Nanopore long-reads reveal fine structure of prokaryotic communities in mangrove sediments, like Illumina short-reads but with twice more taxa

Alice LEMOINNE[1,†,*], Guillaume DIRBERG[1], Myriam GEORGES[2] & Tony ROBINET[1,*]

[1] Laboratoire de Biologie des Organismes et Ecosystèmes Aquatiques (Borea), Muséum National d'Histoire Naturelle, UMR 8067, CNRS, MNHN, IRD, SU, UCN, UA, Station marine de Concarneau, Quai de la Croix, 29900 Concarneau, France

[2] UMS 2700 2AD - Acquisition et Analyse de Données pour l'Histoire naturelle, Station marine de Concarneau, Quai de la Croix, 29900 Concarneau, France

[†] present address: UMR CARRTEL, INRAe, 75 bis Avenue de Corzent, 74200 Thonon les Bains, France

** correspondence:* alice.lemoinne@inrae.com, tony.robinet@mnhn.fr

Abstract

Following the development of high-throughput sequencers, environmental prokaryotic communities can be described by metabarcoding with genetic markers on the 16s domain. However, usual short-read sequencing encounters a limitation in phylogenetic coverage and taxonomic resolution, due to the primers choice and read length. On these critical points, nanopore sequencing, a rising technology, suitable for long-read metabarcoding, was much undervalued because of its relatively higher error rate per read. Here we compared the prokaryotic community structure in samples obtained by short-read metabarcoding on 16sV4-V5 marker (*ca*. 0.4kbp) analyzed by sequencing-by-synthesis (Illumina dye sequencing, MiSeq), with those obtained by nanopore long-read metabarcoding on bacterial nearly complete 16s (*ca*. 1.5 kbp, Oxford Nanopore, MinION, R9.2), i.e. a mock community and 52 sediment samples from two contrasted mangrove sites. Nanopore and Illumina retrieved all the bacterial genus from the mock, although both showing similar deviations from the awaited proportions. From the sediment samples, with a coverage-based rarefaction of reads and after singleton filtering, Illumina and Nanopore recorded 34.7% and 35.4% of unknown OTUs, respectively. Nanopore detected 92.2% of the 309 families detected by Illumina, 87.7% of the 448 genus, and recorded 973 additional taxa not detected by Illumina, among which 91.7% were identified to the genus rank. In spite of primer specificities and read length, probably accountable for these discrepancies, co-inertia and Procrustean tests showed that community structures were significantly similar between technologies, showing both a marked contrast between sites and a coherent sea-land orientation within sites. [242 words]

Keywords: microbial metabarcoding, environmental DNA, methods, primers, diversity

Zenodo: fastq.gz Illumina F+R, Nanopore, sur 54 samples + mock + blancs extr. et PCR
BiorXiv: BIORXIV/2023/541006 - Version 1
GitHub: tables otus et sampledata, scripts

Introduction

The composition and structure of microbial communities are nowadays studied in environmental samples through culture-independent methods, based on nucleic acid sequencing, either in bulk DNA extracted from environmental samples (*metagenomic*) or only for DNA markers of interest (gene fragments), amplified from environmental samples before they are sequenced (*metabarcoding*). The metagenomic approach is exempt from amplification bias inherent to metabarcoding (marker specificities, PCR-induced stochasticity) and can produce metagenome assembled genomes (MAGs), but it still faces technical and cost challenges (Taş et al. 2021). The metabarcoding approach remains more widely used, much cheaper, but amplification bias are recurrent : (i) primers choice is crucial and constrained by the maximum size of inserts for second-generation sequencers (400bp for Ion Torrent PGM, 550bp for Illumina MiSeq, up to 800 bp for Roche 454, Luo et al. 2012) ; (ii) taxa diversity can be overestimated, because of the non-targeted DNA present in the sample (i.e. DNA from the eukaryotic digestive tracts, or extracellular “relic” DNA, Carini et al. 2017) and also because of the ribosomal DNA polymorphism, hidden in individual genomes (intragenomic variability in the number of duplicates of ribosomal operon, differences in allelic variants between copies, Pereira et al. 2020) ; and (iii) relative abundances of reads per taxa are somehow inaccurate, compared to real abundances in the mock samples, a consequence of PCR stochasticity and primers specificity.

In high-throughput sequencing (HTS) metabarcoding, primer choice is known to be crucial for taxa resolution, phylogenetic coverage and sensitivity to fine community structure. For prokaryotes, none of all the primer pairs that amplifies markers at a convenient size for short-reads strategies (> 550bp for Illumina) can give a complete phylogenetic coverage. Primers spanning over more than one 16s V-region are often preferred, because they improve taxonomic resolution. However, each one of these combinations (V1-V2, V3-V4, V4-V5, V6-V8, V7-V9, etc.) showed bias in phylogenetic coverage (Abellan-Schneyder et al. 2021). The 412 bp V4-V5 marker (515F-926R, Parada, Needham, et Fuhrman 2016) covers more broadly the prokaryotic domains (bacteria and archaea), whereas the 438 bp V6-V8 (B969F-BA1406R, Willis, Desai, et LaRoche 2019) amplifies additional bacterial clades, leading some authors to consider as a best method to combine several short regions along the prokaryotic 16s to minimize these bias (Fuks et al. 2018). However, the multiplication of marker standards for bacteria and archaea also plays against works intercomparability.

Third-generation DNA sequencers marked a significant progress for metabarcoding studies, in the fact that the marker size was no longer a technical limitation (up to 30kb for PacBio Sequel II, and no theoretical limit for Nanopore devices), and one can target much more binding sites for primers, improving considerably taxonomic resolution and phylogenetic coverage (Furneaux et al. 2021; Tedersoo et al. 2021; Eshghi Sahraei et al. 2022).

These long-read high-throughput sequencers have been first implemented for sequencing markers from cultivated organisms (Schlaeppi et al. 2016; Loit et al. 2019; Maestri et al. 2019). Long-reads environmental metabarcoding has been usually performed on PacBio sequencers, because the Single-Molecule Real-Time (SMRT) technology offers a read quality similar to those of short-reads platforms. Long-reads metabarcoding is mostly used for taxonomic groups in which short-reads are too short for a descent assignment, like micro-eukaryotes and specially fungi (Tedersoo, Tooming-Klunderud, and Anslan 2018; Furneaux et al. 2021; Kolaříková et al. 2021; Eshghi Sahraei et al. 2022; Gueidan and Li 2022), but also a few bacterial phyla (Katiraei et al. 2022). Despite several published works showed the possibility to use

Nanopore sequencing for environmental or food metabarcoding, by sequencing mock communities of known composition (Benítez-Páez, Portune, and Sanz 2016; Davidov et al. 2020; Urban et al. 2021; Toxqui Rodríguez, Vanhollebeke, and Derycke 2023) or by comparing it with an Illumina library sequenced concurrently (J. Shin et al. 2016; H. Shin et al. 2018) , the great majority of works that we found in literature did not use the Nanopore platform for environmental metabarcoding.

Raw reads accuracy are similar for PacBio (88-90%) and Nanopore (95-98% on the R9 flow-cells, above 99% for R10.4), nonetheless, the fact that PacBio circular consensus sequence technology (CCS) can align several reads of the same amplicon brings it to an accuracy of >99.9% at 10-fold consensus (Tedersoo et al. 2021). The first long-read third-generation sequencer acknowledged to be suitable for metabarcoding was PacBio Sequel II on fungal complete rRNA operon (ca. 3000 bp, Tedersoo, Tooming-Klunderud, et Anslan 2018). Despite its error rate being slightly higher than Illumina, the PacBio long-read sequencing allowed a much better taxonomic resolution, due to the joint powers of ITS1-ITS2 and SSU-LSU flanking regions on the same amplicon.

Promising attempts were made to reach a satisfactory accuracy with Nanopore, by mimicking PacBio with a rolling circle amplification (RCA) or by flanking, at the two first steps of PCR, each single amplicon with a unique molecular identifier (UMI). RCA and UMI methods produce a consensus error rate of 0.7% (coverage > 45x) and 0.01% (> 25x) respectively, offering a quality similar to PacBio or Illumina standards (Baloğlu et al. 2021; Karst et al. 2021). The consensus, compared with BLAST (Camacho et al. 2009) to reference sequences of a curated database, could be assigned more accurately to a taxa than standard short markers do (reviewed by Kerkhof 2021). However, lab and downstream bioinformatic workflows are quite complex to implement for ecology scientists, requiring a higher technicity in library preparations and in downstream bioinformatics than directly sequencing amplicons from environmental samples, as we tested it without success. To date, no environmental metabarcoding based on RCA or UMI protocols has been published.

In community ecology, Nanopore was initially used for barcoding individuals with long-reads (Maestri et al. 2019), but quickly metabarcoding appeared with Nanopore sequencing alone, to detect pathogen bacterial strains, mostly by a metagenomic approach (Brown et al. 2017; Charalampous 2019; Cuscó et al. 2019), or eukaryotic communities on the more or less complete rRNA operon (H. Lu, Giordano, and Ning 2016; Toxqui Rodríguez, Vanhollebeke, and Derycke 2023). For bacterial communities, studies with a metabarcoding workflow on environmental samples and relying only on Nanopore MinION, aimed at characterizing mouse gut or human respiratory bacteriomes (J. Shin et al. 2016; Ibironke et al. 2020), bacteria associated with algae or plastic debris at sea (H. Shin et al. 2018; Davidov et al. 2020; van der Loos et al. 2021), pathogenic bacteria in food (Planý et al. 2023), or pelagic bacteriomes in freshwaters (Urban et al. 2021). In all studies we found, Nanopore was used alone, except for two. (Loit et al. 2019) compared it with PacBio CCS for detecting fungal pathogens in plants, concluding that “MinION could be used for rapid and accurate identification of dominant pathogenic organisms and other associated organisms from plant tissues following both amplicon-based and PCR-free metagenomics approaches”. (J. Lu et al. 2022) characterized mycobiomes of fungal isolates and environmental samples by sequencing in parallel the full rRNA operon on MinION and the shorter ITS2 on Illumina HiSeq. They concluded that “ITS2 sequencing [was] more biased than full operon sequencing”. To date, no published work has compared Nanopore to Illumina bacterial metabarcoding on the same environmental samples, which was attempted here.

Here we propose to test the efficiency of the Nanopore MinION device on environmental samples of marine sediments, by sequencing in parallel Nanopore and Illumina libraries, both made of amplicons from the same DNA extracts and with a similar protocol of library preparation, i.e. without RCA or UMI, adressing two questions : (i) is the beta diversity of prokaryotic communities similar between sequencing strategies, (ii) do the sequencing strategies conserve the differences at a gross-scale in bacterial communities between two mangrove sites, and at a fine for the sea-land orientation of intertidal communities within sites?

Materials and methods

Sampling sites and sample collection

In June 2019, 2 sites were selected in the mangrove of Guadeloupe Island, at 6 km of distance each other, for their a priori difference in the level of direct and indirect human pressures (Fig.1a-b) : the impacted “Rivière salée” site was located on the foreshore of a salty river, close to the city of Pointe-à-Pitre, to its dump and its airport (latitude -61,5469; longitude 16,2594) ; the less-impacted “Babin” site was located in the Ramsar zone (protected area) close to coral reefs (latitude -61,5294 ; longitude 16,3388).

A total of 54 samples of surface sediment were collected, by 3 lines of 3 points each respectively in each site, each line separated by 3 m to the neighboring line ; points were separated each other by 12.5 m within a line. Each point was composed of 3 samples (biological replicates A, B and C), analyzed in the workflow separately (Fig.1c). The line close to the sea was the “seaward line”, those close to the inland mangrove was the “landward line”, and the “middle line” was in between. Therefore, each line showed a different time of marine immersion per day. Each replicate was sampled with a sterile syringe and appropriated microbiological precautions, stored in a 50ml Falcon tube, freezed a couple of hours after sampling and preserved at –20° C.

DNA extraction

Samples were freeze-dried and crushed to powder in a mortar, carefully clean with an alcoholic tissue between each sample. Total genomic DNA from 50mg of dried samples and a standard microbial community (zymoBIOMICS), here named “*Ze*”, were extracted using the NucleoSpin Soil kit (Macherey-Nagel) with a final elution volume of 50 µl following the manufacturer instructions. After this DNA extraction of samples and Ze, nucleic acid yield and purity were checked using a Nanodrop spectrophotometer (Thermo Fisher Scientific) and the concentration of each sample was equalized to final concentration of 10ng/µl on a PCR plate of 96 wells.

Illumina library

In order to limit PCR biases, the first round of PCR consisted in 3 PCR replicates per sample, targeting the DNA coding for the V4-V5 hypervariable region of 16S RNA ribosomal with degenerate primers (Parada, Needham, and Fuhrman 2016) : 515F (GTGYCAGCMGCCGCGGTAA) and 926R

(CCGYCAATTYMTTTRAGTTT). Each primer was flanked in its 5'-end by a nucleotide sequence used for indexing at a later step, according to Nag et al. 2017. At this stage, 2 PCR blanks were done with water instead of extracted DNA. Each 12,5 µl reaction mix contained 1 µl of DNA (~10ng.µl-1), 0,25 µl of forward primer, 0,25 µl of reverse primer (10nM), 6,25µl of 2X Promega Green Master mix G2, 4,25µl of milliQ water. The PCR cycles consisted of of initial denaturing for 2 min at 94°C, followed by 30 cycles (denaturation 30 s at 94°C, hybridization 30 s at 51°C, elongation 45 s at 72 °C) and a final elongation during 5 min at 72°C. First PCR products were verified by electrophoresis on 1% agarose gel, re-amplified if negative until they were positive. Each PCR triplicate was pooled into one before the indexing PCR. Indexation PCR was realized in a 27.5 µl reaction mix containing 2 µl of first PCR products, 5 µl of reverse and forward index, 12,5µl of NEB Q5 2X mix and 8µl of milliQ water. This second PCR consisted of a initial denaturing for 30s at 98°C, followed by 30 cycles (denaturation 20 s at 98°C, hybridization 20 s at 60 °C, elongation 10 s at 72°C) and final elongation 10 s at 72°C. At this stage, one PCR blank was added with water instead of first PCR products. All indexed samples were pooled into a single low-bind tube and purified with magnetic beads (Nucleomag, Macherey Nagel 1:1 ratio). Size range of final PCR products was verified by electrophoresis (Agilent BioAnalyzer High-sensitivity), then pooled in a final library, and sequenced on an Illumina MiSeq (one v3 kit 600 cycles and one nano kit 500 cycles for resequencing) in the Concarneau marine station (MNHN) to product demultiplexed output fastq files.

Nanopore library

The same DNA extracts were processed in parallel for Nanopore sequencing on complete 16s, with the following 16s markers : V1-V9 regions (nearly complete 16s for bacteria, ~1.45 kbp; Weisburg et al. 1991; 27F:AGAGTTTGATCMTGGCTCAG ; 1492R:TACGGYTACCTTGTTACGACTT) and V1-V6 regions (for archaea, ~1 kpb; Bahram et al. 2019; SSU1Ar F: TCCGGTTGATCCYGCBRG ; SSU1000Ar R: GGCCATGCAMYWCCTCTC). PCRs were performed in 3 small-volume replicates of 12,5 µl each, containing 6,25µl of LongAmp Taq 2x Master Mix (NEB), 4,25µl of milliQ water, 1 µl of DNA (~10ng.µl-1), 0,25 µl of forward primer, 0,25 µl of reverse primer (10nM each). PCR cycles consisted of initial denaturing for 3 min at 94 °C (95°C), followed by 30 (32) cycles composed of denaturation for 30 s at 94 °C (95°C), hybridization for 30 s at 51 °C (55°C), and elongation for 45 s at 65 °C and final elongation for 10 min at 65°C. All first PCR products were verified by agarose gel electrophoresis, re-amplified if negative until they were positive, and positive triplicates were pooled into one before the indexation PCR. Concentrations were measured by the Qubit fluorometer (dsDNA BR kit) and reduced to a concentration of 1ng/µl. Indexation PCR was realized according to the Nanopore « PCR barcoding (96) amplicons (SQK-LSK109) » manufacture, by indexing each sample (environmental, blank and Ze) with Nanopore purified barcodes (BC) in 50µl reaction mix containing 20 µl of the first PCR product, 1 µl of reverse and forward BC, 25µl of mix LongAmp Taq 2x Master Mix (NEB) and 4µl of milliQ water. Indexation PCR cycles consisted of initial denaturing for 30s at 95 °C, followed by 25 cycles composed of denaturation for 30 s at 95 °C, hybridization for 1min at 62 °C, and elongation for 1 min at 65 °C and final elongation for 1min50s at 65°C. Indexed amplicons were pooled into one tube per primer/marker and purified with magnetic beads (Nucleomag Macherey Nagel, 1:0.8 ratio). Indexed and purified products were verified on agarose gel electrophoresis. DNA concentration was measured by phospho-luminescence (Qubit), then diluted in order to have 1µg of DNA into 47µl of water. Final ligation of Nanopore sequencing adapters was done following the "SQK-

LSK109 with EXP-PBC096" protocol from Nanopore website. 16S V1-V9 library was sequenced on two R9.4.1 MinION flow cells (half of the samples for each), 16S V1-V6 on a third one. Flow cells were loaded on MinION Mk-1C and sequenced for approximately 48 h, until no further sequencing reads could be collected (quality failed). Fast5 files were basecalled and demultiplexed using Guppy 6.4.2 high-accuracy model on a local GPU (Nvidia Quadro K4000) and DNA sequence reads were output with >Q10 flag, as fastq files.

For Illumina and Nanopore, indexes with less than 1500 reads were re-sequenced, to reach a minimum of 1624 reads (16SV4-V5 Illumina) on which rarefaction of reads per sample was established for comparing both technologies. Further analyses only with Nanopore reads were based on a rarefaction at 5500 reads.

Sequence data are submitted to SRA database and are available with BioProject accession number XXXXXXXX.

Processing of raw reads

Fastq files from Illumina were filtered with R package DADA2 v 1.16.0 (Callahan et al. 2016). Reads 1 and 2 were filtered using the filterAndTrim function (minLen=200, matchIDs=TRUE, maxN=0, maxEE=c(3,3)), then merged to unique sequences (ASVs) with at least 12 overlapping nucleotides between R1 and R2. Chimeric sequences were removed using the *removeBimeraDenovo* function. A matrix of 16sV4-V5 ASVs per sample was obtained and processed by Qiime2 tools, after 16Sv4 ASVs were extracted from fasta files containing sequences from other primers (18SV9 and ITS2, not presented here). Nanopore fastq sequences (>Q10) were filtered with Nanofilt : all reads shorter than 1.4 kbp and longer than 1.6 kbp for 16S V1-V9 and 900 and 1.1kpb for 16 V1-V6 were removed. Then, each ASV table (Illumina 16sV4-V5, Nanopore 16s V1-V9 bacteria and 16sV1-V6 archaea) was clustered into an OTU table with 97% of similarity using the Vsearch tool. OTUs were taxonomically assigned with a trained Qiime2 classifier, inferring to the SILVA NR 99 reference database v138.1 (Quast et al. 2013), formatted for each marker.

Community structures analysis

Chloroplastic, mitochondrial and eukaryotic assignments, contaminants detected from blanks and singletons (OTUs with only one read in all samples) were removed from OTU tables. Tables of filtered OTU read abundances, OTU taxonomy and sample data were imported to make *phyloseq* objects in R, one for each marker (R package *phyloseq*, McMurdie and Holmes 2013). *Ze* samples were able to detect a reliable relative abundance threshold of 1.8% for Illumina 16sV4-V5 and of 1.0% for ONT on the mock bacterial community (Fig. 2), so relative abundances in *phyloseq* objects were filtered in this way. OTU were filtered on a minimum of 50 reads/OTU for Ze.

The prokaryotic community structure of environmental samples depends tightly on the read number in each sample. The conventional rarefaction consists in randomly depleting reads in each sample, until all samples reach the number of reads of the poorest one (Simberloff 1972). This method is known to have major bias: non-reproducibility since reads are removed randomly, and alteration of community structures, specially for rare species (Coddington et al. 2009). In soil or sediment microbiotas, sample OTU richnesses depend

strongly on sample size, therefore we opted for the rarefaction method developed by Chao and Jost (2012), consisting in comparing samples of equal completeness (equal *coverage*), not of equal size. “When samples are standardized by their coverage (a measure of sample completeness [...]) instead of by their size, the estimated richnesses approximately satisfy a replication principle, which is an essential property for characterizing diversity” (Chao and Jost 2012). This coverage-based rarefaction was used by the function *phyloseq_coverage_raref* (R package *metagMisc*, Mikryukov 2019).

Analyses were carried out on filtered OTU tables after coverage-based rarefaction, except for Figure 3, in which both rarefaction methods are shown. In sediment samples, core members were identified by their prevalence in samples (≥50%). For exploring dissimilarities between data sets, a Principal Coordinate Analysis (PCoA, from *phyloseq ordinate* function, equivalent to MDS - Metric Multidimensional Scaling) was performed on matrices of Bray-Curtis distances between communities. To identify the most contributing OTUs to the different parts of the communities, a Principal Component Analysis (PCA, from ade4 package *dudi.pca* function) was performed on relative abundances. In order to assess the similarity of community structures described by both sequencing methods, a Procrustes analysis was carried out on their respective PCoA scores, with *procrustes* and *protest* functions (R package *vegan*, Oksanen et al. 2010). In parallel, a co-inertia analysis on PCA two first components was done, with *coinertia* and *RV.rtest* (999 permutations) from *ade4*. A Mantel permutation test was performed on two matrices of Bray-Curtis distances, for Illumina and Nanopore bacterial communities (Pearson method, 999 permutations). A Mantel test was performed with *vegan*, on the Pearson correlation between Bray-Curtis distance matrices for OTU tables obtained by Illumina and Nanopore. Classification trees were used to characterize the genus and families contributing the most to the [site x (sea-land orientation] effect in each dataset by the R package *randomForest* (Liaw and Wiener 2002). The circular diagram showing archaean OTUs the most contributing to PCA structure in each sample was obtained with the *ord_plot_iris* function from *microViz* R package (Barnett 2023).

Results

Mock Community

The 8 prokaryotic taxa of the Ze community have all been found with each sequencing method (Fig. 2). However, the proportions of reads assigned at each taxa were not those initially introduced (12% for each prokaryote). The Illumina and the ONT methods both overestimated some genus *(Lactobacillus*, *Limosilactobacillus, Salmonella*) and underestimated some others (*Escherichia*, *Listeria, Enterococcus*). Both methods found undetermined reads (5.1% for Illumina, 11.0 and 10.4% for Nanopore_1 and Nanopore_2, respectively). The re-sequencing of Ze on two different Nanopore flow-cells showed a stability in the results.

Samples read coverage

With conventional read rarefaction, for bacteria only, all samples were standardized at 1582 reads for both sequencers, resulting in a total of 570 (16sV4-V5) and 967 (full-length 16s) bacterial species, so in proportion full-length 16s (Nanopore) counted 170% of the species detected by 16sV4-V5 (Illumina). With coverage-based rarefaction, Illumina samples presented 749 species for 2609 reads in average (min 1338, max 4991), Nanopore samples 1495 species for 5108 reads in average (min 4451, max 6019, Table 1). In proportion with this rarefaction method, Nanopore counted 200% of the species detected by Illumina. For the rest of this section, only results obtained by the coverage-based rarefaction method are presented.

Phylogenetic diversity

Bacteria. Over the 56 bacterial phyla detected in total, 54 were detected by Nanopore and 45 by Illumina (Table 1). At high taxonomic levels, Illumina 16sV4-V5 and Nanopore full 16s were approximately ⅘ alike for phyla, Nanopore detected 11 exclusive phyla over a total of 54 for this platform (20% of exclusive among those detected by Nanopore), when Illumina only had two (4.4%). The 11 phyla only detected by Nanopore were *Acetothermia*, *WS2*, *LCP-89*, *WOR-1*, *Armatimonadota*, *Margulisbacteria*, *Nitrospinota*, *Fermentibacterota*, *Methylomirabilota*, *Caldatribacteriota*, *WPS-2*, whereas the only 2 only detected by Illumina were *Cloacimonadota* and *CK-2C2-2,* with a coverage-based rarefaction (Fig. 4a). At lower taxonomic levels, 92.2% and 87.7% of respectively the family and bacterial genus detected by Illumina 16sV4-V5 were detected by Nanopore full-length 16s. Nanopore detected twice more species than Illumina, with only 34.9% of the species and 50.1% of the genus detected shared with Illumina. The trend that Nanopore detected almost all Illumina taxa *plus* a certain number of Nanopore original taxa decreased with lowering taxonomic ranks (Fig. 3a-b).

All the 54 Nanopore-detected phyla were more diversified based on full-length 16s, but four : *NB1-j*, *SAR324*, *Dadabacteria* and *Hydrogendentes*. The most diversified phylum, the *Proteobacteria*, presented more than 4 times more species with full-length 16s than with 16sV4-V5. Overall, communities described by the two sequencers were phylogenetically very similar when considering shared taxa at the family and genus level (92.2% and 87.7% of taxa similarity for Illumina *vs.* Nanopore, respectively) and also the composition of core microbiotas for shared taxa (100% of Illumina core-phyla were also core-phyla for Nanopore).

11.7% of the Nanopore reads (bacterial 16s) were unassigned at the phylum level, versus 0.36% for Illumina. 53.1% of the Nanopore reads unassigned at the species level (35.4% of total Nanopore bacterial OTUs), versus 46.0% for Illumina (34.7% of total Illumina bacterial OTUs, Table 1). For shared genus, the unassigned reads were much lower for Illumina (5.8%) than Nanopore (35.5%). All core-phyla detected by Illumina were also parts of core-phyla detected by Nanopore, whatever the rarefaction method used (Fig. 4a-b).

Archaea. With 16sV4-V5 degenerated primers on a single flow-cell, Illumina read coverage for archaea (mean 193 reads/sample, min 29 - max 620) was much lower than those of Nanopore for archaea (mean 4817 reads/sample, min 2384 - max 6701), the latter with specific archaean 16s primers and a dedicated

flow-cell. Therefore Illumina archaean taxa are just mentioned here, but not interpreted. Nanopore detected 171 archaean OTUs in 11 phyla, almost all belonging to core communities in samples (Fig. 4b ; Table 1).

Variations in community structures

Community composition and multivariate analyses showed that both technologies detected a marked difference between bacterial communities from Babin and Rivière salée sites, but also their fine-scale orientation, from sea- to land-oriented samples. Bacterial communities sequenced by Illumina and Nanopore described the same global patterns, i.e. a preponderance of Pirellulales in Rivière salée, of Pseudomonadales and Nitrosopumilales in Babin, separating clearly the two sites in ordination (Fig. 5a-b). Babin showed the most structured community along the tidal gradient, with the presence of Pseudomonadales in seaward samples and of Bacteroidales in landward samples. Biological replicates were relatively close to each other in the PCoAs (Fig. 6a-b), but Bray-Curtis dissimilarity indexes of communities within replicates were always higher for Illumina than for Nanopore, either for Babin or Rivière salée (Fig. 6c, anova $p<0.001$).

The Procrustes analysis of the two first axes of multivariates showed a significantly strong similarity between structures drawn by Illumina and Nanopore (Fig. 6d-e, $p<0.001$), confirmed by a co-inertia analysis on PCA's two first axes ($p<0.001$). The Mantel test indicated a significant correlation coefficient of 0.7248 ($p<0.001$) between the Bray-Curtis dissimilarity matrices obtained from Nanopore and Illumina communities at species rank. In order to point out the similarity of taxa contributing to the [site x (sea-land orientation] effect, classification trees were made by a random forest approach on the 393 genus and 285 families shared between Illumina and Nanopore. Models found 48% of similarity among the top-100 contributing genus and 63% among the top-100 contributing families between sequencers. However, taxa contributing in the same way to the [site x (sea-land orientation] effect were scarce (Fig. S1, Table S1). Archaean communities described with Nanopore specific primers followed roughly the structure obtained with bacteria (Fig. S2).

Discussion

In this study, bacteria (and secondarily archaea) were amplified on their rRNA gene by 16sV4-V5 and full-length 16s primers from the same DNA extractions of environmental samples, sequenced on Illumina and Nanopore respectively, and assigned on the same database of reference sequences (Silva 138.1 SSU LR99). Filtered and standardized with a coverage-based rarefaction, the bacterial communities described by both sequencing tools were similar in their coarse structure (site effect) and fine structure (sea-land orientation), with nevertheless a couple of constant differences, already noticed with short-read *vs*. long-read sequencing on PacBio (Katiraei et al. 2022): (i) communities described by full-length 16s were twice more species-diversified than those described by 16sV4-V5, (ii) and abundances of OTUs based on long-reads were slightly less variable within biological replicates than those based on short-reads. These differences reflected probably more a direct effect of read length than a sequencing-platform effect. This work suggests that Nanopore long-read can be used for metabarcoding environmental samples, with the advantage of a lower cost and field-lab portability.

Long-reads outperformed short-reads for taxonomic diversity

Katiraei et al. (2022) sequenced full 16s amplicons, on a PacBio system, and extracted afterward in silico the 16sV4 fragments. In-silico-extracted V4 dataset had approximately half of the read count per sample, compared to those of the full-length 16s PacBio dataset, indicating that a significant proportion of the taxa that were identified by full-length 16s were not detected by extracting the V4-region from the same initial sequences. In this way, the length of the 16s fragment can modify the taxonomic assignment, a longer fragment increasing the diversity of taxa assigned. Our study confirmed that there were much more taxa detected by full-length 16s than by 16sV4-V5, but also that a small proportion of taxa sequenced with 16sV4-V5 were not detected with full-length 16s dataset (30.3% of the species, 13.3% of the genus).

When considering non-shared taxa, the present study illustrated the power of a longer bacterial 16s rRNA, compared to a restricted 16s V-region, incidentally acknowledged to have the most appropriate cover for bacteria and archaea among short-reads primers (Parada, Needham, and Fuhrman 2016; Walters et al. 2016; Willis, Desai, and LaRoche 2019). Taxa assignment rates were lower at species level whatever the read length, probably due to pseudogenes and intra-genome 16s polymorphism (Pei et al. 2010; Větrovský and Baldrian 2013), impossible to evaluate with our approach.

Our study on marine sediment samples could not provide evidence that full-length 16s improved the taxonomic assignment, as it was done with human gut microbial communities (Jeong et al. 2021; Matsuo et al. 2021). However, genus level is considered as the maximum resolution of 16S sequencing, so a correct platform comparison should start from genus toward higher levels. The fact that read assignment was always lower for Nanopore-exclusive taxa probably reflects more the fact that mangrove sediments contain a high diversity of uncultivated microbes with presently unavailable full-length 16s in reference databases, than a lower sequencing accuracy of Nanopore (and therefore a sequencing-platform effect).

Similar site- and sea-orientation patterns, based on Nanopore and Illumina

Coarse and fine spatial structures were overall significantly similar, since the site effect and the sea-land orientation were conserved in ordinations.

Differences in abundances for the same taxa were obvious in the structure of mock communities, i.e. coming from the same DNA extraction but followed by separate amplification on different primers, different library preparation and sequencing. This discrepancy is typical and outlines the semi-quantitative trait of any microbial HTS sequencing. However, all qualitative elements (beta-diversity) of mocks were preserved, allowing us to extend this assumption to communities described from environmental samples processed with the same workflow as for the mock. This assumption may explain differences observed in top-20 more abundant bacterial orders, and is reinforced by the relative orientation of samples, preserved between the two sequencing workflows on the same ordination.

On the other hand, it is noteworthy that Nanopore communities contained twice more species than Illumina and this did not change the overall structure of ordinations, providing evidence that core-communities in both sequencing strategies were congruent and that additional taxa detected by Nanopore behaved ecologically like those shared with Illumina. In an other perspective, 16sV4-V5 Illumina's communities, albeit reduced, were sufficient and contained the smallest share of taxa needed to correctly describe the assemblages at play.

Toward a better portablity of metabarcoding

At the time of this study, there was no sequencing platform in Guadeloupe Island, where mangrove samples have been collected. The estimated cost of 1 Gb PacBio sequencing (17€) was lower than Illumina NovaSeq (44€) and MiSeq (56€), but the accessibility to a PacBio sequencer was difficult for this remote place, because of the instrument cost (650 k€ for a PacBio Sequel II) and technicity. Today, the MinION device of Oxford Nanopore Technologies is accessible for 900€, the estimated cost for 1Gb is about 12€, and its smartphone size allows scientists to use it as a field lab device. The portability of the MinIon device is advantageous for molecular ecology scientists located far away from a research center, opening possibilities for studying microbial communities from a field lab, i.e. equiped with usual devices for DNA extraction (mortar, mini-centrifuge, spectrophotometer for DNA drops), PCR (freezer, thermocycler, electrophoresis tank, UV table, ultra-pure water), and libraries making (DNA fluorometer, DNA dryer). Such a field lab is affordable and quite simple to set up for molecular ecologists in remote places or for proposing environmental metabarcoding in the frame of engineering consultancy.

381 # References cited

Abellan-Schneyder, Isabel, Monica S. Matchado, Sandra Reitmeier, Alina Sommer, Zeno Sewald, Jan Baumbach, Markus List, and Klaus Neuhaus. 2021. 'Primer, Pipelines, Parameters: Issues in 16S RRNA Gene Sequencing'. Edited by Susannah Green Tringe. *MSphere* 6 (1): e01202-20. https://doi.org/10.1128/mSphere.01202-20.

Bahram, Mohammad, Sten Anslan, Falk Hildebrand, Peer Bork, and Leho Tedersoo. 2019. 'Newly Designed 16S RRNA Metabarcoding Primers Amplify Diverse and Novel Archaeal Taxa from the Environment'. *Environmental Microbiology Reports* 11 (4): 487–94. https://doi.org/10.1111/1758-2229.12684.

Baloğlu, Bilgenur, Zhewei Chen, Vasco Elbrecht, Thomas Braukmann, Shanna MacDonald, and Dirk Steinke. 2021. 'A Workflow for Accurate Metabarcoding Using Nanopore MinION Sequencing'. Edited by Erica Leder. *Methods in Ecology and Evolution* 12 (5): 794–804. https://doi.org/10.1111/2041-210X.13561.

Barnett, David. (2020) 2023. 'MicroViz'. R. https://github.com/david-barnett/microViz.

Benítez-Páez, Alfonso, Kevin J Portune, and Yolanda Sanz. 2016. 'Species-Level Resolution of 16S RRNA Gene Amplicons Sequenced through the MinION™ Portable Nanopore Sequencer'. *GigaScience* 5 (1): s13742-016-0111-z. https://doi.org/10.1186/s13742-016-0111-z.

Brown, Bonnie L., Mick Watson, Samuel S. Minot, Maria C. Rivera, and Rima B. Franklin. 2017. 'MinION™ Nanopore Sequencing of Environmental Metagenomes: A Synthetic Approach'. *GigaScience* 6 (3). https://doi.org/10.1093/gigascience/gix007.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. 'DADA2: High-Resolution Sample Inference from Illumina Amplicon Data'. *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. 'BLAST+: Architecture and Applications'. *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Carini, Paul, Patrick J. Marsden, Jonathan W. Leff, Emily E. Morgan, Michael S. Strickland, and Noah Fierer. 2017. 'Relic DNA Is Abundant in Soil and Obscures Estimates of Soil Microbial Diversity'. *Nature Microbiology* 2 (3): 16242. https://doi.org/10.1038/nmicrobiol.2016.242.

Chao, Anne, and Lou Jost. 2012. 'Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size'. *Ecology* 93 (12): 2533–47. https://doi.org/10.1890/11-1952.1.

Charalampous, Themoula. 2019. 'Nanopore Metagenomics Enables Rapid Clinical Diagnosis of Bacterial Lower Respiratory Infection'. *Nature Biotechnology* 37: 14.

Coddington, Jonathan A., Ingi Agnarsson, Jeremy A. Miller, Matjaž Kuntner, and Gustavo Hormiga. 2009. 'Undersampling Bias: The Null Hypothesis for Singleton Species in Tropical Arthropod Surveys'. *Journal of Animal Ecology* 78 (3): 573–84. https://doi.org/10.1111/j.1365-2656.2009.01525.x.

Cuscó, Anna, Anna Salas, Celina Torre, and Olga Francino. 2019. 'Shallow Metagenomics with Nanopore Sequencing in Canine Fecal Microbiota Improved Bacterial Taxonomy and Identified an *Uncultured CrAssphage*'. Preprint. Microbiology. https://doi.org/10.1101/585067.

Davidov, Keren, Evgenia Iankelevich-Kounio, Iryna Yakovenko, Yuri Koucherov, Maxim Rubin-Blum, and Matan Oren. 2020. 'Identification of Plastic-Associated Species in the Mediterranean Sea Using DNA Metabarcoding with Nanopore MinION'. *Scientific Reports* 10 (1): 17533. https://doi.org/10.1038/s41598-020-74180-z.

Eshghi Sahraei, Shadi, Brendan Furneaux, Kerri Kluting, Mustafa Zakieh, Håkan Rydin, Håkan Hytteborn, and Anna Rosling. 2022. 'Effects of Operational Taxonomic Unit Inference

Methods on Soil Microeukaryote Community Analysis Using Long-Read Metabarcoding'. *Ecology and Evolution* 12 (3): e8676. https://doi.org/10.1002/ece3.8676.
Fuks, Garold, Michael Elgart, Amnon Amir, Amit Zeisel, Peter J. Turnbaugh, Yoav Soen, and Noam Shental. 2018. 'Combining 16S RRNA Gene Variable Regions Enables High-Resolution Microbial Community Profiling'. *Microbiome* 6 (1): 17. https://doi.org/10.1186/s40168-017-0396-x.
Furneaux, Brendan, Mohammad Bahram, Anna Rosling, Nourou S. Yorou, and Martin Ryberg. 2021. 'Long- and Short-Read Metabarcoding Technologies Reveal Similar Spatiotemporal Structures in Fungal Communities'. *Molecular Ecology Resources* 21 (6): 1833–49. https://doi.org/10.1111/1755-0998.13387.
Gueidan, Cécile, and Lan Li. 2022. 'A Long-Read Amplicon Approach to Scaling up the Metabarcoding of Lichen Herbarium Specimens'. *MycoKeys* 86 (February): 195–212. https://doi.org/10.3897/mycokeys.86.77431.
Ibironke, Olufunmilola, Lora R McGuinness, Shou-En Lu, Yaquan Wang, Sabiha Hussain, Clifford P Weisel, and Lee J Kerkhof. 2020. 'Species-Level Evaluation of the Human Respiratory Microbiome'. *GigaScience* 9 (4): giaa038. https://doi.org/10.1093/gigascience/giaa038.
J, Oksanen. 2010. 'Vegan : Community Ecology Package'. *Http://CRAN.R-Project.Org/Package=vegan*. https://cir.nii.ac.jp/crid/1571980076147214336.
Jeong, Jinuk, Kyeongeui Yun, Seyoung Mun, Won-Hyong Chung, Song-Yi Choi, Young-do Nam, Mi Young Lim, et al. 2021. 'The Effect of Taxonomic Classification by Full-Length 16S RRNA Sequencing with a Synthetic Long-Read Technology'. *Scientific Reports* 11 (1): 1727. https://doi.org/10.1038/s41598-020-80826-9.
Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, Emil A. Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. 2021. 'High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or PacBio Sequencing'. *Nature Methods* 18 (2): 165–69. https://doi.org/10.1038/s41592-020-01041-y.
Katiraei, Saeed, Yahya Anvar, Lisa Hoving, Jimmy F. P. Berbée, Vanessa van Harmelen, and Ko Willems van Dijk. 2022. 'Evaluation of Full-Length Versus V4-Region 16S RRNA Sequencing for Phylogenetic Analysis of Mouse Intestinal Microbiota After a Dietary Intervention'. *Current Microbiology* 79 (9): 276. https://doi.org/10.1007/s00284-022-02956-9.
Kerkhof, Lee J. 2021. 'Is Oxford Nanopore Sequencing Ready for Analyzing Complex Microbiomes?' *FEMS Microbiology Ecology* 97 (3): fiab001. https://doi.org/10.1093/femsec/fiab001.
Kolaříková, Zuzana, Renata Slavíková, Claudia Krüger, Manuela Krüger, and Petr Kohout. 2021. 'PacBio Sequencing of Glomeromycota RDNA: A Novel Amplicon Covering All Widely Used Ribosomal Barcoding Regions and Its Applicability in Taxonomy and Ecology of Arbuscular Mycorrhizal Fungi'. *New Phytologist* 231 (1): 490–99. https://doi.org/10.1111/nph.17372.
Liaw, Andy, and Matthew Wiener. 2002. 'Classification and Regression by RandomForest' 2.
Loit, Kaire, Kalev Adamson, Mohammad Bahram, Rasmus Puusepp, Sten Anslan, Riinu Kiiker, Rein Drenkhan, and Leho Tedersoo. 2019. 'Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens'. Edited by Irina S. Druzhinina. *Applied and Environmental Microbiology* 85 (21): e01368-19, /aem/85/21/AEM.01368-19.atom. https://doi.org/10.1128/AEM.01368-19.
Loos, Luna M. van der, Sofie D'hondt, Anne Willems, and Olivier De Clerck. 2021. 'Characterizing Algal Microbiomes Using Long-Read Nanopore Sequencing'. *Algal Research* 59 (November): 102456. https://doi.org/10.1016/j.algal.2021.102456.

Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. ‘Oxford Nanopore MinION Sequencing and Genome Assembly’. *Genomics, Proteomics & Bioinformatics* 14 (5): 265–79. https://doi.org/10.1016/j.gpb.2016.05.004.

Lu, Jingjing, Xudong Zhang, Xuan Zhang, Linqi Wang, Ruilin Zhao, Xiao Yong Liu, Xinzhan Liu, et al. 2022. ‘Nanopore Sequencing of Full RRNA Operon Improves Resolution in Mycobiome Analysis and Reveals High Diversity in Both Human Gut and Environments’. *Molecular Ecology* n/a (n/a). https://doi.org/10.1111/mec.16534.

Luo, Chengwei, Despina Tsementzi, Nikos Kyrpides, Timothy Read, and Konstantinos T. Konstantinidis. 2012. ‘Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample’. *PLOS ONE* 7 (2): e30087. https://doi.org/10.1371/journal.pone.0030087.

Maestri, Cosentino, Paterno, Freitag, Garces, Marcolungo, Alfano, et al. 2019. ‘A Rapid and Accurate MinION-Based Workflow for Tracking Species Biodiversity in the Field’. *Genes* 10 (6): 468. https://doi.org/10.3390/genes10060468.

Matsuo, Yoshiyuki, Shinnosuke Komiya, Yoshiaki Yasumizu, Yuki Yasuoka, Katsura Mizushima, Tomohisa Takagi, Kirill Kryukov, et al. 2021. ‘Full-Length 16S RRNA Gene Amplicon Analysis of Human Gut Microbiota Using MinION™ Nanopore Sequencing Confers Species-Level Resolution’. *BMC Microbiology* 21 (1): 35. https://doi.org/10.1186/s12866-021-02094-5.

McMurdie, Paul J., and Susan Holmes. 2013. ‘Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data’. *PLOS ONE* 8 (4): e61217. https://doi.org/10.1371/journal.pone.0061217.

Mikryukov, Vladimir. 2019. ‘MetagMisc: Miscellaneous Functions for Metagenomic Analysis’.

Parada, Alma E., David M. Needham, and Jed A. Fuhrman. 2016. ‘Every Base Matters: Assessing Small Subunit RRNA Primers for Marine Microbiomes with Mock Communities, Time Series and Global Field Samples: Primers for Marine Microbiome Studies’. *Environmental Microbiology* 18 (5): 1403–14. https://doi.org/10.1111/1462-2920.13023.

Pei, Anna Y., William E. Oberdorf, Carlos W. Nossa, Ankush Agarwal, Pooja Chokshi, Erika A. Gerz, Zhida Jin, et al. 2010. ‘Diversity of 16S RRNA Genes within Individual Prokaryotic Genomes’. *Applied and Environmental Microbiology* 76 (12): 3886–97. https://doi.org/10.1128/AEM.02953-09.

Pereira, Tiago José, Alejandro De Santiago, Taruna Schuelke, Sarah M. Hardy, and Holly M. Bik. 2020. ‘The Impact of Intragenomic RRNA Variation on Metabarcoding-derived Diversity Estimates: A Case Study from Marine Nematodes’. *Environmental DNA* 2 (4): 519–34. https://doi.org/10.1002/edn3.77.

Planý, Matej, Jozef Sitarčík, Jelena Pavlović, Jaroslav Budiš, Janka Koreňová, Tomáš Kuchta, and Domenico Pangallo. 2023. ‘Evaluation of Bacterial Consortia Associated with Dairy Fermentation by Ribosomal RNA (Rrn) Operon Metabarcoding Strategy Using MinION Device’. *Food Bioscience* 51 (February): 102308. https://doi.org/10.1016/j.fbio.2022.102308.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. ‘The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools’. *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Schlaeppi, Klaus, S. Franz Bender, Fabio Mascher, Giancarlo Russo, Andrea Patrignani, Tessa Camenzind, Stefan Hempel, Matthias C. Rillig, and Marcel G. A. van der Heijden. 2016. ‘High-Resolution Community Profiling of Arbuscular Mycorrhizal Fungi’. *New Phytologist* 212 (3): 780–91. https://doi.org/10.1111/nph.14070.

Shin, HyeonSeok, Eunju Lee, Jongoh Shin, So-Ra Ko, Hyung-Seok Oh, Chi-Yong Ahn, Hee-Mock Oh, Byung-Kwan Cho, and Suhyung Cho. 2018. ‘Elucidation of the Bacterial Communities Associated with the Harmful Microalgae Alexandrium Tamarense and Cochlodinium

Polykrikoides Using Nanopore Sequencing'. *Scientific Reports* 8 (1): 5323. https://doi.org/10.1038/s41598-018-23634-6.
Shin, Jongoh, Sooin Lee, Min-Jeong Go, Sang Yup Lee, Sun Chang Kim, Chul-Ho Lee, and Byung-Kwan Cho. 2016. 'Analysis of the Mouse Gut Microbiome Using Full-Length 16S RRNA Amplicon Sequencing'. *Scientific Reports* 6 (1): 29681. https://doi.org/10.1038/srep29681.
Simberloff, Daniel. 1972. 'Properties of the Rarefaction Diversity Measurement'. *The American Naturalist* 106 (949): 414–18. https://doi.org/10.1086/282781.
Taş, Neslihan, Anniek EE de Jong, Yaoming Li, Gareth Trubl, Yaxin Xue, and Nicholas C Dove. 2021. 'Metagenomic Tools in Microbial Ecology Research'. *Current Opinion in Biotechnology* 67 (February): 184–91. https://doi.org/10.1016/j.copbio.2021.01.019.
Tedersoo, Leho, Mads Albertsen, Sten Anslan, and Benjamin Callahan. 2021. 'Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology'. *Applied and Environmental Microbiology* 87 (17): e00626-21. https://doi.org/10.1128/AEM.00626-21.
Tedersoo, Leho, Ave Tooming-Klunderud, and Sten Anslan. 2018. 'PacBio Metabarcoding of Fungi and Other Eukaryotes: Errors, Biases and Perspectives'. *New Phytologist* 217 (3): 1370–85. https://doi.org/10.1111/nph.14776.
Toxqui Rodríguez, Maria del Socorro, Joran Vanhollebeke, and Sofie Derycke. 2023. 'Evaluation of DNA Metabarcoding Using Oxford Nanopore Sequencing for Authentication of Mixed Seafood Products'. *Food Control* 145 (March): 109388. https://doi.org/10.1016/j.foodcont.2022.109388.
Urban, Lara, Andre Holzer, J Jotautas Baronas, Michael B Hall, Philipp Braeuninger-Weimer, Michael J Scherm, Daniel J Kunz, et al. 2021. 'Freshwater Monitoring by Nanopore Sequencing'. Edited by María Mercedes Zambrano, Bavesh D Kana, María Mercedes Zambrano, and Alejandro Sanchez-Flores. *ELife* 10 (January): e61504. https://doi.org/10.7554/eLife.61504.
Větrovský, Tomáš, and Petr Baldrian. 2013. 'The Variability of the 16S RRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses'. *PLOS ONE* 8 (2): e57923. https://doi.org/10.1371/journal.pone.0057923.
Walters, William, Embriette R. Hyde, Donna Berg-Lyons, Gail Ackermann, Greg Humphrey, Alma Parada, Jack A. Gilbert, et al. 2016. 'Improved Bacterial 16S RRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys'. Edited by Holly Bik. *MSystems* 1 (1). https://doi.org/10.1128/mSystems.00009-15.
Weisburg, W G, S M Barns, D A Pelletier, and D J Lane. 1991. '16S Ribosomal DNA Amplification for Phylogenetic Study'. *Journal of Bacteriology* 173 (2): 697–703. https://doi.org/10.1128/jb.173.2.697-703.1991.
Willis, Ciara, Dhwani Desai, and Julie LaRoche. 2019. 'Influence of 16S RRNA Variable Region on Perceived Diversity of Marine Microbial Communities of the Northern North Atlantic'. *FEMS Microbiology Letters* 366 (13): fnz152. https://doi.org/10.1093/femsle/fnz152.

Acknowledgements

This work was lead by the GT DCE Mangroves (Groupe de Travail Directive Cadre sur l'Eau Mangroves), founded by the OFB (Office Français de la Biodiversité, Olivier Monnier), lead by the MNHN (Muséum National d'Histoire Naturelle, GD). Lab work was achieved by AL during her Master practice, with BOREA financial support. Thanks to our colleague Cédric Hubas (BOREA), for discussion and statistical advice.

Data accessibility

All data presented and scripts in this manuscript are available from the GitHub repository : https://github.com/tonyrobinet/nanopore_metabarcoding. Raw sequences (fastq format) are submitted to SRA database, available with BioProject accession number XXXXXXXX.

Author contributions

G.D. and T.R. designed the study and seeked for funds ; G.D. conducted the fieldwork and collected samples; T.R. designed the lab protocols, with support of A.L. ; A.L. performed the lab work, with the supervision of M.G. and T.R. ; T.R. and A.L. performed the statistical analysis ; T.R. wrote the manuscript, corrected by A.L. and G.D.

Competing interests

The authors declare no competing interests.

Funding

This work has been funded by the OFB (Office Français de la Biodiversité, Groupe de Travail Mangroves) for field work and G.D. appointment, and by our research unit BOREA for A.L. Master grant and genetics consumables.

Figures & Tables

Figure 1. (**a**) Location of sampling sites on Guadeloupe Island (red squares) ; (**b**) zoom on the two sampling sites, with site names ; (**c**) sampling protocol in each site : 3 lines of 3 points, each composed of 3 biological replicates (a, b and c), at 12.5m of distance between each point on each line.
Figure 2. Abundance of bacterial genus retrieved from a mock community sample, by illumina and nanopore.

Figure 2. Relative abundances found in a mock community sequenced by Illumina (once, 16sV4) and Nanopore (twice, complete 16s), after singletons filtering ; the theoretical abundances are shown at left. *Eukaryotic taxa, which are present in the mock but not supposed to be amplified with 16s bacterial markers. Mock datasets were filtered at a minimum depth of 50 reads per OTU, at species level (97%), then aggregated to genus level.

Figure 3. (**a**) Venn diagrams showing the proportions of bacterial taxa shared and unshared between both sequencers, at each phylogenetic rank (numbers in the discs refer to the numbers of taxa of the portion of the disc it is written on) ; (**b**) proportion of bacterial taxa shared between both sequencers (lines, left axis), number of taxa shared and unshared (bars, right axis) at each phylogenetic rank ; more details in Table S1.

Figure 4. Number of OTUs (97% similarity, singleton-filtered) for each prokaryotic phylum in environmental samples analyzed here (bacterial + archaeal dataset), depending on the sequencing device : (**a**) with conventional equal-rarefaction (1582 reads for all samples of both sequencers, see Methods section for comments on inner bias) ; (**b**) with coverage-based read rarefaction (Chao and Jost 2012 ; Illumina 2609 reads in average [min 1338, max 4991], Nanopore 5108 reads in average [min 4451, max 6019]). Phylum names in red or blue were detected only by Illumina or only by Nanopore, respectively. Red or blue dots indicate core-phyla, i.e. phyla with a minimum prevalence of 50% in the respective datasets. Red or blue arrows indicate phyla that were not detected with read equal-rarefaction, for Illumina or Nanopore respectively.

Figure 5. (**a**) Top-20 bacterial orders in samples for both sequencing devices, ranked by their overall relative abundances in samples ; (**b**) biplot of sample scores from a nMDS on abundances of bacterial OTUs agglomerated at genus level, for both sequencing devices (stress=14.1%) ; for this common ordination, shared OTUs were named differently between Illumina and Nanopore on purpose, in order to separate the two datasets for a better visualization. Number of reads per sample was rarefied with the coverage-based method (Chao and Jost 2012).

Figure 6. (**a-b**) PCoA on coverage-based rarefied abundances of bacterial communities at species level, (**a**) sequenced by Illumina (16sV4), showing biological replicates (polygons) ; (**b**) sequenced by Nanopore (complete 16s) ; (**c**) Dispersion of Bray-Curtis dissimilarity index within biological replicates, salmon boxplots for Illumina, cyan for Nanopore ; thick horizontal lines : mean ; box plots : 75% range ; whiskers : 95% range ; dots: outliers ; (**d**) Procrustes analysis of the 2 first components of both PCoAs (presented in a-b), showing the degree of matching between the two ordinations ; empty dots show the position of the samples in the Nanopore ordination and arrows point to their positions in the Illumina ordination ; the plot also shows the rotations between the axis (solid *vs*. dashed), necessary to make ordinations match as closely as possible ; (**e**) residuals for each sample between the ordinations (this time, on the 20 first axis); the horizontal lines, from bottom to top, are the 25% (dashed), 50% (solid), and 75% (dashed) quantiles of the residuals.

Table 1. Statistics on bacterial taxa detected by Illumina and Nanopore sequencers, shared or unshared between sequencers, by taxonomic rank, percentage of taxa or reads assigned. Here only the bacterial dataset was coverage-based rarefied and singletons filtered, leading to a slightly different result than that presented in Figure 4, on which a coverage-based rarefaction and singletons filtering has been done on all prokaryotic dataset.

Supplementary material

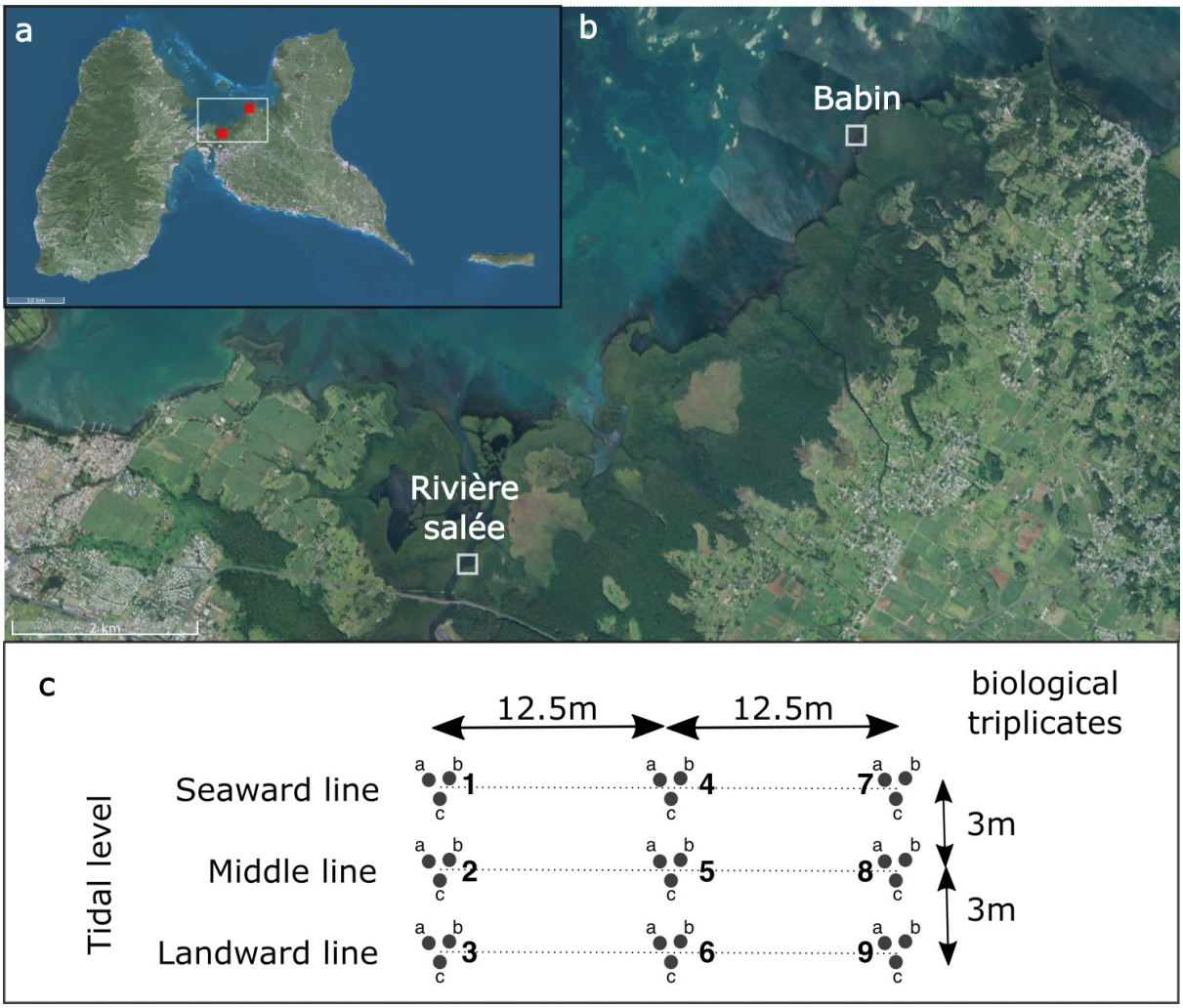
Available at https://github.com/tonyrobinet/nanopore_metabarcoding

Figure S1. Contribution of Mean Decrease Gini coefficient (MDG) of common genus (**a**) and common families (**b**) sequenced by Illumina and Nanopore, for [site+(sea-land orientation)] predictors (see details in Table S2). Mean Decrease Gini is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest (see Methods for details) ; the higher the value of MDG score, the higher the importance of the variable in the model.

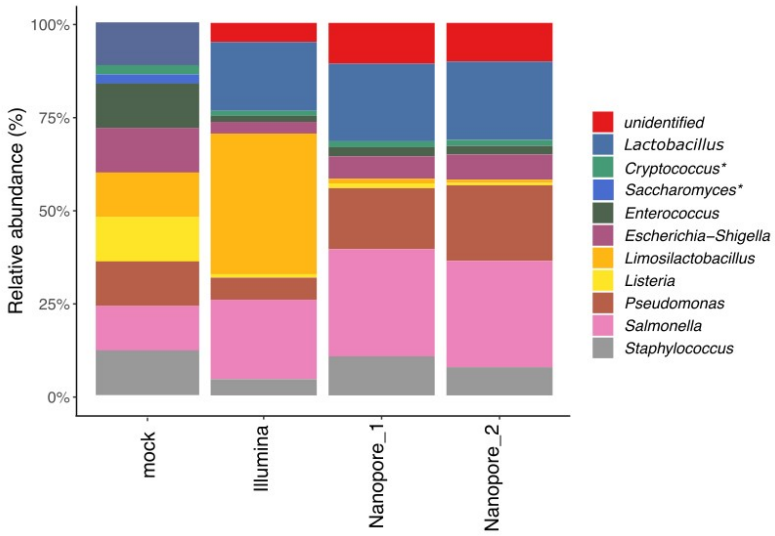
Figure S2. (**a-b**) Archaean taxa (genus level) contributing to structuring the communities in samples sequenced by Nanopore (rarefied at 5500 reads per sample, 97% OTUs with a minimum coverage of 50 reads) : (**a**) PCA on relative abundances, (**b**) iris plot of the relative abundances for taxa the most contributing to the PCA in (a). (**c-d**) same for bacterial taxa (genus level), sequenced by Nanopore.

Table S1. Archaea detected by Illumina were mentioned but the read coverage by sample was much lower than those for Nanopore Archaea.

Table S2. Bacterial genus contributing the most importantly to the site effect, after a random forest analysis on Illumina and Nanopore datasets. In green : OTUs common to both datasets. MDG : mean decrease in Gini coefficient, a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest ; the higher the value of MDG score, the higher the importance of the variable in the model.



464 **Figure 1**. (**a**) Location of sampling sites on Guadeloupe Island (red squares) ; (**b**) zoom on the two
465 sampling sites, with site names ; (**c**) sampling protocol in each site : 3 lines of 3 points, each com-
466 posed of 3 biological replicates (a, b and c), at 12.5m of distance between each point on each line.



467 **Figure 2**. Relative abundances found in a mock community sequenced by Illumina (once, 16sV4)
468 and Nanopore (twice, complete 16s), after singletons filtering ; the theoretical abundances are
469 shown at left. *Eukaryotic taxa, which are present in the mock but not supposed to be amplified
470 with 16s bacterial markers. Mock datasets were filtered at a minimum depth of 50 reads per OTU,
471 at species level (97%), then aggregated to genus level.

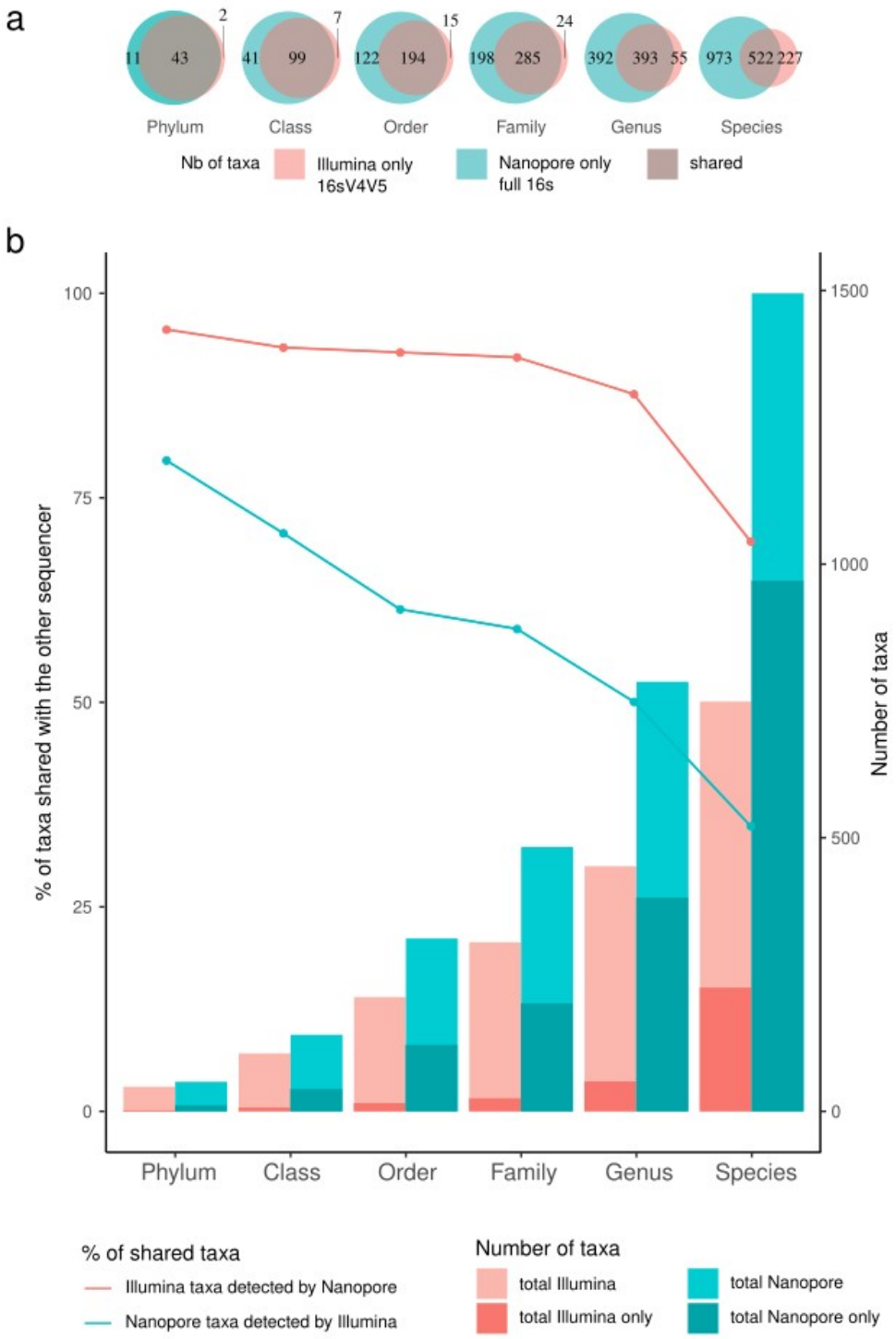


Figure 3. (**a**) Venn diagrams showing the proportions of bacterial taxa shared and unshared between both sequencers, at each phylogenetic rank (numbers in the discs refer to the numbers of taxa of the portion of the disc it is written on) ; (**b**) proportion of bacterial taxa shared between both sequencers (lines, left axis), number of taxa shared and unshared (bars, right axis) at each phylogenetic rank ; more details in Table S1.

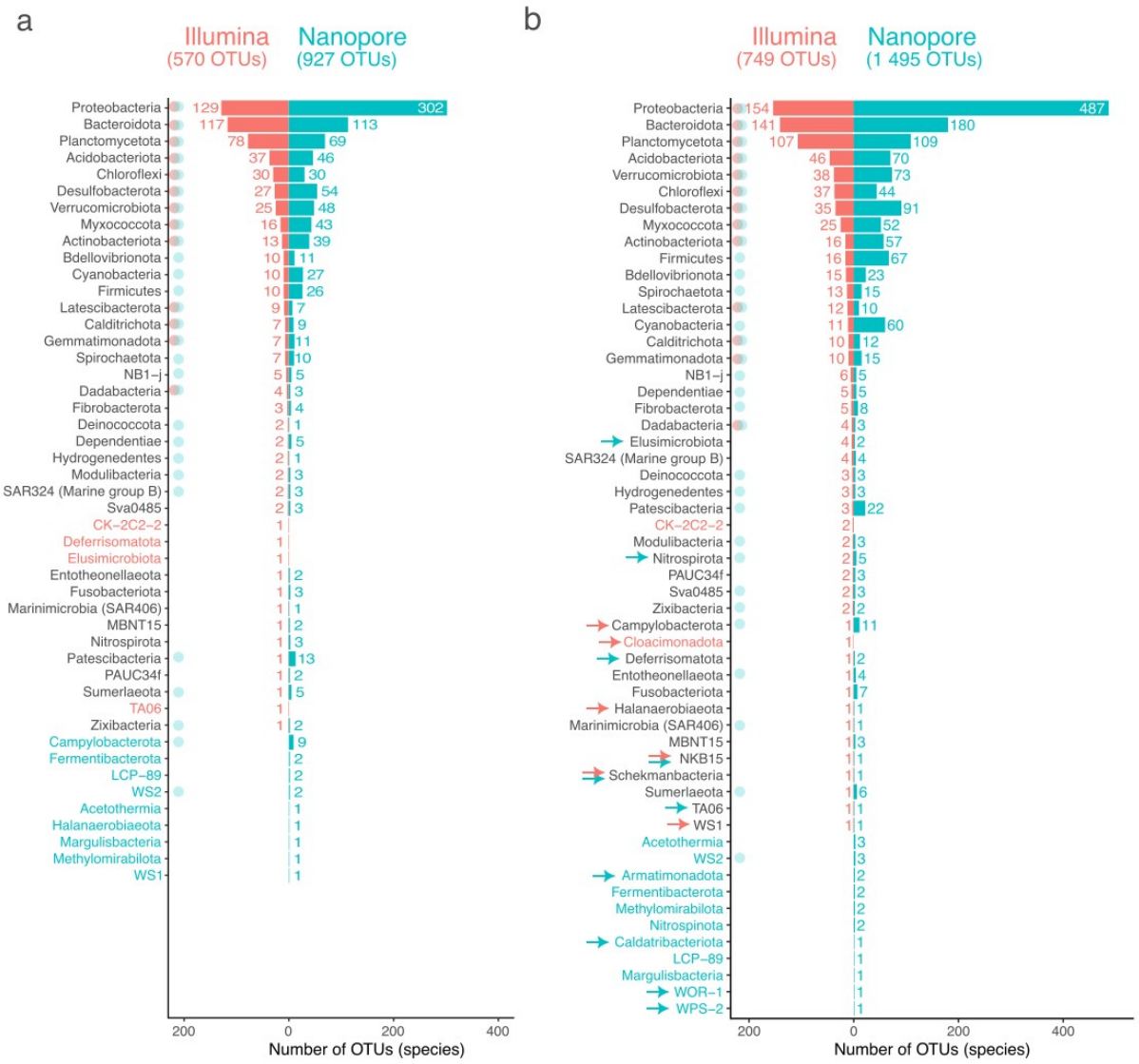


Figure 4. Number of OTUs (97% similarity, singleton-filtered) for each prokaryotic phylum in environmental samples analyzed here (bacterial + archaeal dataset), depending on the sequencing device : (**a**) with conventional equal-rarefaction (1582 reads for all samples of both sequencers, see Methods section for comments on inner bias) ; (**b**) with coverage-based read rarefaction (Chao and Jost 2012 ; Illumina 2609 reads in average [min 1338, max 4991], Nanopore 5108 reads in average [min 4451, max 6019]). Phylum names in red or blue were detected only by Illumina or only by Nanopore, respectively. Red or blue dots indicate core-phyla, i.e. phyla with a minimum prevalence of 50% in the respective datasets. Red or blue arrows indicate phyla that were not detected with read equal-rarefaction, for Illumina or Nanopore respectively.

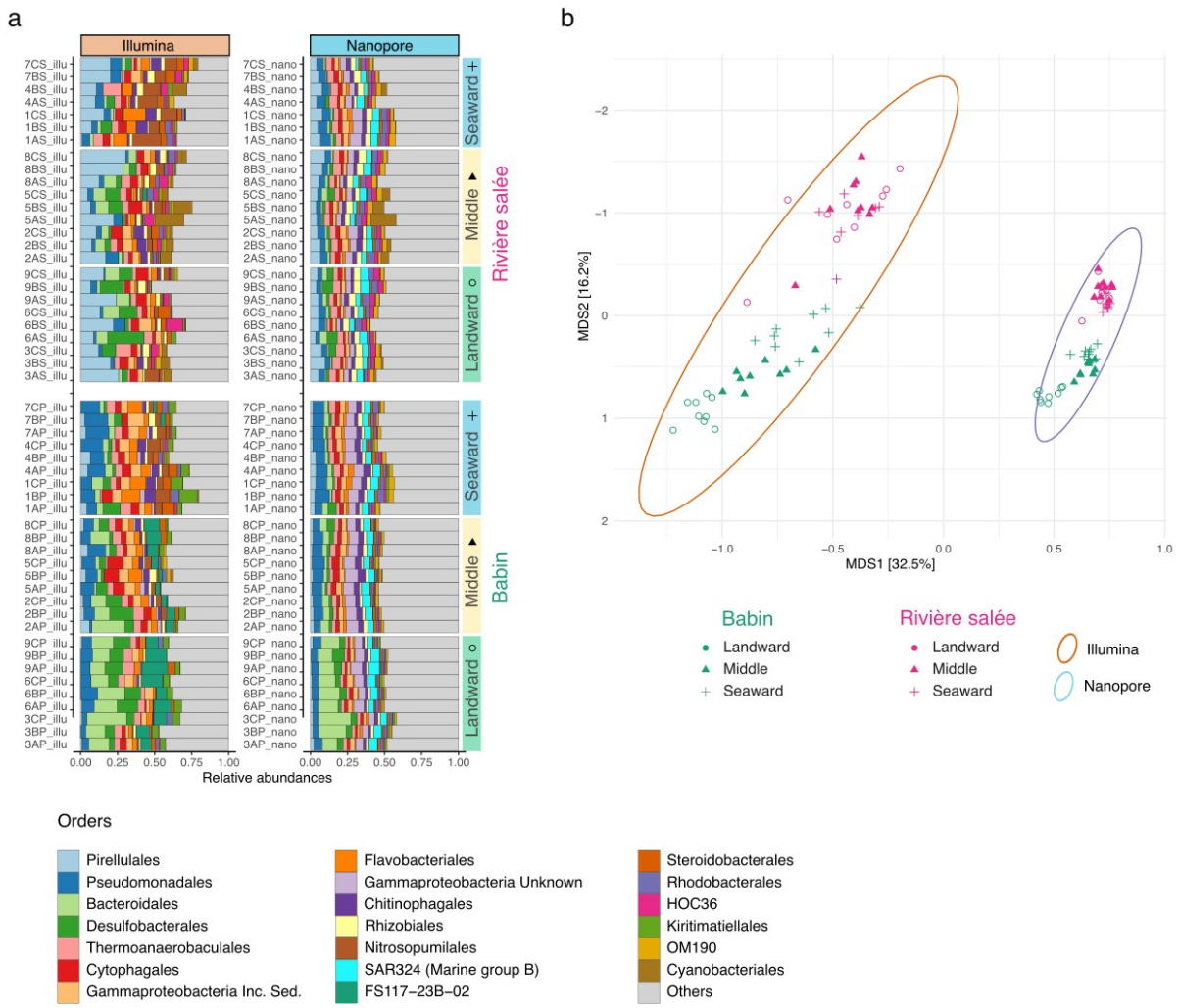


Figure 5. (**a**) Top-20 bacterial orders in samples for both sequencing devices, ranked by their overall relative abundances in samples ; (**b**) biplot of sample scores from a nMDS on abundances of bacterial OTUs agglomerated at genus level, for both sequencing devices (stress=14.1%) ; for this common ordination, shared OTUs were named differently between Illumina and Nanopore on purpose, in order to separate the two datasets for a better visualization. Number of reads per sample was rarefied with the coverage-based method (Chao and Jost 2012).

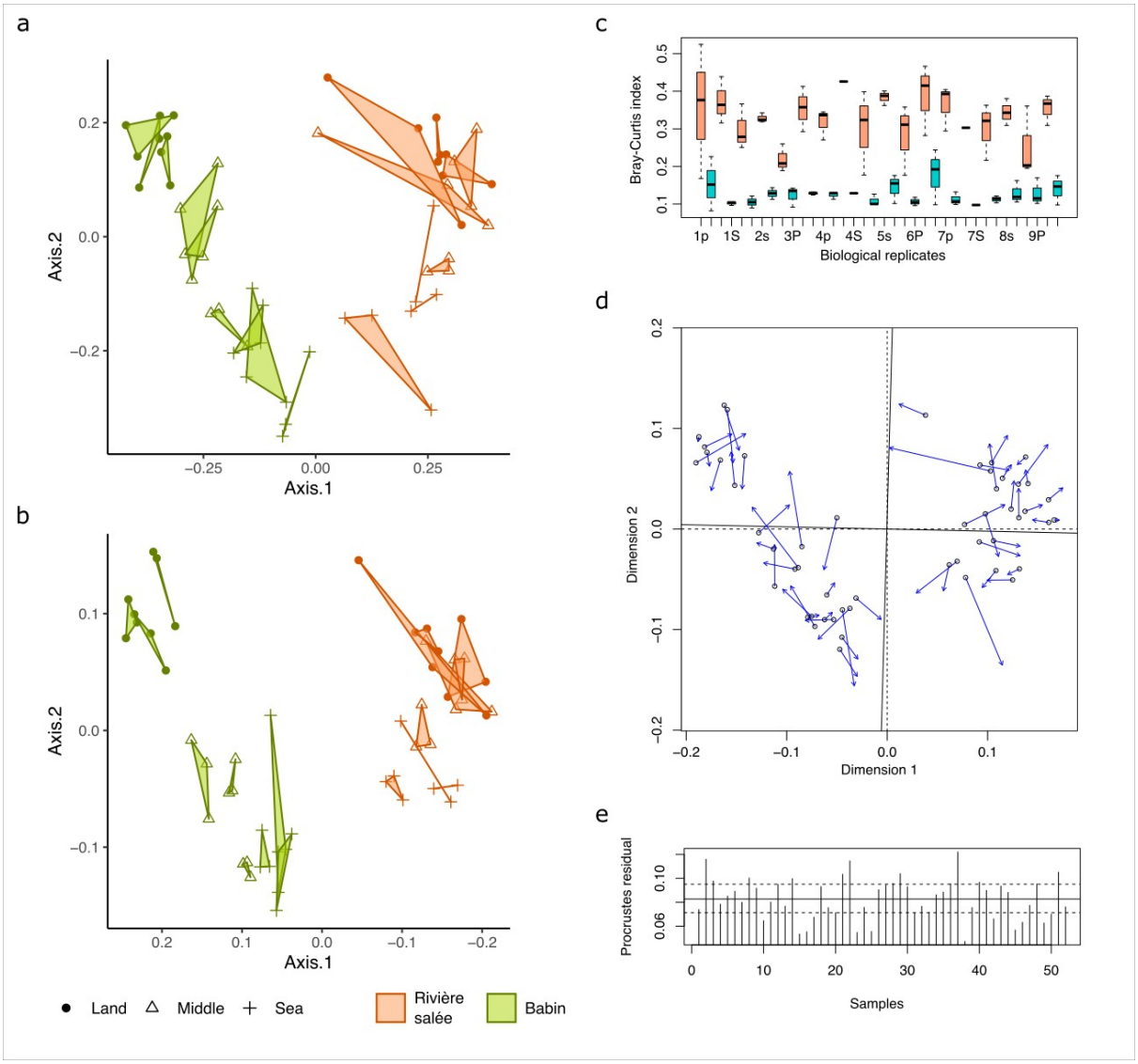


Figure 6. (**a-b**) PCoA on coverage-based rarefied abundances of bacterial communities at species level, (**a**) sequenced by Illumina (16sV4), showing biological replicates (polygons) ; (**b**) sequenced by Nanopore (complete 16s) ; (**c**) Dispersion of Bray-Curtis dissimilarity index within biological replicates, salmon boxplots for Illumina, cyan for Nanopore ; thick horizontal lines : mean ; box plots : 75% range ; whiskers : 95% range ; dots: outliers ; (**d**) Procrustes analysis of the 2 first components of both PCoAs (presented in a-b), showing the degree of matching between the two ordinations ; empty dots show the position of the samples in the Nanopore ordination and arrows point to their positions in the Illumina ordination ; the plot also shows the rotations between the axis (solid *vs.* dashed), necessary to make ordinations match as closely as possible ; (**e**) residuals for each sample between the ordinations (this time, on the 20 first axis); the horizontal lines, from bottom to top, are the 25% (dashed), 50% (solid), and 75% (dashed) quantiles of the residuals.

Kingdom	sequencer	Taxa / *reads*	Phylum	Class	Order	Family	Genus	Species
Bacteria	Illumina (515F + 926R)	detected (% taxa assigned) *(% reads assigned)*	45 (97.8%) *(99.6%)*	106 (92.5%) *(99.6%)*	209 (91.9%) *(98.3%)*	309 (88.0%) *(98.0%)*	448 (82.4%) *(94.0%)*	749 (65.3%) *(54.0%)*
		unshared (% taxa assigned) *(% reads assigned)*	2 (100%) *(100%)*	7 (100%) *(100%)*	15 (93.3%) *(97.1%)*	24 (87.5%) *(83.9%)*	55 (81.8%) *(84.5%)*	227 (80.6%) *(85.8%)*
	Nanopore (27F + 1492R)	detected (% taxa assigned) *(% reads assigned)*	54 (98.1%) *(88.3%)*	140 (89.3%) *(86.0%)*	316 (89.2%) *(74.9%)*	483 (85.9%) *(73.3%)*	785 (80.9%) *(67.8%)*	1495 (64.6%) *(46.9%)*
		unshared (% taxa assigned) *(% reads assigned)*	11 (100%) *(100%)*	41 (82.9%) *(70.5%)*	122 (85.2%) *(87.2%)*	198 (82.8%) *(83.7%)*	392 (79.3%) *(81.8%)*	973 (67.8%) *(67.4%)*
	shared taxa (% all detected taxa shared) (% taxa assigned)		43 (76.8%) (97.6%)	99 (67.3%) (91.9%)	194 (58.6%) (91.8%)	285 (56.2%) (88.1%)	393 (46.8%) (82.4%)	522 (30.3%) (58.6%)
	% of shared taxa for Illumina *(% reads assigned among shared)*		95.6% *(99.6%)*	93.4% *(99.6%)*	92.8% (98.3%)	92.2% (98.0%)	87.7% *(94.2%)*	69.7% *(50.6%)*
	% of shared taxa for Nanopore *(% reads assigned among shared)*		79.6% *(88.3%)*	70.7% *(86.1%)*	61.4% (74.4%)	59.0% (72.8%)	50.1% *(66.5%)*	34.9% *(41.1%)*

Table 1. Statistics on bacterial taxa detected by Illumina and Nanopore sequencers, shared or unshared between sequencers, by taxonomic rank, percentage of taxa or reads assigned. Here only the bacterial dataset was coverage-based rarefied and singletons filtered, leading to a slightly different result than that presented in Figure 4, on which a coverage-based rarefaction and singletons filtering has been done on all prokaryotic dataset.