Fine-scale congruence in bacterial community structure from marine sediments sequenced by short-reads on Illumina and long-reads on Nanopore

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 are usually described by metabarcoding with genetic markers on the 16S domain. However, 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 a mock community and 52 sediment samples from two contrasted mangrove sites, described by short-reads on 16SV4-V5 marker (*ca*. 0.4kpb) analyzed by Illumina sequencing (MiSeq, V3), with those described by long-reads on bacterial nearly complete 16S (*ca*. 1.5 kpb) analyzed by Oxford Nanopore (MinION, R9.2). Short- and long-reads retrieved all the bacterial genera 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 singletons filtering, co-inertia and Procrustean tests showed that bacterial community structures inferred from short- and long-reads were significantly similar, showing both a comparable contrast between sites and a coherent sea-land orientation within sites. In our dataset, 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. Primer specificities of long-16S allowed it to detect 92.2% of the 309 families and 87.7% of the 448 genera that were detected by the short 16SV4-V5. Long-reads recorded 973 additional taxa not detected by short-reads, among which 91.7% were identified to the genus rank, some belonging to 11 exclusive phyla, albeit accounting for only 0.2% of total long-reads. [276 words]

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

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 the useful MAGs (metagenome assembled genomes), 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, an intragenomic variability in the number of duplicates of ribosomal operon, bearing differences in allelic variants between copies (Pereira et al. 2020) ; and (iii) relative abundances of reads per taxa are somehow inaccurate, compared to awaited abundances in the mock samples, a probable 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 (SR) 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, any 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 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 (tens of kpb 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 (LR) 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). LR 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 SR platforms. LR metabarcoding is mostly used for taxonomic groups in which SR 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 LR metabarcoding.

Despite raw reads accuracy are similar for PacBio (88-90%) and Nanopore (95-98% on the R9 flow-cells, above 99% for R10.4), 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 LR third-generation sequencer acknowledged to be suitable for metabarcoding was PacBio Sequel II on fungal complete rRNA operon (ca. 3000 bp, Tedersoo, Tooming-Klunderud and Anslan 2018). Despite its error rate being slightly higher than Illumina, the PacBio LR 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, Baloğlu et al. 2021) or by flanking, at the two first steps of PCR, each single amplicon with a unique molecular identifier (UMI, Karst et al. 2021). 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 for directly sequencing amplicons from environmental samples, as we tested it without success. To date, no environmental metabarcoding based on RCA or UMI protocols have been published.

In community ecology, Nanopore was initially used for simply barcoding individuals with LR (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), fungal communities and biotic interactions (Vass et al. 2022) 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".

At the time of this study, there was no sequencing platform in the research center nearby the place where sediment samples were collected and stored (Guadeloupe Island). 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. equipped 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.

To date, no published work has compared Nanopore LR to Illumina SR bacterial metabarcoding on the same environmental samples, perhaps because of the impossibility to disentangle the effects of primers from those of sequencing technologies. Here we propose to overcome this issue by evaluating the similarity of bacterial communities described from marine sediments sequenced by 16S LR on Nanopore MinION device, and by sequencing in parallel the 16S-V4V5 SR amplified from same DNA extracts on Illumina MiSeq, and by addressing a simple question: will these sequencing strategies conserve (i) the structures of bacterial communities between two neighboring mangrove sites, and (ii) the sea-land orientation of bacterial communities within sites?

Materials and methods

Sampling sites and sample collection

In June 2019, two 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 a Ramsar-protected area close to coral reefs (latitude -61,5294 ; longitude 16,3388).

A total of 54 samples of surface sediment were collected on intertidal zone, on 3 lines of 3 points each respectively in each site, each line separated by 3 m to the neighboring line. Points were separated by 12.5 m within a line. Each point was composed of 3 biological replicates (a, b and c), analyzed in the workflow separately (Fig.1c). The line closest to the sea was the "seaward line", those closest to the inland mangrove was the "landward line", 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 sent in France, freeze-dried and crushed to powder in a mortar, carefully cleaned with an alcoholic tissue between each sample processing. Total genomic DNA from 50mg of dried samples and a standard microbial community (zymoBIOMICS Microbial Community Standard D6300, by ZYMO RESEARCH), 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^{-1} 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). Two other primer pairs (18SV9 and ITS2) were amplified, added to libraries and sequenced together with 16SV4-V5. They were in raw reads but not presented here. Each primer was flanked in its 5'-end by a nucleotide sequence used for indexing at a later step, according to a protocol proposed by Nag et al. 2017. At this stage, 2 additional 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 2× 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 20s at 98°C, hybridization 20s at 60 °C, elongation 10s at 72°C) and final elongation 10s 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), with an waited size peak around 420bp, then pooled in a final library, and sequenced on an Illumina MiSeq (one Miseq Reagent v3 kit 600 cycles and one nano MiSeq Reagent kit v2 kit 500 cycles for resequencing) in the Concarneau marine station (MNHN) to output demultiplexed fastq files.

Nanopore library

The same DNA extracts were processed in parallel for Nanopore LR sequencing, with the following 16S markers : V1-V9 regions (nearly complete 16S for bacteria, ~1.45 kpb; Weisburg et al. 1991; 27F:AGAGTTTGATCMTGGCTCAG ; 1492R: TACGGYTACCTTGTTACGACTT). PCRs were performed in 3 small-volume replicates of 12,5 µl each, containing 6,25µl of LongAmp Taq 2×

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, followed by 30 cycles composed of denaturation for 30s at 94°C, hybridization for 30s at 51°C, and elongation for 45s 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 brought back to a concentration of 1ng.µl^{-1}. Indexation PCR was realized according to the Nanopore « PCR barcoding (96) amplicons (SQK-LSK109) » protocol. 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. 16S V1-V9 library was sequenced on two R9.4.1 MinION flow cells (half of the samples + Ze for each). Flow cells were loaded on MinION Mk-1C and sequenced for approximately 48h, until no further sequencing reads could be collected above Q10 quality score. 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 SR and Nanopore LR, indexes with less than 1500 reads were re-sequenced.

Sequence data are available in NCBI with BioProject accession number PRJNA985243.

Processing of raw reads

Fastq files from Illumina SR were filtered with R package DADA2 v 1.16.0 (Callahan et al. 2016). Reads R1 and R2 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-V5 ASVs were extracted from fasta files containing sequences from other primers (18SV9 and ITS2, not presented here). Nanopore LR fastq sequences (>Q10) were filtered with Nanofilt : all reads shorter than 1.4 kpb and longer than 1.6 kpb for 16S V1-V9 were removed. Then, ASVs with 97% of similarity were clustered into OTUs using the Vsearch (version 2023.7.0) 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 this specific 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). 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.

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 due to the random sorting of rare species (Coddington et al. 2009). In soil or sediment microbiotas, sample OTU richness depends 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).

Since the SR primers (16SV4-V5) amplified both bacteria and archaea, but LR primers (16SV1-V9) amplified bacteria only, archaeal taxa obtained by SR were removed for the present analysis (deleted after rarefaction). Further analyses with archaeal taxa are however proposed in Supplementary Materials.

Analyses were carried out on filtered OTU tables after coverage-based rarefaction, except for Fig. 5, in which both rarefaction methods are shown. In sediment samples, core members were identified by their prevalence among all samples (≥50%, i.e. they were present in 50% or more of the samples, Fig. 5). For exploring dissimilarities between datasets, 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 R package ade4, *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 SR and LR bacterial communities (Pearson method, 999 permutations), with vegan. Classification trees were used to characterize the genus and species contributing the most to the [site x (sea-land orientation] effect in each dataset by the R package randomForest (Liaw and Wiener 2002).

Results

Mock Community

OTUs were filtered on a minimum of 50 reads/OTU for Ze. 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 supposed to have been initially introduced (12% for each prokaryote). The SR and LR methods both overestimated some genus *(Lactobacillus*, *Limosilactobacillus, Salmonella*) and underestimated some others (*Escherichia*, *Listeria, Enterococcus*). Both methods found undetermined reads (5.1% for SR, 11.0 and 10.4% for LR_1 and LR_2, respectively). The re-sequencing of Ze on two different LR 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 counted 170% of the species detected by 16SV4-V5. With coverage-based rarefaction, samples analyzed by SR presented 749 species for 2609 reads in average (min 1338, max 4991), samples analyzed by LR showed 1495 species for 5108 reads in average (min 4451, max 6019, Table 1). In proportion with this rarefaction method, LR detected twice more species than SR. For the rest of this section, only results obtained by the coverage-based rarefaction method are presented.

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. Communities sequenced by SR and LR described the same global patterns, i.e. a preponderance of Pirellulales (Planctomycetota) in Rivière salée, of Pseudomonadales (Gammaproteobacteria) and Bacteroidales (Bacteroidota) in Babin, separating clearly the two sites in ordination (Fig. 3a-b). Babin showed the most structured community along the tidal gradient, with the presence of Pseudomonadales in seaward samples and of Bacteroidales (Bacteroidota) in landward samples. Biological replicates were relatively close to each other in the PCoAs (Fig. 4a-b), but Bray-Curtis dissimilarity indexes of communities within replicates were always higher for SR than for LR, either for Babin or Rivière salée (Fig. 4c, anova $p<0.001$).

The Procrustes analysis of the two first axes of multivariates showed a significantly strong similarity between structures drawn by SR and LR (Fig. 4d-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 LR and SR communities using OTUs at species taxonomic rank. In order to point out the similarity of taxa

contributing to the [site × (sea-land orientation] effect, classification trees were made by a random forest approach on the 393 genus and 285 families shared between SR and LR. 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 × (sea-land orientation] effect were scarce (Fig. S2, Table S1). Archaean communities described with LR specific primers followed roughly the structure obtained with bacteria (Fig. S3).

Phylogenetic diversity

Over the 56 bacterial phyla detected in total, 54 were detected by LR and 45 by SR (Table 1). At high taxonomic levels, 16SV4-V5 and full-16S were approximately ⅘ alike for phyla, LR detected 11 exclusive phyla over a total of 54 for this primer (20% of exclusive among those detected by LR), when SR only had 2 (4.4%). However, LR-exclusive subcommunity represented only 0.2% of reads in the full community (Fig. S1). The 11 phyla only detected by LR were *Acetothermia*, *WS2*, *LCP-89*, *WOR-1*, *Armatimonadota*, *Margulisbacteria*, *Nitrospinota*, *Fermentibacterota*, *Methylomirabilota*, *Caldatribacteriota*, *WPS-2*, whereas the only 2 detected by SR were *Cloacimonadota* and *CK-2C2-2,* with a coverage-based rarefaction (Fig. 5). At lower taxonomic levels, 92.2% and 87.7% of respectively the family and bacterial genus detected by 16SV4-V5 were detected by full-16S. LR detected twice more species than SR, with only 34.9% of the species and 50.1% of the genus detected shared with SR. The trend that LR detected almost all SR taxa *plus* a certain number of LR original taxa decreased with lowering taxonomic ranks (Fig. 6).

All the 54 LR-detected phyla were more diversified based on full-16S, but four : *NB1-j*, *SAR324*, *Dadabacteria* and *Hydrogendentes*. The most diversified phylum, the *Proteobacteria*, presented more than 4 times more species with full-16S than with 16SV4-V5. Overall, communities described by the two primer-sets were phylogenetically very similar when considering shared taxa at the family and genus level (92.2% and 87.7% of taxa similarity for SR *vs.* LR, respectively).

11.7% of the LR (bacterial 16S) were unassigned at the phylum level, versus 0.36% for SR. 53.1% of the LR unassigned at the species level (35.4% of total LR bacterial OTUs), versus 46.0% for SR (34.7% of total SR bacterial OTUs, Table 1). For shared genera, the unassigned reads were much lower for SR (5.8%) than LR (35.5%). All core-phyla detected by SR were also parts of core-phyla detected by LR, whatever the rarefaction method used (Fig. 5).

Discussion

In this study, bacteria were amplified on their rRNA gene by 16SV4-V5 and full-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 primer pairs and sequencing tools were similar in their coarse structure (site effect) and fine

structure (sea-land orientation), with nonetheless a couple of constant differences, already noticed with SR *vs*. LR sequencing on PacBio (Katiraei et al. 2022): (i) communities described by full-16S were twice more species-diversified than those described by 16SV4-V5, reflecting differences in primer pairs ; (ii) abundances of OTUs based on LR were slightly less variable within biological replicates than those based on SR. The present work suggests that LR can be used for characterizing bacteria communities by metabarcoding environmental samples on a Nanopore sequencing device.

Similar site- and sea-orientation patterns, based on short- or long-reads

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 observation to communities described from environmental samples processed with the same workflow as for the mock. This assumption of a correct taxa detection in spite of abundance discrepancies may explain differences observed in top-20 bacterial taxa influencing structures (Fig. 3a), and is reinforced by the relative orientation of samples, preserved between the two sequencing workflows on the same ordination (Fig. 3b).

On the other hand, it is noteworthy that LR communities contained twice more species than SR ones 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 LR did not significantly changed this ordination. In another 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.

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-16S PacBio dataset, indicating that a significant proportion of the taxa that were identified by full-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, albeit not figuring if they were true taxa or not. Our study confirmed that there were much more taxa detected by full-16S than by 16SV4-V5, but also that a certain proportion of taxa sequenced with 16SV4-V5 were not detected with full-16S dataset (30.3% of the species and 13.3% of the genus). However, this observation is much tempered when the proportion of reads involved in these non-detected taxa is considered, concerning 15.3% of the reads for species and only 1.2% for genus.

When considering non-shared taxa, the present study illustrated the assignation power of a longer bacterial 16S rRNA, compared to a restricted 16S V-region, incidentally acknowledged to have the most appropriate cover for bacteria among SR 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 the incomplete databases that are constantly being updated, or 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-16S improved the taxonomic assignment, as it was done with human gut microbial communities (Jeong et al. 2021; Matsuo et al. 2021). However, if genus level is considered as the maximum resolution of 16S sequencing for a correct taxonomic assignment, a proper primers comparison should start from genus toward higher levels. The fact that read assignment was always lower for LR-exclusive taxa probably reflects more the fact that mangrove sediments contain a high diversity of uncultivated microbes with presently unavailable full-16S in reference databases, than a lower sequencing accuracy of Nanopore (and therefore a plausible sequencing-platform effect).

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 of16S RRNA Gene Amplicons Sequenced through the MinIONTM 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. 'MinIONTM 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'. *https://cran.r-project.org/web/packages/vegan/index.html*

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, 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 MinIONTM 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.

Vass, Máté, Eriksson, Karolina, Carlsson-Graner, Ulla, Wilkner, Johan, and Andersson Agneta. 2022. Co-occurrences enhance our understanding of aquatic fungal metacommunity assembly and reveal potential host–parasite interactions. *FEMS Microbiology Ecology*, 98 (11), fiac120, https://doi.org/10.1093/femsec/fiac120

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. Thanks to associate-editor and reviewers of PCI microbiology for their pertinent suggestions.

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 available in NCBI with BioProject accession number PRJNA985243.

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

Supplementary material

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

Figure S1. Mean relative abundance of bacterial phyla obtained with LR on Nanopore ; exclusiveLR phyla are in red.

Figure S2. Contribution of Mean Decrease Gini coefficients of common species (a) and common genus (b) sequenced by short-reads (SR) and long-reads (LR), 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.

Methods and results for archaea

Figure S3. (**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 SR were mentioned but the read coverage by sample was much lower than those for archaeal LR.

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.

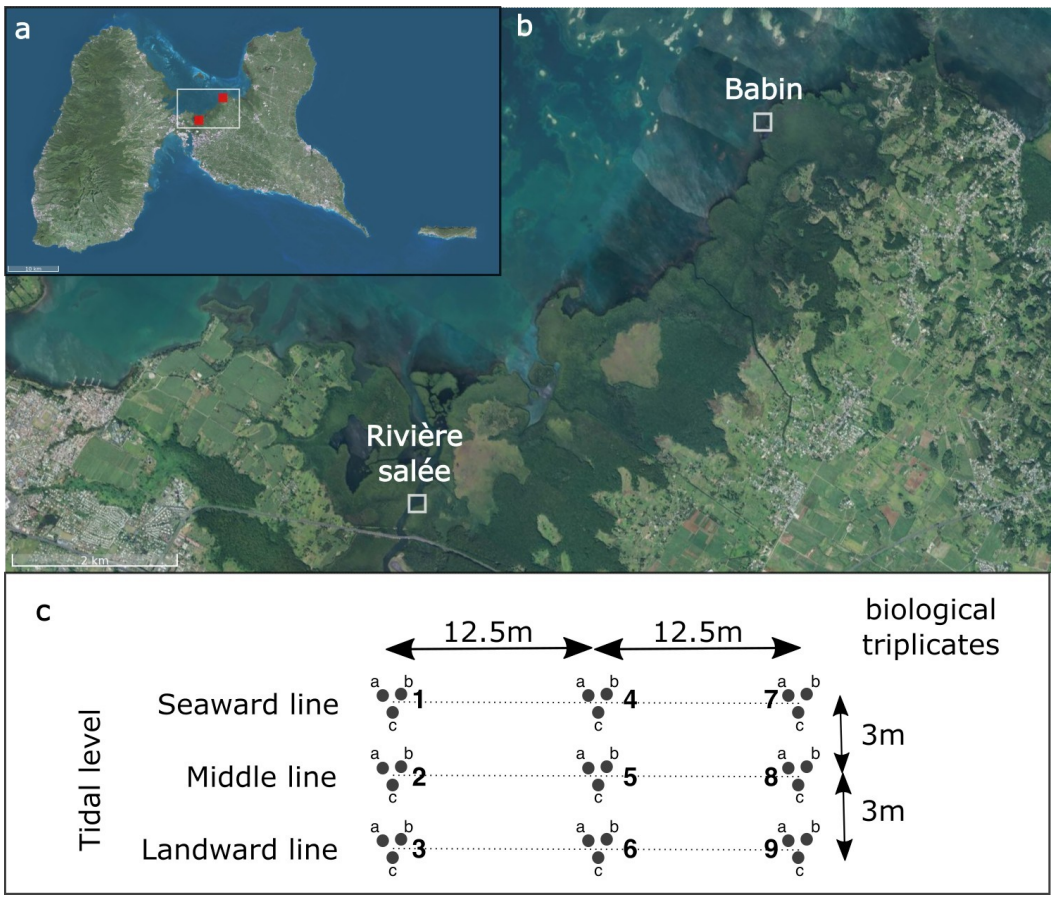


Figure 1. (**a**) Location of sampling sites on Guadeloupe Island (red squares) ; (**b**) zoom on the two sampling sites, with site names ; (**c**) sampling protocol in each site : 3 lines of 3 points, each composed of 3 biological replicates (a, b and c), at 12.5m of distance between each point on each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

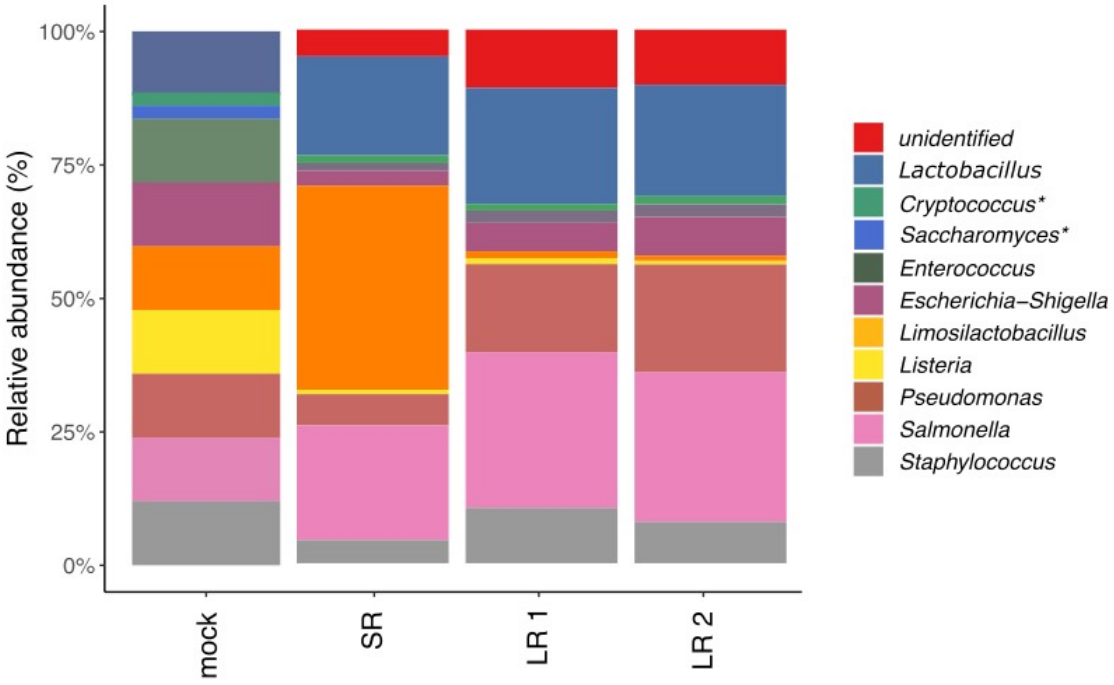


Figure 2. Relative abundances found in a mock community sequenced by SR (1 Illumina flow-cell, 16SV4), LR1 and LR2 (2 Nanopore flow-cells, full-16S), after singletons filtering ; the theoretical abundances are shown at left. *Eukaryotic taxa, which are present in the mock but not supposed to be amplified 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.

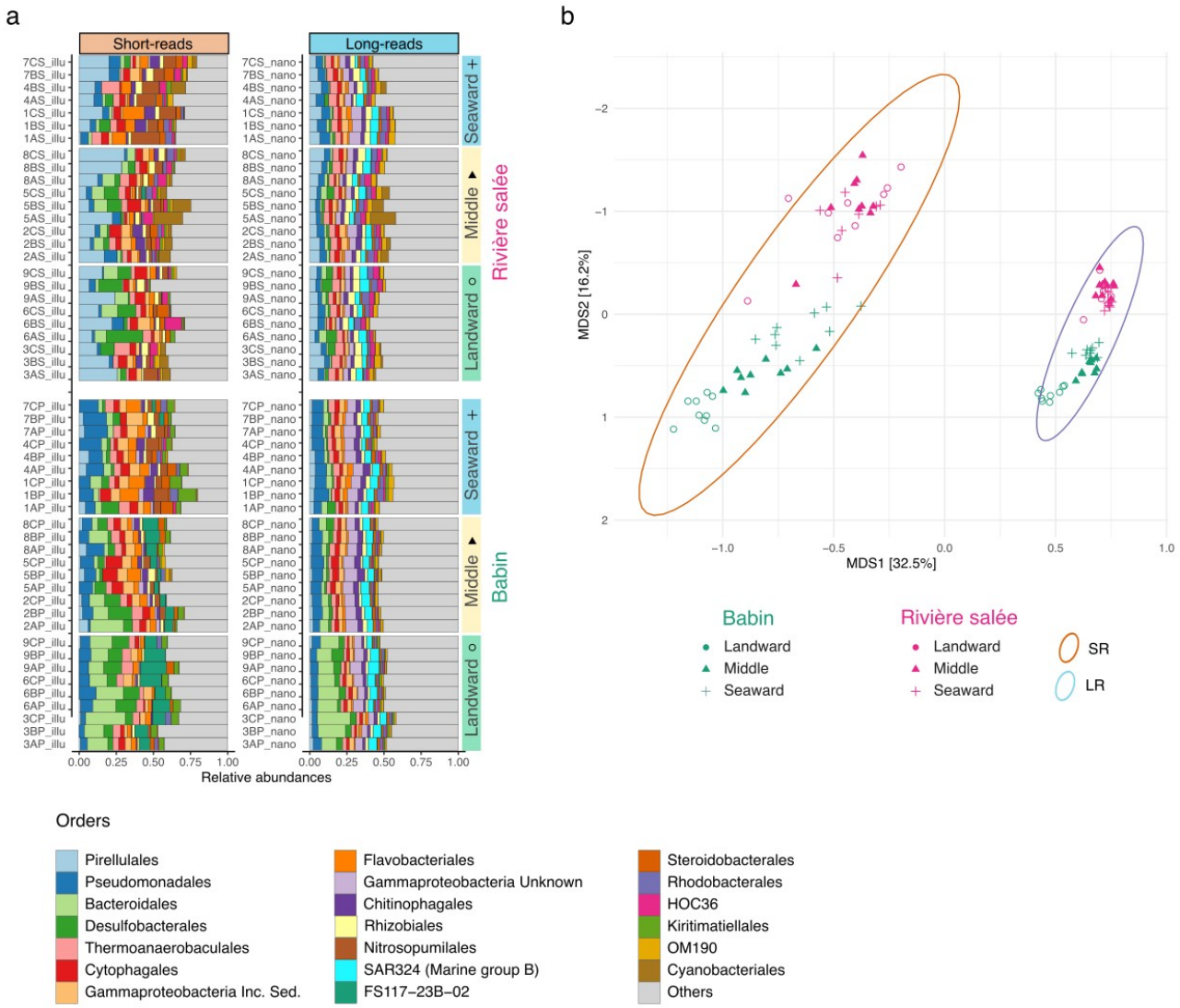


Figure 3. (**a**) Top-20 bacterial orders in samples for both sequencing devices, ranked by their overall relative abundances in samples ; (**b**) biplot of sample scores from a nMDS on abundances of bacterial OTUs agglomerated at genus level, for both sequencing devices (stress=14.1%) ; for this common ordination, shared OTUs were named differently between SR and LR 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).

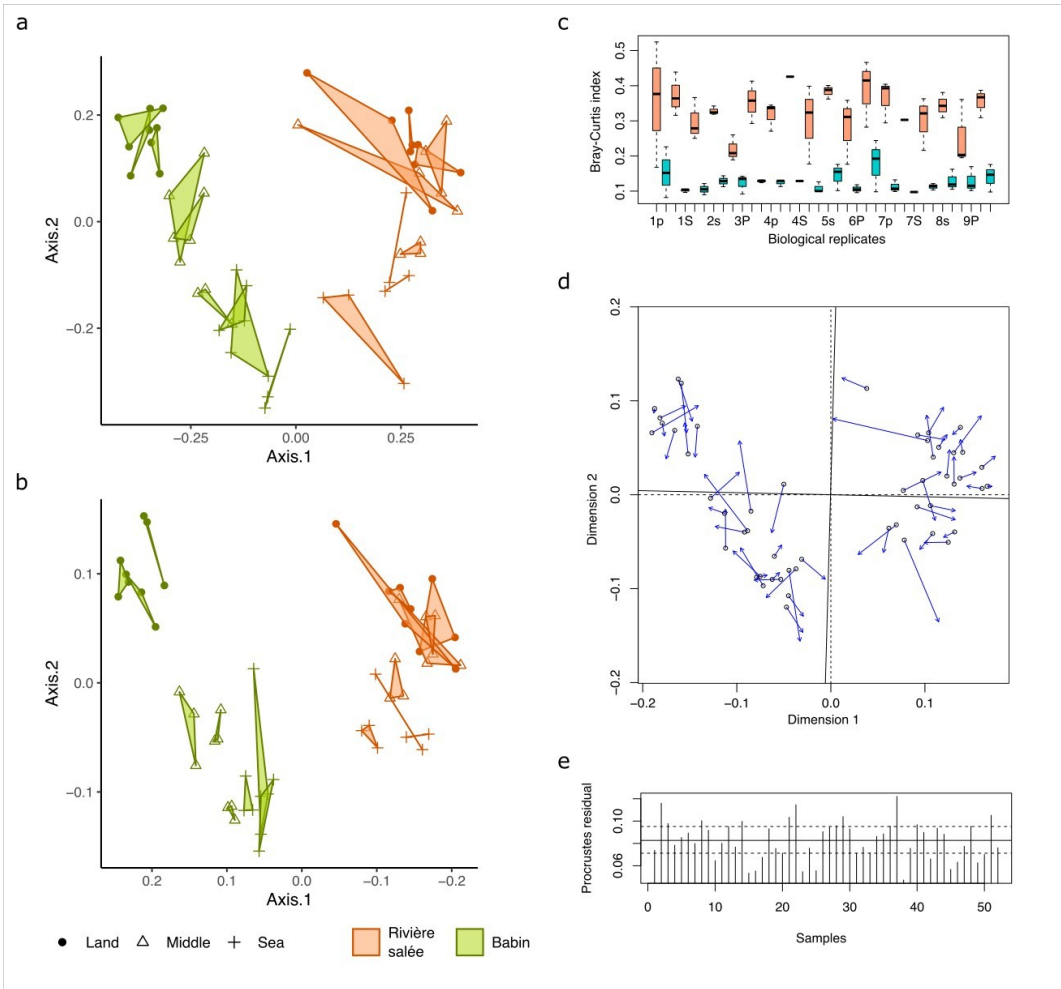


Figure 4. (**a-b**) PCoA on coverage-based rarefied abundances of bacterial communities at species level, (**a**) sequenced by SR, showing biological replicates (polygons) ; (**b**) sequenced by LR ; (**c**) Dispersion of Bray-Curtis dissimilarity index within biological replicates, salmon boxplots for SR, cyan for LR ; 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 LR ordination and arrows point to their positions in the SR 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.

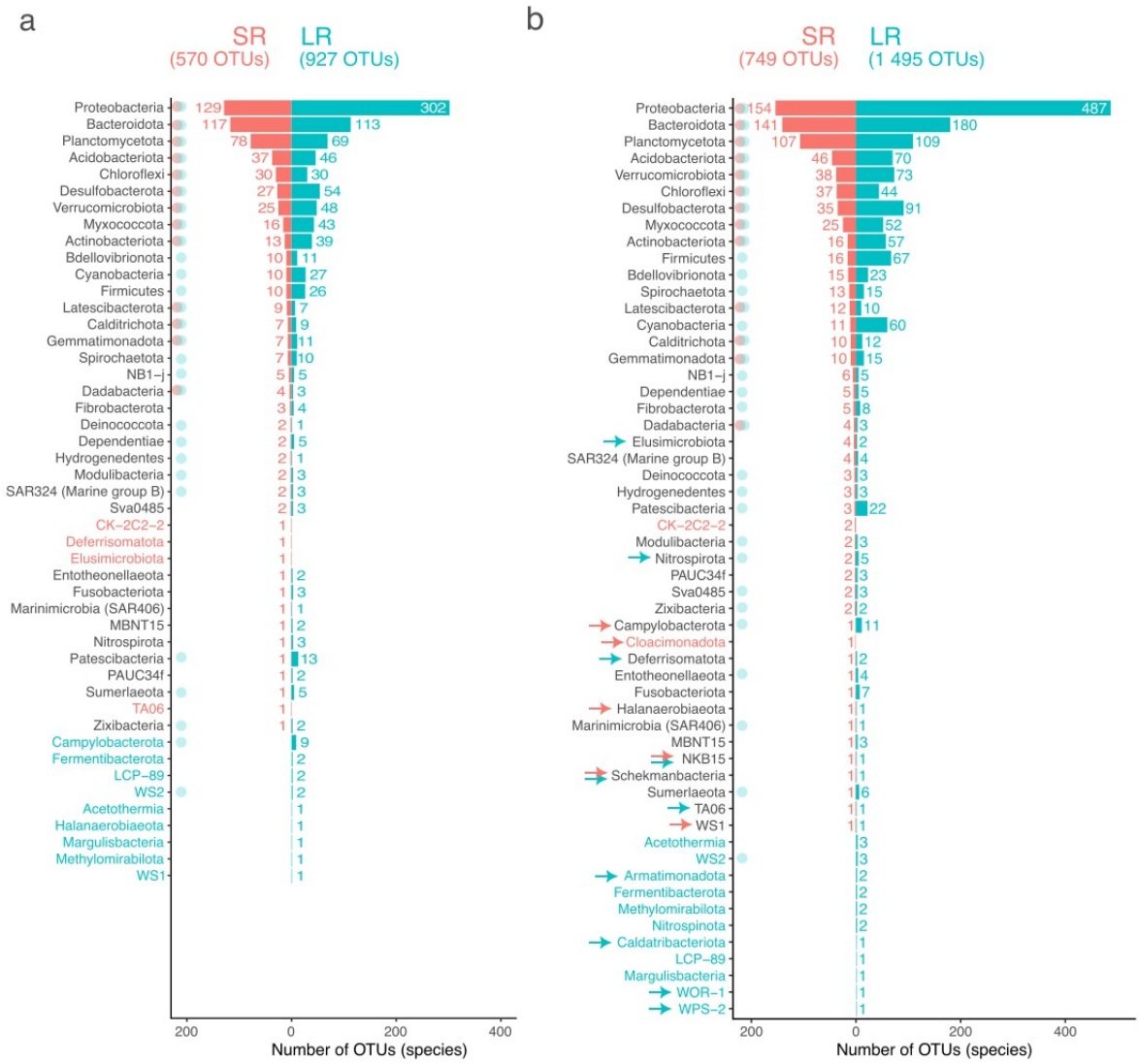


Figure 5. Number of OTUs (97% similarity, singleton-filtered) for each prokaryotic phylum in environmental samples analyzed here (bacteria only), depending on the sequencing strategy : (**a**) with conventional equal-rarefaction (1582 reads for all samples of both strategies, see Methods section for comments on inner bias) ; (**b**) with coverage-based read rarefaction (see Results section for details). Phylum names in red or blue were detected only by SR or only by LR, 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 SR or LR respectively.

Figure 6 (next page). (**a**) Venn diagrams showing the proportions of bacterial taxa shared and unshared between both primer pairs, 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**) On left axis, proportion of bacterial taxa shared between both primer pairs (lines with dots) and proportion of reads assigned to taxa shared between both primer pairs (lines with stars), number of shared and unshared taxa (bars, right axis), at each phylogenetic rank ; more details in Table S1.

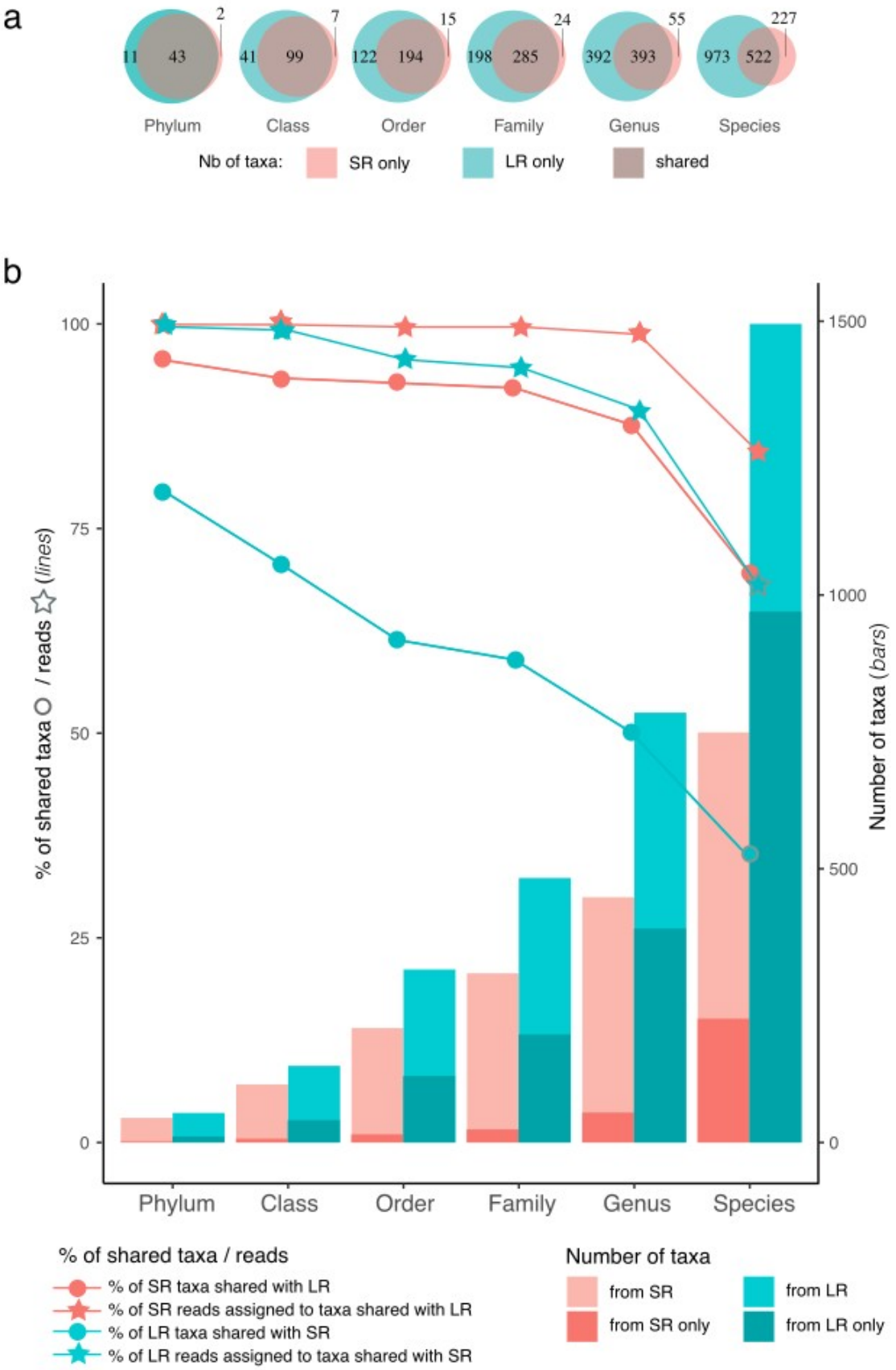
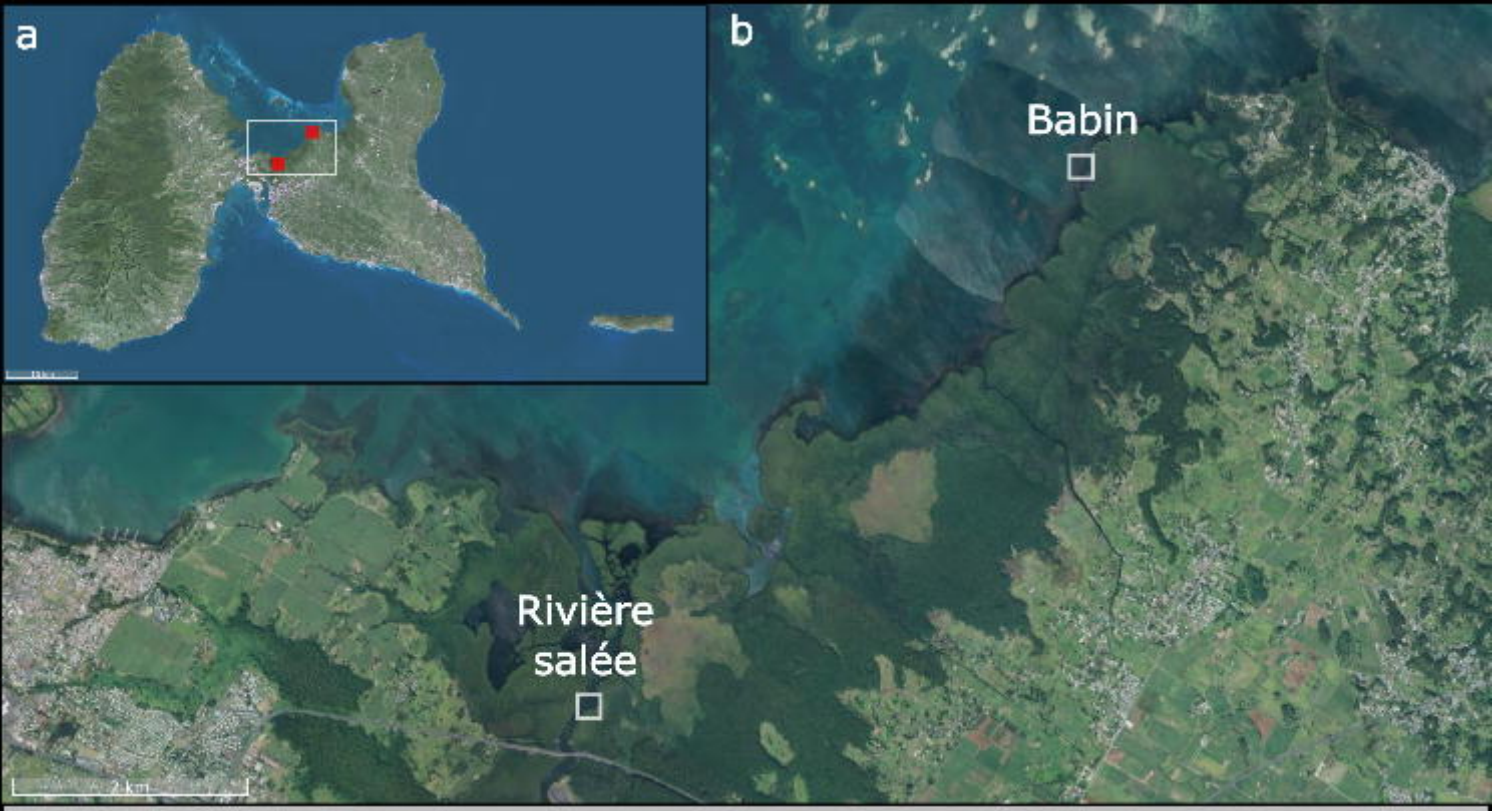
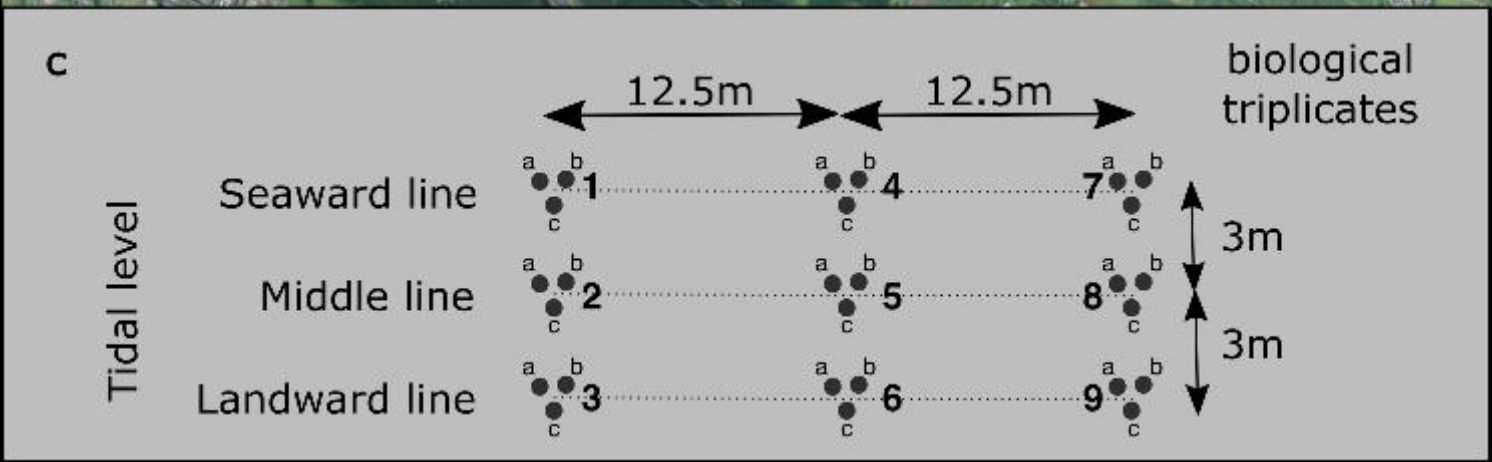

a
2
11
43
7
41
99
15
122
194
24
198
285
55
392
393
227
973
522
Phylum
Class
Order
Family
Genus
Species
Nb of taxa:
SR only
LR only
shared
b
100
75
50
25
0
1500
1000
500
0
% of shared taxa / reads (lines)
Number of taxa (bars)
Phylum
Class
Order
Family
Genus
Species
% of shared taxa / reads
% of SR taxa shared with LR
% of SR reads assigned to taxa shared with LR
% of LR taxa shared with SR
% of LR reads assigned to taxa shared with SR
Number of taxa
from SR
from SR only
from LR
from LR only

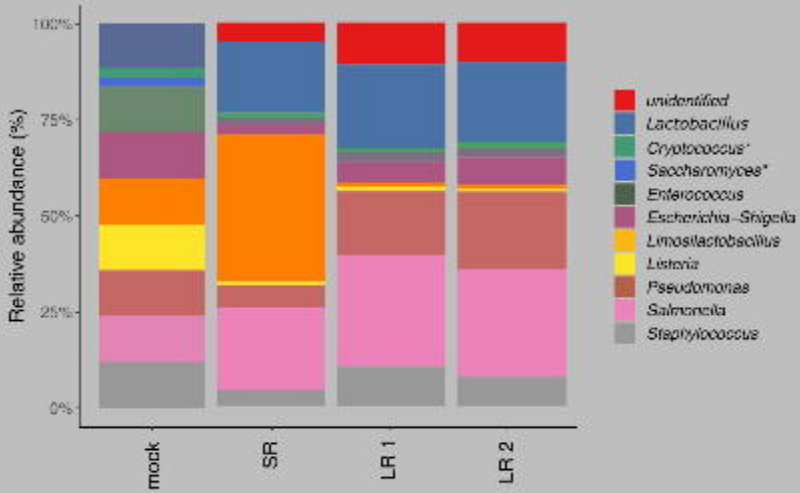
Figure 6.

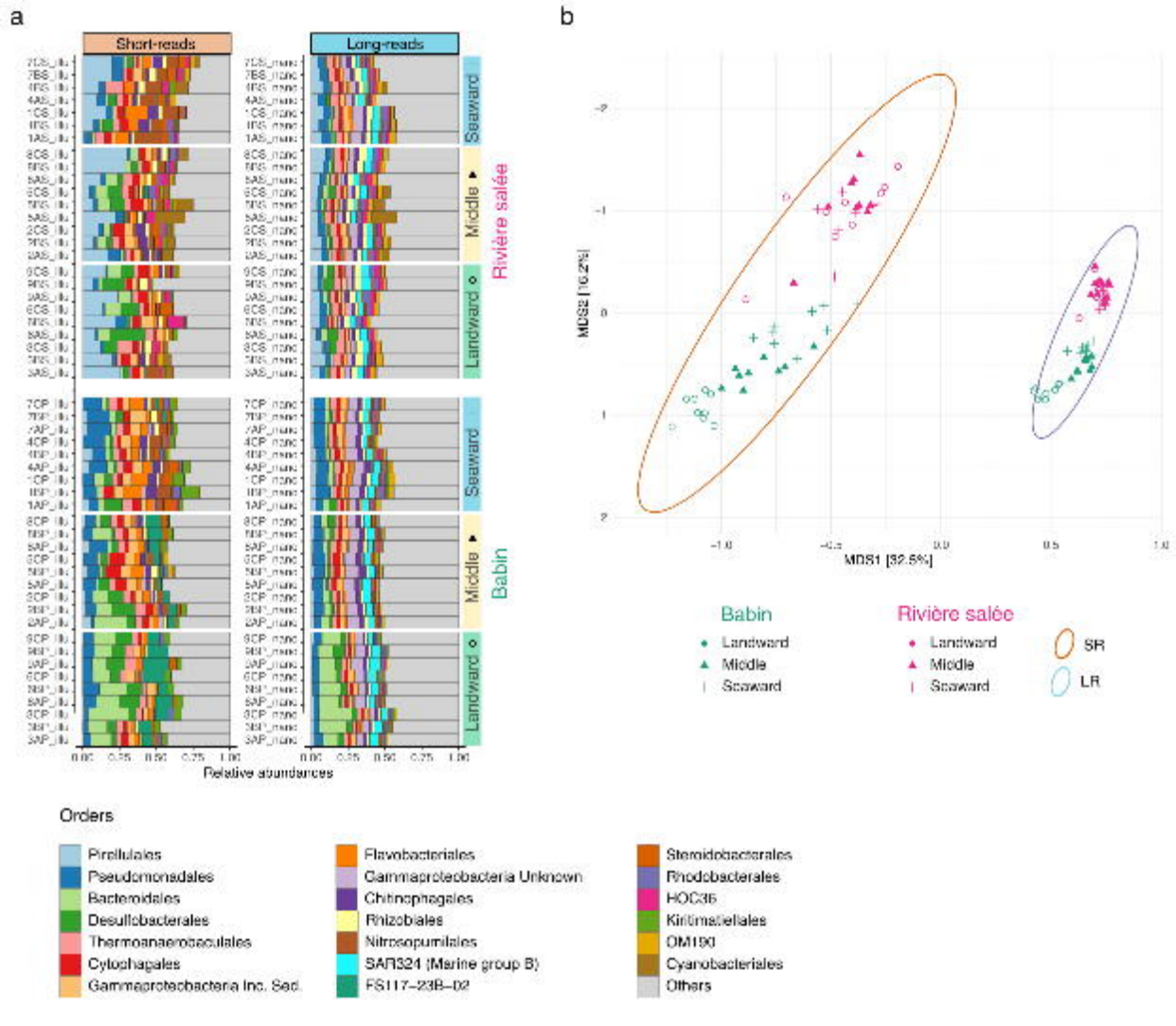
Primers / Sequencer	Taxa / Reads	Phylum	Class	Order	Family	Genus	Species
Short-reads 135 692 reads Illumina (515F + 926R)	**Taxa detected** % taxa assigned % reads assigned	**45** 97.8% 99.6%	**106** 92.5% 99.6%	**209** 91.9% 98.3%	**309** 88.0% 98.0%	**448** 82.4% 94.0%	**749** 65.3% 54.0%
	Taxa exclusive % 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%
	% taxa shared with long-reads % reads with same assignment than LR*	95.6% 99.99%	93.4% 99.97%	92.8% 99.7%	92.2% 99.7%	87.7% 98.8%	69.7% 84.7%
Long-reads 265 650 reads Nanopore (27F + 1492R)	**Taxa 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%
	Taxa exclusive % 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%
	% taxa shared with short-reads % reads with same assignment than SR*	79.6% 99.7%	70.7% 99.3%	61.4% 95.7%	59.0% 94.7%	50.1% 89.6%	34.9% 68.5%
Shared taxa	**Shared taxa** % taxa shared % taxa assigned among shared	**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%

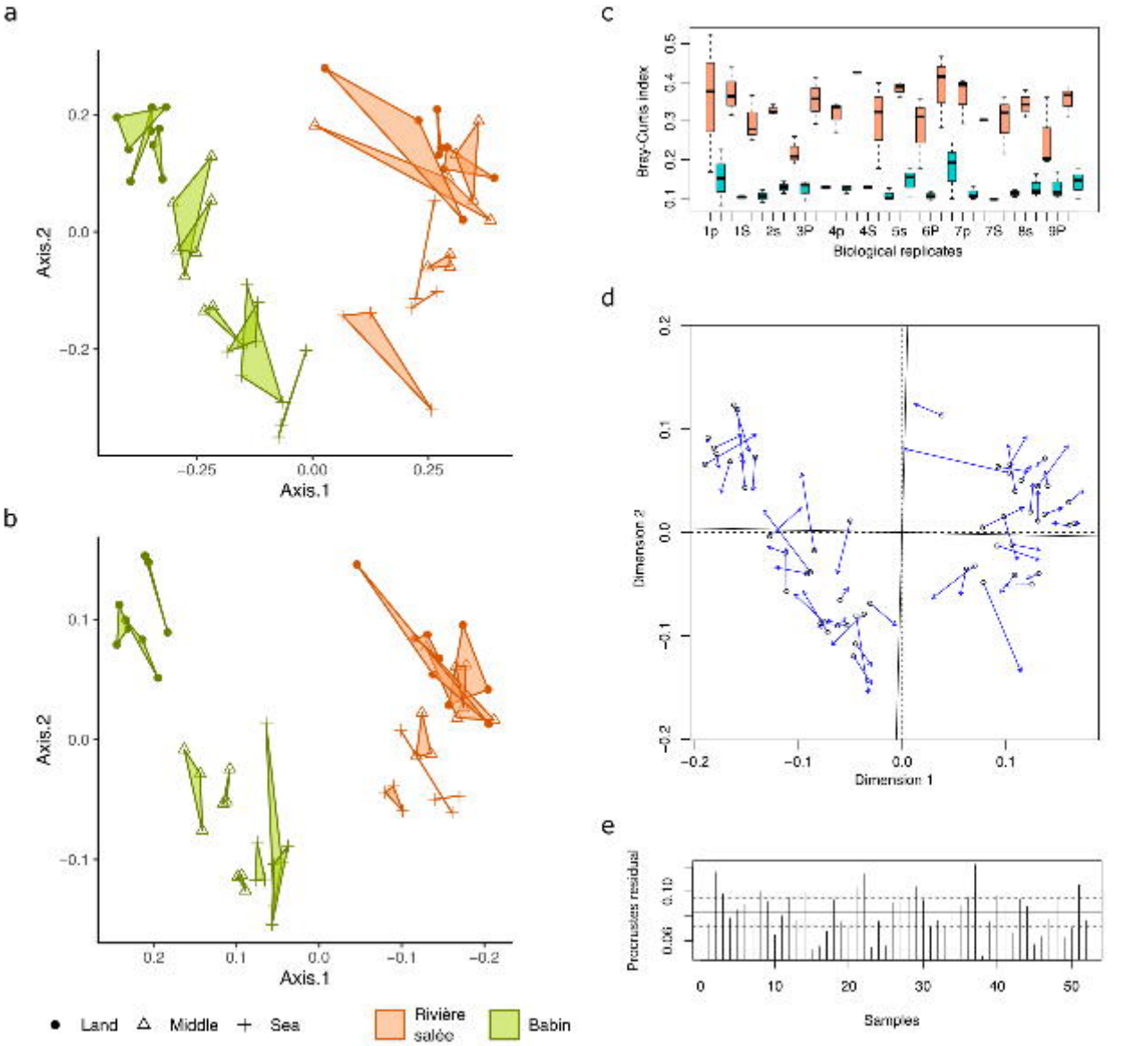
Table 1. Statistics on bacterial taxa detected by SR and LR primer pairs, shared or unshared taxa between primer pairs, by taxonomic rank, percentage of taxa or reads assigned. Here only the bacterial dataset was coverage-based rarefied and singletons filtered. *among taxa assigned at this rank (non-assigned are not counted). In the bacterial communities analyzed, 84.7 and 98.8% of SR were assigned to the same species and genus, respectively, than those detected by LR (in orange). Conversely, 68.5 and 89.6% of LR were assigned to the same species and genus, respectively, than those detected by LR (in blue).

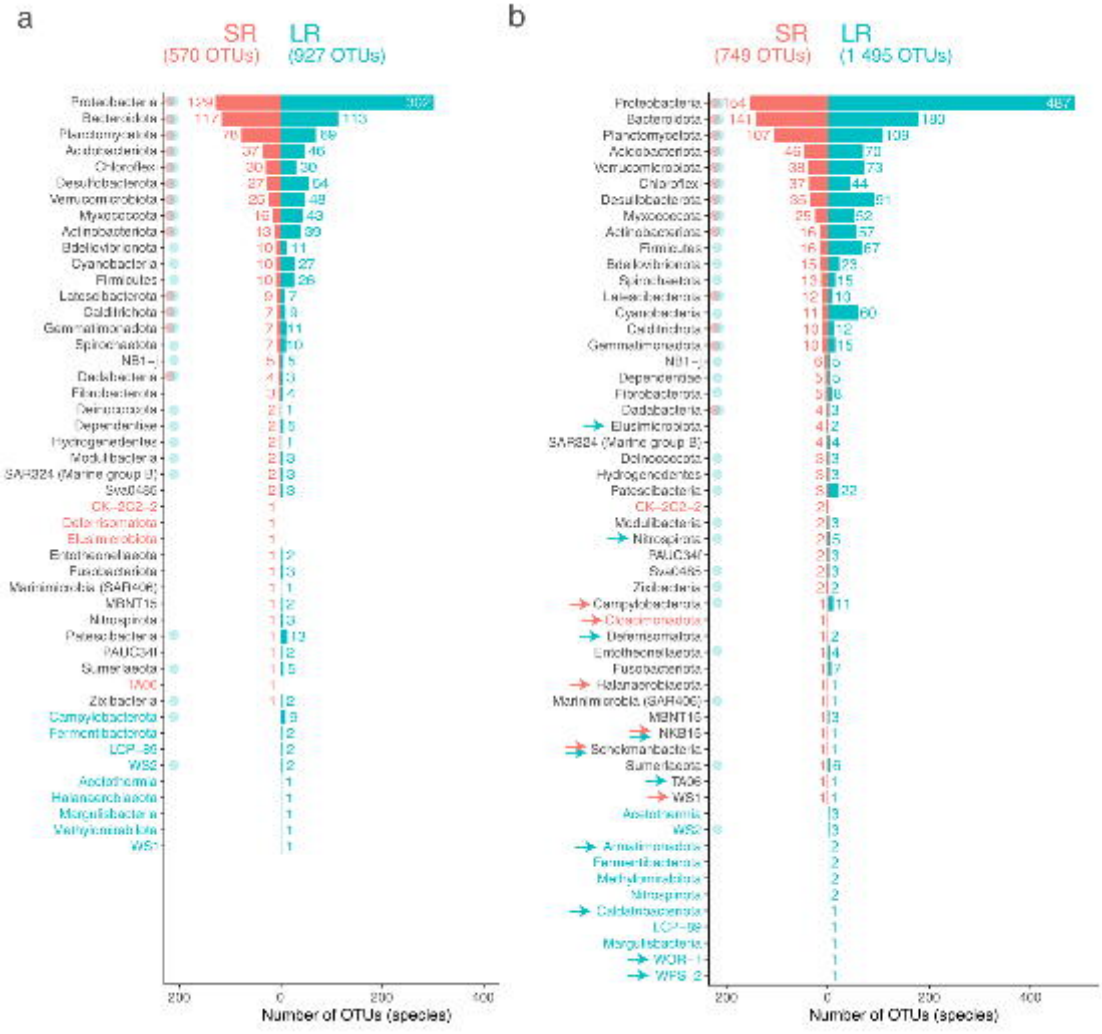

a
b
Babin
Rivière
salée
2 km

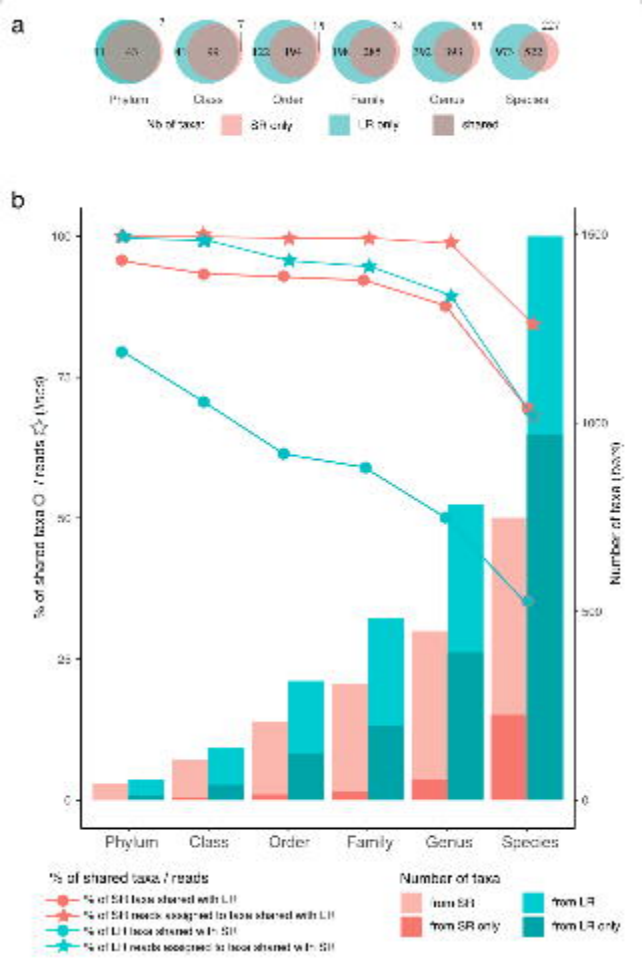

c
12.5m
12.5m
biological
triplicates
Tidal level
Seaward line
Middle line
Landward line
a
b
c
1
2
3
4
5
6
7
8
9
3m
3m


Relative abundance (%)
100%
75%
50%
25%
0%
mock
SR
LR 1
LR 2
unidentified
Lactobacillus
Cryptococcus*
Saccharomyces*
Enterococcus
Escherichia–Shigella
Limosilactobacillus
Listeria
Pseudomonas
Salmonella
Staphylococcus


a
Short-reads
Long-reads
Seaward
Middle
Landward
Rivière salée
Seaward
Middle
Landward
Babin
0.00 0.25 0.50 0.75 1.00
Relative abundances
b
MDS2 [16.2%]
MDS1 [32.5%]
Babin
Landward
Middle
Seaward
Rivière salée
Landward
Middle
Seaward
SR
LR
Orders
Pirellulales
Pseudomonadales
Bacteroidales
Desulfobacterales
Thermoanaerobacullales
Cytophagales
Gammaproteobacteria Inc. Sed.
Flavobacteriales
Gammaproteobacteria Unknown
Chitinophagales
Rhizobiales
Nitrosopumilales
SAR324 (Marine group B)
FS117-23B-02
Steroidobacterales
Rhodobacterales
HOC36
Kiritimatiellales
OM190
Cyanobacteriales
Others


a
Axis.2
Axis.1
b
Axis.2
Axis.1
Land
Middle
Sea
Rivière salée
Babin
c
Bray-Curtis index
Biological replicates
1p
1S
2s
3P
4p
4S
5s
6P
7p
7S
8s
9P
d
Dimension 2
Dimension 1
e
Procrustes residual
Samples


a
SR
(570 OTUs)
LR
(927 OTUs)
Proteobacteria 129 397
Bacteroidota 117 113
Planctomycetota 78 89
Acidobacteriota 37 46
Chloroflexi 30 30
Desulfobacterota 27 54
Verrucomicrobiota 25 48
Myxococcota 16 43
Actinobacteriota 13 39
Bdellovibrionota 10 11
Cyanobacteria 10 27
Firmicutes 10 26
Latescibacterota 9 7
Calditrichota 7 3
Gemmatimonadota 7 11
Spirochaetota 7 10
NB1-j 5 5
Dadabacteria 4 3
Fibrobacterota 3 4
Deinococcota 2 1
Dependentiae 2 5
Hydrogenedentes 2 1
Modulibacteria 2 3
SAR324 (Marine group B) 2 3
Sva0485 2 3
CK-2C2-2 1
Calditenuicollota 1
Elusimicrobiota 1
Entotheonellaeota 1 2
Fusobacteriota 1 3
Marinimicrobia (SAR406) 1 1
MBNT15 1 2
Nitrospinota 1 3
Patescibacteria 1 13
PAUC34f 1 2
Sumerlaeota 1 5
TA06 1
Zixibacteria 1 2
Campylobacterota 3
Fermentibacterota 2
LCP-89 2
WS2 2
Aerophobetes 1
Heimdallarchaeota 1
Margulisbacteria 1
Methylomirabilis 1
WS1 1
200 0 200 400
Number of OTUs (species)
b
SR
(749 OTUs)
LR
(1 495 OTUs)
Proteobacteria 104 487
Bacteroidota 141 180
Planctomycetota 107 109
Acidobacteriota 48 70
Verrucomicrobiota 38 73
Chloroflexi 37 44
Desulfobacterota 35 81
Myxococcota 25 52
Actinobacteriota 16 57
Firmicutes 16 67
Bdellovibrionota 15 23
Spirochaetota 13 15
Latescibacterota 12 13
Cyanobacteria 11 60
Calditrichota 10 12
Gemmatimonadota 10 15
NB1-j 6 5
Dependentiae 5 5
Fibrobacterota 5 6
Dadabacteria 4 3
Elusimicrobiota 4 2
SAR324 (Marine group B) 4 4
Deinococcota 3 3
Hydrogenedentes 3 3
Patescibacteria 3 32
CK-2C2-2 2
Modulibacteria 2 3
Nitrospirota 2 5
PAUC34f 2 3
Sva0485 2 3
Zixibacteria 2 2
Campylobacterota 1 11
Calditenuicollota 1
Deferrisomatota 1 2
Entotheonellaeota 1 4
Fusobacteriota 1 7
Halanaerobiaeota 1 1
Marinimicrobia (SAR406) 1 1
MBNT15 1 3
NKB15 1 1
Schekmanbacteria 1 1
Sumerlaeota 1 6
TA06 1 1
WS1 1 1
Aerophobetes 3
WS2 3
Armatimonadota 2
Fermentibacteria 2
Methylomirabilota 2
Nitrospinota 2
Calditenuicollota 1
LCP-89 1
Margulisbacteria 1
WOR-1 1
WPS-2 1
200 0 200 400
Number of OTUs (species)


a
Phylum
Class
Order
Family
Genus
Species
No. of taxa:
SR only
LR only
shared
b
% of shared taxa / reads
Number of taxa (log10)
Phylum
Class
Order
Family
Genus
Species
% of shared taxa / reads
% of SR taxa shared with LR
% of SR reads assigned to taxa shared with LR
% of LR taxa shared with SR
% of LR reads assigned to taxa shared with SR
Number of taxa
from SR
from SR only
from LR
from LR only