42. Wolf A, Fritze A, Hagemann M, *et al.*: **Stenotrophomonas rhizophila sp. nov., a novel plant-associated bacterium with antifungal properties.** *Int J Syst Evol Microbiol.* 2002; **52**(Pt 6): 1937–1944.
PubMed Abstract | Publisher Full Text

43. Yan W, Xiao X, Zhang Y: **Complete genome sequence of the *Sporosarcina psychrophila* DSM 6497, a psychrophilic *Bacillus* strain that mediates the calcium carbonate precipitation.** *J Biotechnol.* 2016; **226**: 14–15.
PubMed Abstract | Publisher Full Text

44. Ceuppens S, Boon N, Uyttendaele M: **Diversity of *Bacillus cereus* group strains is reflected in their broad range of pathogenicity and diverse ecological lifestyles.** *FEMS Microbiol Ecol.* 2013; **84**(3): 433–450.
PubMed Abstract | Publisher Full Text

45. Seite S, Flores GE, Henley JB, *et al.*: **Microbiome of affected and unaffected skin of patients with atopic dermatitis before and after emollient treatment.** *J Drugs Dermatol.* 2014; **13**(11): 1365–1372.
PubMed Abstract

46. Dekio I, Sakamoto M, Hayashi H, *et al.*: **Characterization of skin microbiota in patients with atopic dermatitis and in normal subjects using 16S rRNA gene-based comprehensive analysis.** *J Med Microbiol.* 2007; **56**(Pt 12): 1675–1683.
PubMed Abstract | Publisher Full Text

47. Tena D, Martínez NM, Losa C, *et al.*: **Skin and soft tissue infection caused by Achromobacter xylosoxidans: report of 14 cases.** *Scand J Infect Dis.* 2014; **46**(2): 130–135.
PubMed Abstract | Publisher Full Text

48. Fernández-Garayzábal JF, Dominguez L, Pascual C, *et al.*: **Phenotypic and phylogenetic characterization of some unknown coryneform bacteria isolated from bovine blood and milk: description of *Sanguibacter* gen.nov.** *Lett Appl Microbiol.* 1995; **20**(2): 69–75.
PubMed Abstract | Publisher Full Text

49. Ivanova N, Sikorski J, Sims D, *et al.*: **Complete genome sequence of *Sanguibacter keddieii* type strain (ST-74).** *Stand Genomic Sci.* 2009; **1**(2): 110–118.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Busse HJ: **Review of the taxonomy of the genus *Arthrobacter*, emendation of the genus *Arthrobacter sensu* lato, proposal to reclassify selected species of the genus *Arthrobacter* in the novel genera *Glutamicibacter* gen. nov., *Paeniglutamicibacter* gen. nov., *Pseudoglutamicibacter* gen. nov., *Paenarthrobacter* gen. nov. and *Pseudarthrobacter* gen. nov., and emended description of *Arthrobacter roseus*.** *Int J Syst Evol Microbiol.* 2016; **66**(1): 9–37.
PubMed Abstract | Publisher Full Text

51. Irlinger F, Bimet F, Delettre J, *et al.*: ***Arthrobacter bergerei* sp. nov. and *Arthrobacter arilaitensis* sp. nov., novel coryneform species isolated from the surfaces of cheeses.** *Int J Syst Evol Microbiol.* 2005; **55**(Pt 1): 457–462.
PubMed Abstract | Publisher Full Text

52. Brewer TE, Albertsen M, Edwards A, *et al.*: **Unlinked rRNA genes are widespread among Bacteria and Archaea.** *BioRxiv.* 2019.
Publisher Full Text

53. Calus ST, Ijaz UZ, Pinto AJ: **NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform.** *GigaScience.* 2018; **7**(12): 1–16.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Karst SM, Dueholm MS, McIlroy SJ, *et al.*: **Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias.** *Nat Biotechnol.* 2018; **36**(2): 190–195.
PubMed Abstract | Publisher Full Text

55. Karst SM, Ziels RM, Kirkegaard RH, *et al.*: **Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing.** *Biorxiv.* 2019.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ✔ ? ?

Version 2

Reviewer Report 20 September 2019

https://doi.org/10.5256/f1000research.21950.r51954

?

Lee J. Kerkhof
Department of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA

The authors have begun to put their findings into context with prior studies in the literature using MinION and rRNA genes or operons. However, I disagree with their conclusions that the primers used for rRNA operon amplification need improvement rather than the choice of software and database. Specifically, Cuscó *et al.* report that mock databases containing only those targets in the mock community can correctly assign OTUs using WIMP and MiniMap2 (Fig 3 and 4). While larger databases such as a rRNA operon database or the NCBI Refseq database incorrectly assign taxa (from 12-55% of reads). This may be true for the software they employed. However, other studies have used QIIME, BLASTN, Centrifuge, and Discontinuous MegaBLAST to identify OTUs for the MinION platform for 16S rRNA genes or ribosomal operons. I think an acknowledgement that these various software packages have also been employed to analyze the MinION reads in the scientific literature would benefit the F1000 readership.

Especially since the authors indicate that our study (Kerkhof *et al.*, 2017[1]) supports their findings which is incorrect. We actually found the databases which provided the most number of OTU calls for our raw reads were the NCBI 16S rRNA database (>18K entries) and the ARB/Silva NR99 database (>645K entries). The intermediate size databases (Greengenes and RDP) did not work as well using Discontinuous Megablast. I suspect the issue with incorrect OTU calling stems more from the software used to screen the database than from the primers or the size of the database being used.

A brief discussion of variability resulting from different PCR amplification reagents and what might be considered a best practice for data analysis for future studies would be helpful.

Other Specific Comments:

1. Figure 3 has the calculated target gene concentration “Reference” on the right side of the heat map while Figure 4 has the “Reference” on the left side. This makes it hard for the reader to figure out what should be for the WIMP or MiniMap2 results as a basis for comparison.

References

1. Kerkhof L, Dillon K, Häggblom M, McGuinness L: Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome*. 2017; **5** (1). Publisher Full Text

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular ecology of microbial systems

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 21 February 2019

https://doi.org/10.5256/f1000research.18384.r43567

?

Lee J. Kerkhof
Department of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA

F1000 Research Cuscó et al. (https://doi.org/10.12688/f1000research.16817.1)

Comments to the authors:

The manuscript describes a study assaying 2 mock bacterial communities or 2 complex skin microbiome samples from dogs (chin or dorsal back) using both near-full length 16S rRNA genes and near-full length rRNA operons with the Oxford Nanopore MinION. The authors employ a library preparation method generating either 16S amplicons (1400 bp) or rRNA operons (4500 bp) including barcoding with a 1D ligation/sequencing kit and FLO-MIN 106 cells. The data analysis pipeline utilized Albacore basecalling, near-full length amplicon size selection, and screening by What's in my Pot (WIMP) and Minimap2 against both NCBI and rrn databases. The authors demonstrate increased resolution at the species level with longer reads, that there can be large losses of raw sequence reads by size selection for rrn amplicons in their hands, and that the data analysis software and database can influence the results of MinION bacterial community analysis.

It would have been very helpful for the authors to put these findings into context with other papers in the literature using MinION and rRNA genes. For example, their results support what others directly sequencing near-full length16S amplicons (e.g. Shin *et al.* (2016[1]), Mitsuhashi *et al.* (2017[2]), and Benitez-Paez *et al.* (2016[3])) or rRNA operons (e.g. Benitez-Paez *et al.* (2017[4]), Kerkhof *et al.* (2017[5])) have shown in mock communities or complex samples with respect to species-level resolution. Additionally, the screening of MinION reads with different 16S rRNA databases has also been described in the supplementary figures of Kerkhof *et al.* (2017[5]). Likewise, an acknowledgement of the various

software packages that has been employed to analyze the MinION reads in the scientific literature would benefit the readership. It appears that QIIME, BLASTN, Centrifuge, LAST aligner, Discontinuous MegaBLAST, WIMP, and MiniMap2 have all been used to identify OTUs for the MinION platform for 16S rRNA genes or rrn operons. As the authors have shown, the software/database being used can be very influential in the results of MinION screens and a synopsis of what they have found in context with other investigators (% bacterial assignment vs. % error) may point to a best practice for future studies.

Other Specific Comments:

1. **Page 3:** I find it awkward/confusing to indicate the number of operons per microorganism per microliter here for the mock communities. Bacteria generally have 1-15 ribosomal operons in their genomes. I think it is clearer to just indicate the number of target rRNA operons is 10^3-10^6 for this particular DNA mixture.

2. **Page 3:** The barcoding expansion pack (EXP-PBC001) requires that the primers contain overhangs attached to the rRNA primers. This is not mentioned by the authors. Did they put overhangs on 27F/1492R/2241R? If so, the first round of target amplification may be affected by the presence of these overhangs. This should be indicated.

3. **Page 4:** The authors clearly show the danger of performing PCR and only characterizing the amplification product by Qubit fluorescence. If they had done agarose gels on the PCR reactions, they may have detected the short amplification products in their initial rrn operon reactions. Furthermore, these short reads are preferentially ligated using the SQK-LSK108 sequencing kit since there are more picomole ends. This best practice of visualizing PCR amplifications for size determinations before sequencing should be explicitly stated.

4. **Page 4:** I am a little confused by the 0.5 nM notation for PCR product in the barcoding reaction. If the authors used 50 microliter reactions, did they put 25 ng of 1st round PCR product in their barcoding reactions for a 15 cycle amplification? Can the authors just state the mass of DNA used to barcode? Secondly, Table 2 indicates BC1, BC2, and BC3 were not used. Was there a reason these barcodes were not utilized?

5. **Page 6:** Stating that the rrn operon profiling was more biased probably because of the lower sequencing depth does not recognize that others have not reported comparable bias or that it is probably a reflection of their potentially compromised amplification efficiencies. This conclusion should be viewed with caution, given the amplification issues noted above.

6. **Page 11:** The running of shorter (1500 bp) and longer (4500 bp) libraries on the same flow cell at the same time should enrich for the shorter reads. The MinION uses electrophoresis to move DNA molecules through the pores and smaller fragments should mobilize easier.

References

1. Shin J, Lee S, Go M, Lee S, Kim S, Lee C, Cho B: Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific Reports*. 2016; **6** (1). Publisher Full Text
2. Mitsuhashi S, Kryukov K, Nakagawa S, Takeuchi J, Shiraishi Y, Asano K, Imanishi T: A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Scientific Reports*. 2017; **7** (1). Publisher Full Text
3. Benítez-Páez A, Portune KJ, Sanz Y: Species-level resolution of 16S rRNA gene amplicons sequenced

through the MinION™ portable nanopore sequencer.*Gigascience*. 2016; **5**: 4 PubMed Abstract | Publisher Full Text
4. Benítez-Páez A, Sanz Y: Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. *GigaScience*. 2017; **6** (7). Publisher Full Text
5. Kerkhof L, Dillon K, Häggblom M, McGuinness L: Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome*. 2017; **5** (1). Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular ecology of microbial systems

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 18 February 2019

https://doi.org/10.5256/f1000research.18384.r43564

?

Kon Chu
Department of Neurology, Seoul National University, Seoul, South Korea

The study compared the results of microbiota profiling using two different markers (16S rRNA and the rrn operon) and different classification methods. Because other reviewers have already made comprehensive reviews and comments including several critical points, I would like to add only a few

minor points to the manuscript:

1. Figure 2: according to the text, *Actinomyces odontolyticus* was detected using the 16S rRNA gene, however, '0' in the figure can create confusion. It would be better to represent the number of copies of *Actinomyces odontolyticus* using more decimal places or adding a caption for this species.

2. Figure 3a:

   - It would be better to change the figure (e.g. heatmap) to make it easier for readers to recognize under-represented and over-represented bacteria. *Listeria monocytogenes* also seems under-represented in the analyses using the mock database and rrn database, and the corresponding sentence in Page 7 may be changed.

   - Include the classification method (WIMP, minimap2) along with the name of the database, as in figure 4, to allow general readers to more easily match the methods and the database.

3. In the last paragraph of page 7, it seems that the criteria of the percentage of wrongly assigned species for the rrn operon are different from that for the 16S rRNA gene.

4. Table 3: I suggest making a caption for the difference between 'Staphylococcus' and 'Other Staphylococcus'.

5. If the authors would like to insist on better resolution by using the rrn operon, they need to demonstrate the data of the analysis using multiple species including species that tend to be under-represented or over-represented.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: neuroinfection, encephalitis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 08 February 2019

https://doi.org/10.5256/f1000research.18384.r43563

Amanda Warr
Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK

Cusco *et al.* evaluate methods for long read sequencing and classification of marker genes from microbial communities, both for mock communities of known microbial composition and complex communities from dog skin from two anatomical locations, chin and back. They find that long read sequencing of 16S and the *rrn* operon is sufficiently accurate to classify microbes and that *rrn* is more sensitive at the species level.

This work demonstrates a valuable option for species identification from microbial samples using long reads, overcoming the current high error rate through covering a larger region. The work also highlights some of the issues that can arise from multiplexing amplicons of differing lengths when using Nanopore sequencing.

Overall the paper is well written and detailed, however there are a few details I feel could be addressed:

1. **16S length reads in *rrn* barcodes:** Do you expect this to be entirely from barcode misassignment or were these shorter fragments produced during PCR? You state that the loss of *rrn* amplicons during the length trimming step was probably due to over-representation of 16S amplicon on the flow cell, and most of the reads lost were roughly 16S amplicon sized - are you suggesting that there are large numbers of 16S reads that are assigned to *rrn* barcodes after 2 rounds of demultiplexing? Are these shorter reads actually whole 16S amplicons or fragments of *rrn*?

2. **Expected sensitivity given read count:** The authors state that failure to detect the less abundant species from the mock community in the *rrn* dataset was "probably" due to their being fewer reads. As the proportions of the species in the mock samples are known, theoretically what total number of reads would be necessary to detect the less abundant species? Given the number of rrn reads obtained, did the authors detect as many species as they would expect to detect and what is the minimum total number of reads they would need to be likely to detect the lowest abundance species?

3. **Differences in classification methods:** Differences in classifications between the mock community database/rrn database and the NCBI database may be attributable to differences in the tools, with minimap2 being used for the mock and *rrn* databases and WIMP (based on centrifuge)

being used for the NCBI database. My understanding is that the authors are mainly interested in classifications from different databases rather than differences between methods. While the authors do not directly compare the classification results between these different methods in text, some of the figures appear to imply that these results are directly comparable (e.g. Figure 3a). It would be useful if either all three databases were used with a single method (for example, using centrifuge with all three databases) or if these were at least more obviously separated or marked as coming from different classification methods in the figures.

4. **Classification rates against NCBI:** The authors should further discuss ways to improve the classification rates, will the biggest improvements come from reduced error rate, better classification tools, improving species representation in databases? The authors conclude that in the future we should aim to improve accuracy, but one of the main results here is that sequencing the full 16S/*rrn* overcomes the problem of the current error rate - perhaps highlight benefits such as improved barcode assignment and emphasise that while this works well classification against a large database would likely improve with increased accuracy. The authors also conclude that *rrn* offers higher resolution at species level, however I suspect that currently more species have 16S sequences in databases than *rrn*.

Additionally, I have a few minor corrections mainly around small grammatical errors and figure/table modifications:

- Page 5: Paragraph beginning "To assign taxonomy...", change "to strategies" to "two strategies". Also I would change the last sentence on the page to say "some of the reads excluded were the expected length of the 16S rRNA gene rather than the *rrn* operon". Figure 1 should also be labelling Albacore as the basecaller.
- Page 6: change "would allow us determining" to "would allow us to determine".
- Page 11, column 2, line 2: change "associated to" to "associated with".
- Figure 3a would benefit from separating the reference bar from the other bars or adding this bar to the other two plots (currently it is grouped with Mock database, but it is also relevant to the *rrn* database and the NCBI database).
- Figure 4 text is quite difficult to read.
- Table 2: the title of the final column isn't clear. Is this the % of reads that pass the quality filters before chimera detection? Could another column be added showing number of reads that pass this filter?
- Figure 5: there are several different colours of 0 in this heat map?

In the conclusion the authors have suggested ways to improve accuracy of this method in the future, I would add the R2C2 method (Volden *et al.*, 2018[1]) as an option to improve consensus accuracy here also, while designed for cDNA it could be applied to fragments of genomic DNA.

References

1. Volden R, Palmer T, Byrne A, Cole C, Schmitz R, Green R, Vollmers C: Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*. 2018; **115** (39): 9726-9731 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, long read sequencing, microbiome assembly

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 04 February 2019

https://doi.org/10.5256/f1000research.18384.r43566

?

Rasmus H. Kirkegaard
Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University (AAU), Aalborg, Denmark

Title:
"Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon".

Summary of the key results:
The study demonstrates the use of nanopore sequencing for characterising low biomass samples with high levels of host DNA using a primer-based approach targeting the entire 16S rRNA gene or the 16S rRNA gene and the 23S rRNA gene.

Furthermore, it evaluates the ability of these methods in the context of known references using mock communities and a pure culture using both the WIMP software and a custom mapping-based approach.

The study demonstrates that nanopore sequencing can give accurate classifications even at the current level of error rate if the reference database contains the right sequences. The study also shows how sequencing the longer fragment spanning both the 16S rRNA and 23S rRNA genes improves the taxonomic classification when the database contains a matching sequence.

Is the work clearly and accurately presented and does it cite the current literature?
The study mentions that the classification methods rely heavily on reference databases so it would be relevant to include citations for papers with methods for producing new reference sequences for both 16S rRNA and the longer fragment in the discussion (metagenomics, artificial long reads, primer free methods). Methods for improving read accuracy are also mentioned as important but the only methods mentioned are future upgrades from the company, relevant existing literature is not included (INC-seq, UMIs etc.). The study concludes that sequencing the entire “rrn operon” would be the best choice but it would be relevant to compare the size of current databases for the 16S rRNA gene versus the rrn operon. The presence of conserved sites for designing better primers is also extremely important but not discussed. Furthermore, there is evidence that quite a few organisms have unlinked rRNA genes, which will thus be missed by a full operon approach.

Citations are also needed for bioinformatics tools for both processing and visualisation of the data.

Is the study design appropriate and is the work technically sound?
The study uses mapping to a reference database to point out that the sequences can get genus- and species-level classification. However, the method will always report a genus and a species even in the absence of the correct sequence in the reference database as indicated from the sequencing of the *S. pseudintermedius* pure culture with the “rrn” method. It will be important to simulate the impact on the results when there is no closely related sequences in the database. This could be done by removing all reference sequences within the Gammaproteobacteria and mapping the HM-783D to the modified database and monitor where the reads end up. It would also be helpful if there was a way to distinguish between reads that have the “correct” match and reads that just happen to map because the 16S rRNA gene is extremely conserved. Something similar would be relevant for the EPI2ME workflow but as the authors cannot control the reference database, it is probably not feasible. One of the advantages of the mock communities should be information about the copy numbers for the rRNA genes but there is no information on this included in the study and how it affects the results.

Are sufficient details of methods and analysis provided to allow replication by others?
The methods section lacks information about what happens after mapping the reads. How are the figures generated, what software is used, etc.? It would also be helpful if the specific scripts/commands used to run the bioinformatics analysis were available.

Figures:

Figure 1: bioinformatic workflow:
The figure gives a decent overview of the bioinformatics processing but seems to miss the visualisation tools used. The main role of Albacore is basecalling the raw data not just demultiplexing. The figure could be improved further if you include the wet lab part of the work, so it becomes clear why the demultiplexing step is included and where the raw data comes from. A mapping step is integrated in the chimera detection (removal?) workflow but it might be better to omit mentioning mapping in that step as it can be confusing that the figure has two mapping steps in a row.

Figure 2: heatmap mock community:
The caption needs to explain what the numbers represent e.g. percentage of sequenced reads/mapped reads. It would be great if the heatmap included the “true” composition of the mock community for comparison. Copy number for each organism in the mock would also be relevant to include in the figure. Since there are only two columns, it would be better to have the sample labels at the top and with horizontal text preferably with a name that makes it easier to interpret the figure.

Figure 3a: stacked bar chart:
Even though stacked bar charts are very common it is not easy to read as they lack a common baseline for most of the values (See https://solomonmg.github.io/blog/2014/when-to-use-stacked-barcharts/ and https://peltiertech.com/stacked-bar-chart-alternatives/). I suggest that you use more of the heatmaps instead of introducing bar charts.

Figure 3b: rarefaction curves:
It would be great if you could add a dashed line for the expected “true” value for the mock community.

Figure 3c: WIMP tree:
This figure is quite complex to read. If the point with running both WIMP and a mapping-based approach with the two different amplicon types is to compare the methods, I suggest that you try to integrate the information better into one combined figure. This way you can help the reader to understand your message.

Figure 4a: stacked bar chart+heatmap dog samples:
Remove the stacked bar chart.

Figure 4b: stacked bar chart+heatmap dog samples:
Remove the stacked bar chart.

Getting rid of the bar charts would allow for making a big heatmap with the data from Figure 4A and 4B combined. This way the reader can also compare the results from the two different sample sites. A naming system that makes it clearer that “_1” and “_2” are replicates would also help the reader interpret the figure. Presenting results aggregated at different levels, which could be included in one another is a bit confusing e.g., “*Bacillus cereus*” could be included in “*Bacillus*” which again could be included in “*Bacillales*”.

Figure 5: heatmap mock community contamination:
It is confusing that several cells in the heatmap have a value of “0” but with very different colours. Adding some meaningful labels with the contamination vs no contamination on the top could help the reader understand the figure without reading the caption.

Tables:

Table 1: Primer sequences:
Fine but could be moved to supplementary.

Table 2: Samples and QC:
Make headers easier to understand e.g. ”% seq 1st QC” could be “% of reads passing QC”, “Albacore

pass" could just be "# reads after basecalling" etc. Where is the number after chimera detection?

Add a column with data accession ID and move the table to supplementary then the sample names can also be expanded so the reader does not have to look to the bottom for an explanation of abbreviations. I suggest adding a column at the end with the number of reads mapping/classified for each sample so the reader know what fraction is included.

Table 3: Pure culture comparison WIMP vs. mapping:
You need to make it clear in the table that *Staphylococcus pseudintermedius* is missing from the "rrn" database. As the paper mentions genus- and species-level classification as the target you may benefit from aggregating the values for *S. pseudintermedius* and *S. pseudintermedius* HKU10-03 as splitting this into strains makes it more confusing as your numbers in the text do not match the ones in the table.

Supplementary Table 1:
It would be great to include the mock communities in this table as well.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: I am a co-owner of DNASense ApS (www.dnasense.com)

Reviewer Expertise: microbial biotechnology, nanopore sequencing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 19 November 2018

https://doi.org/10.5256/f1000research.18384.r40373

Alfonso Benítez-Páez
Microbial Ecology, Nutrition and Health Research Unit, Institute of Agrochemistry and Food Technology-Spanish National Research Council (IATA-CSIC), Valencia, Spain

Cusco and co-workers present an evaluation of both a mock community and the dog skin associated microbiota. The authors made use of the single-molecule Nanopore DNA sequencing technology and compared two different technical approaches by studying the nearly-full 16S rRNA bacterial gene and the nearly-full bacterial rRNA operon.

In my opinion, this work represents an important advance regarding the application of nanopore technology in the field of microbiome research.

The main strength of the work is its detailed technical description regarding the protocols for library preparation, sequencing and basecalling, that altogether facilitate the reproducibility. Moreover, the genetic data generated was properly deposited in a specialized database for public accession to whomever may want to replicate the analysis of long reads by similar approaches or new ones.

The figure quality is good and the information disclosed by them is well accompanied with appropriate captions.

Notwithstanding, I have some minor concerns about the work that should be clarified, at least for me:

1. The last paragraph of page 7 describes the level of reads correctly assigned to species level for the microbial isolate *Staphylococcus pseudintermedius*. However, some of the values cited in the text do not match, at least, explicitly in Table 3. So, the authors should revise this issue or better describe the information obtained.

2. The authors found that the study of a nearly-full 16S rRNA gene reflects in a better way the expected abundances of microbial species present in the mock community tested. This comparative analysis with regard to the results obtained by using the rrn operon should be accompanied by a linear regression analysis, declaring respective Pearson's "r" coefficients, that can measure more accurately the efficiency of both methods and better support the authors' observations and conclusions.

3. Additionally to the observed richness (observed species) and Shannon diversity, the authors could also include a microbial community evenness evaluation of reference and observed microbiome data from the different approaches evaluated in the study, so that additional conclusions could be addressed.

4. Given the issues with underrepresentation of "rrn" data as a consequence of mixing this type of synthetic DNA with nearly-full 16S rRNA amplicons, the authors should highlight this observation as a major issue of this approach and state a clear recommendation to avoid this type of multiplexing for future studies.

5. It is necessary to better describe the contamination issues described in the last paragraph of the results (page 9). I'm not sure if this cross-contamination came from re-utilization of a flowcell or if this came from contamination of the barcoded-primer, used during nested PCR, with amplicons/DNA from the mock community. In a similar manner, the estimation of 6% of contamination has to be explained in detail (species/proportions discarded or having been taken into account to calculate this percentage).

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human microbiome, Microbial genomics, Nanopore sequencing, Computational biology, Metagenomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.