Decision for round #1 : *Revision needed*
Revision of your manuscript

Dear authors,
Two reviewers have now evaluated your manuscript and given useful comments for improving its quality.
I suggest you take all the reviewers' comments into account for providing a revised version of the manuscript. More specifically, I would ask you to particularly take into account reviewer 2's comments regarding your experimental design (confounding factors between technologies and primers choice) and elaborate upon this in the revised version.
Sincerely yours,
Aymé Spor
*by **Aymé Spor**, 17 Jul 2023 09:58*
Manuscript: **https://doi.org/10.1101/2023.06.06.541006**
version: 1

(Author's replies in green)

Dear Editor,

We address sincere thanks to you and to both reviewers for their work on our manuscript and constructive suggestions, followed almost integrally.

We fully agree with reviewer #2 that primer effects cannot be disentangled from those of sequencing devices, a point that we should have stated clearly throughout the manuscript. However, this element does not fundamentally change the scope of the study, which was to provide evidence that sediment samples metabarcoded by long-reads on Nanopore have shown a similar bacterial community structure within samples than did the same samples metabarcoded by short-reads on Illumina. So we modified all statements in the manuscript that were abusively attributed only to the sequencing device (Nanopore or Illumina), and reformulated in "short-reads" or "long-reads" each time. Title and abstract have been changed in this perspective, and the difference in diversity for communities described by long-reads was not emphasized.

In the first draft, focus was made on taxa diversity, but actually it appeared that taxa exclusively detected by short- or long-reads did represent a much lower proportion of reads than their proportion among taxa. The 11 phyla exclusively detected by long-reads represented only 0.2% of the reads. Moreover, 84.7 and 98.8% of the short-reads were assigned strictly to the same species and genus, respectively, than those detected by long-reads. So new pieces of results have been added in Table 1 (colored lines), Figure 6 (stars), and were mentioned in the text and the abstract.

The authors

Review by anonymous reviewer 1, 30 Jun 2023 06:38

Dear Authors,

Thank you for the opportunity to review your manuscript. It shows detailed comparison of two sequencing approaches with potentially large impact on microbial alpha diversity of biological samples. Higher number of taxa obtained using long read sequencing is shown after rigorous analysis. On the other hand, similarity of major patterns of community composition between short and long read approaches is presented.

Although authors avoid giving strong recommendations on which method should be used, I think that presented results show valuable comparison which is useful for readers oriented on methodological papers.

I listed my line by line comments below. I noted two major issues from which one is about PCR cycling conditions (L152, L158, L173, L182) and the second is about input data into random forest analysis (L311). I ask authors to consider these and other comments below.
Based on this, I think that the current version of the manuscript needs minor revisions.

L22, L25 - genera instead of genus. DONE (L27, L34)

L27 - Here the statement can be stronger, I would omit "probably" since these are real reasons for discrepancies. DONE

L60 - please omit "works". DONE

L62 - If there is a length limit for PacBio, it is for sure longer. I suggest to avoid specific number (as it might be obsolete soon) and mention "tens of kbp" or similar. DONE

L106, L109 - please omit parentheses at the beginning of sentences. DONE

L139 - Please specify type of ZymoBIOMICS mock sample as there are more types on the manufacturer's website. DONE

L144 - Starting from this section but also in previous parts of the manuscript, the typographic corrections need to be applied widely. This includes 16S instead of 16s, dot as decimal point, en dashes where appropriate, English quotation marks, multiplication sign instead of x, etc. DONE

L152, L158 - Altogether 60 cycles of amplification after primary and secondary PCR seems like a lot of cycles. The authors want to avoid PCR biases and do triplicate PCRs (L145) which is certainly good. But then DNA goes through so many cycles which increase the chance for bias and contamination amplification. Could you please include a reference, if there is such an approach recommended?
We followed a standard protocol proposed by ONT (« PCR barcoding (96) amplicons SQK-LSK109 ») for the indexation of samples in a library, consisting in a first amplification of the marker and a second amplification that labels amplicons with a unique tag for each sample,

allowing multiplexed sequencing. This protocol also helps for increasing and standardizing the quantity of amplicons in all the samples. The Nextera protocol for indexing Illumina libraries is similar. Another protocol for indexing ONT library would be to ligate a nucleotidic tag to each sample, but is more expensive than PCR and does not standardize the library.

L163 - Please include full names of sequencing kits as they are written at manufacturer's webpage. DONE

L169 - I agree with usage of specific primers for Archaea, but I am missing an explanation of such an approach in Introduction or in Methods. Could you please include one sentence why archaeal primers were used in the case of Nanopore?
A sentence was added L205 ("However these primers were designed for bacteria and do not amplify archaea, unlike the primers pair used with Illumina, so a second marker was chosen for archaea (V1-V6 regions, ~1 kpb; Bahram et al. 2019; SSU1Ar F: TCCGGTTGATCCYGCBRG ; SSU1000Ar R: GGCCATGCAMYWCCTCTC).").

L173, L182 - Here amplicons for Nanopore went through 55 cycles which poses the same question if it is necessary to cycle so many times.
We followed the standard protocol recommended by ONT.

L173, L174 - unclear meaning of values in brackets, please clarify. DONE

L177 - diluted is maybe better then reduced. Replaced by "brought back to".

L188 - "protocol from Nanopore website" There is no need to specify from which website the protocol was downloaded. Alternatively, you can include link as proper reference. DONE
L195 - please check the number 1624, L257 and L479 mention different level of rarefaction.
DONE, the sentence was deleted from this section ; the correct number is 1582 reads for conventional rarefaction, as written in Results : Samples read coverage section.

L198 - PRJNA985243 checked and fastq files are available together with clear labelling of individual samples. OK

L202, L208 - the connection of ASVs and OTUs is a little bit confusing here. I understand that DADA2 was used to merge pair-end reads (L202) and maybe to correct errors. Individual sequences were then clustered at 97% threshold. If it is so, please clarify the paragraph. It was not ASV table which was clustered (L208) but rather individual sequences, right? DONE

L207 - please correct typo in kpb. DONE (kbp)

L208 - please consider to include vsearch version. DONE

L215-218 - The sentence needs clarification, its meaning is unclear. DONE

L230 - Please check, maybe Figure 4 was meant. DONE

L232 - Please specify core threshold. Is it >=50% in each sample, >=50% of all samples or something else? DONE

L240-L243 - The sentence is duplicated, please correct. DONE
Figure 2 - Please include description in figure caption that "once" means one flow-cell (L467) and "twice" means sequenced on two flow-cells (L468). DONE

L250 - I wonder if the manufacturer took into account the copy number of 16S genes of individual genomes present in the mock sample. Taxon with 2 copies of 16S rRNA will show higher relative abundance in final sequences than taxon with one copy. This was probably considered during mock preparation but it might be one of the explanations for changed proportions. However, I am aware that there is a nicely looking barplot with even distribution of taxa. Agree.

L261 - This is side note, but Table 1 partly duplicates information in Figure 3. That's right : Table 1 gives precise numbers and Figure 3 allows a rapid overview ; Table 1 can be moved in Supplemantary material.

L267 - I suggest to include information what is the mean abundance of 11 phyla detected exclusively in Nanopore data. This might provide an idea about the size of this Nanopore-detected subcommunity. DONE, size is very small (0.2%)

L270 - please omit only. DONE

L276 - please consider to reorder Figure 4 and Figure 3. In the current version, Fig 3 is referenced after Fig 4. DONE

L282 and L288 - information here is repeated, please correct. DONE

L286 - genera. DONE

L296 - L305 - For reader's reference, I suggest to include phylum names of individual orders in brackets. DONE

L309 - Does the "species rank" means that OTUs served as input into Mantel test? Yes. Please clarify. DONE

L311 - Genera and families are arbitrary groups and as such I am not sure if they can enter random forest. I suggest to test the same effects with OTUs which are exactly defined. At species level, the error rate of random forest model was 26.92% (21.15% at genus level).

L315 - The archaeal sentence sounds a little bit vague. I suggest to include at least information on how many phyla were detected as Nanopore-only. We decided to exclude archaea from the main text, because archaean long-reads were obtained with archaean specific primers and a dedicated flow-cell, which much enhanced the quantity of archaean reads and taxa with long-reads. All archaean results are presented in Supplementary Material.

L323 - L325 - nicely summarized output which applies also in this manuscript.

L342 - Please consider to add that another reason might be due to incomplete databases. This is written a few lines below.

L346 - What do you mean by maximum resolution? I feel that this sentence needs reformulating. DONE

L359 - L361 - I understand what was meant here but I feel that this sentence needs reformulating. DONE

L362-L364 - Nice key output of the study.

L365 - Ecology of Nanopore-only taxa can not be inferred based on the fact that the rest of core community was similar between Nanopore and Illumina. E.g. Nitrospinota detected by Nanopore might represent low-density nitrifiers with potentially high impact on N cycling in sediment. Corrected.

L368-L380 - The last paragraph seems out of context, please consider mentioning portability in Introduction if you prefer to keep it. I think that manuscript has same quality even without portability section. Right, the paragraph was moved to Introduction.

Review by anonymous reviewer 2, 11 Jul 2023 21:23

The present study compares short read sequencing with long read sequencing from ONT on environmental (marine) sediment samples. The authors conclude in this comparison that ONT works as good as Illumina with even covering more diversity. The articles writing is okay, and the findings are concisely presented. I think the study design as it is presented is however not correct, while the conclusions are partially valid (see below). I have one important methodological question and one important question related to the mock community.

Line 87-88: please give the respective references for RCA and UMI already in this sentence. DONE

Line 105: you may add https://doi.org/10.1093/femsec/fiac120 to the list. And I think there are also others that are more and more using it. DONE

Line 113: The Study design description is not quite accurate: how do the authors disentangle the effects of primers from the effects of sequencing technologies? This can't be done with this data. (except maybe with an in silico PCR and Illumina simulation on the Nanopore reads). Therefore, what the authors really compare were short amplicons with primer pair A with long amplicons with primer pair B. Since the study of Parada et al. 2015, we know that even a single nucleotide in one primer can have a tremendous effect on diversity estimates from sequencing. Here, the authors compare different primer sets, which renders a comparison of sequencing technologies not quite on

the point. Therefore, it is rather a feasibility study on long read sequencing with ONT that shows, that it produces similar results as established primers for short read sequencers. A direct comparison with numbers (alpha diversity estimates) is not advised, because it is like comparing apples with oranges. Therefore, I am afraid that the aim and the writing of the manuscript needs to be revised accordingly (and rather extensively). DONE It's totally right, we've understood this major ambiguity without formulating it as clearly. Thank you for clarifying it. So we have modified the title and the abstract, putting forward the comparability of bacterial community structure between the binomial made by primers and sequencing platform. We also modified the aim of the study in this way. In results, similarity tests of the community structures were moved before the phylogenetic diversity section. However we believe that specific results related to the primers sequences are also worth to be given in the abstract and text, because long-read primers are inherent to the Nanopore platform.

Method: LSK109 with 10% error rate; clustering at 97% will result in spurious OTUs, even with singletons across all samples removed LSK109 is today around 6% error rate, but the power of the length of long-reads may have a decisive effect on taxonomic assignment. The BLAST of a long read with 6 or even 10% error rate on a known cultivated taxa has evidenced it: the assignment is the same than those based on a short-read sequence. However, even if the high error rate would create artificial taxa, our study shows that this artifact does not really affect the structure of the community.

Line 221: I would advise to look into the new publications from Patrick Schloss (doi: https://doi.org/10.1101/2023.06.23.546313) considering this argument, in particular for the alpha diversity estimates, since much weight is put on it in the author's manuscript. Thank you for this new reference. Two points have weighted in our choice to use coverage-based rarefaction instead of conventional rarefaction : (1) conventional rarefaction is made randomly, inducing that different rarefactions on the same dataset will give slightly different analysis ; (2) our dataset showed a small number of samples with a reduced number of reads, conventional rarefaction would have wasted a lot of reads from the other samples if based on them.

Line 246. Relative proportions of the mock community are one thing, but not really that relevant since we are talking about compositional data. More important is the matching of OTU numbers with actual # of taxa in the Mock community and the detection of all taxa. I can imagine that there are vast differences between the amplicons. Please amend these missing results, even if they represent a weak point for Nanopore R9. We acknowledge that, but we also acknowledge the assumption that sequencing a mock community allows us to detect the threshold in relative abundances below which we can do a lower cut of OTUs in samples. That's why we did not present all the taxa detected in mocks. See the first paragraph of section "Community structures analysis" in Mat&Met : "Bacterial taxa known to be present in Ze samples were all above a relative abundance threshold of 1.8% for Illumina 16SV4-V5 and of 1.0% for ONT (Fig. 2), so relative abundances in all phyloseq objects were filtered above these thresholds." Moreover, given the strong sequencing effort for mocks, OTUs were filtered on a minimum of 50 reads/OTU in mocks before applying the threshold mentioned above.

Line 254: This comparison with percentage suggests that more species are better. This is however not the case. An accurate estimate of the taxa in a given sample is important. More OTUs may for

instance mean more artifacts, less true taxa. Indeed, we also believe that Nanopore may have a high false discovery rate for OTUs, but our protocol could not test it. Nevertheless, we don't agree that the text L254 would suggest that more species is better.

Line 266: Nanopore detected these 11 phyla or did the primer system detect these phyla. I would argue that the primer pair detected it. ONT is just the tool to read out these sequences. I would suggest revising this terminology by replacing "Nanopore" with e.g. "long amplicons", which is more objective. This is right, we've replaced it when possible. Our results intend to promote Nanopore because it's more affordable for small labs like us, and we'd like to tell it to the scientific community.

Line 386: I consider OTU/ASV tables with taxonomy classifications, read abundances per sample, and one representative FASTA as mandatory supplemental item. Please amend as an annotated .csv file. OTU tables are available on the github repo.

Thanks!