OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae642

https://doi.org/10.1093/bib/bbae642

Problem Solving Protocol

Breaking free from references: a consensus-based approach for community profiling with long amplicon nanopore data

Willem Stock [1,*], Coralie Rousseau[2], Glen Dierickx[3,4], Sofie D'hondt[1], Luz Amadei Martínez[5], Simon M. Dittami[2], Luna M. van der Loos[1], Olivier De Clerck[1]

[1]Phycology Research Group, Ghent University, 9000 Gent, Belgium
[2]Sorbonne University, CNRS, Laboratory of Integrative Biology of Marine Models (LBI2M, UMR 8227), Station Biologique de Roscoff (SBR), Roscoff, France
[3]Research Group Mycology, Ghent university, 9000 Gent, Belgium
[4]Research Unit Forest Ecology and Management, Research Institute for Nature and Forest, Geraardsbergen, Belgium
[5]Laboratory of Protistology and Aquatic Ecology, Ghent University, 9000 Gent, Belgium

*Corresponding author. Willem Stock, +32 9 264 85 07; Krijgslaan 281-Sterre-S8 (Phycology); 9000 Gent, Belgium. E-mail: Willem.Stock@ugent.be

Abstract

Third-generation sequencing platforms, such as Oxford Nanopore Technology (ONT), have made it possible to characterize communities through the sequencing of long amplicons. While this theoretically allows for an increased taxonomic resolution compared to short-read sequencing platforms such as Illumina, the high error rate remains problematic for accurately identifying the community members present within a sample. Here, we present and validate CONCOMPRA, a tool that allows the detection of closely related strains within a community by drafting and mapping to consensus sequences. We show that CONCOMPRA outperforms several other tools for profiling bacterial communities using full-length 16S rRNA gene sequencing. Since CONCOMPRA does not rely on a sequence database for profiling communities, it is applicable to systems and amplicons for which little to no reference data exists. Our validation test shows that the amplification of long PCR products is likely to produce chimeric byproducts that inflate alpha diversity and skew community structure, stressing the importance of chimera detection. CONCOMPRA is available on GitHub (https://github.com/willem-stock/CONCOMPRA).

Keywords: Oxford nanopore technology; amplicon sequencing; chimera; consensus sequence

Introduction

Oxford Nanopore Technology (ONT) allows for the sequencing of much longer amplicons than second generation sequencing platforms. Several studies have attempted to leverage the increased length of marker amplicons to increase the taxonomic resolution at which microbial communities can be characterized. For instance, [1] sequenced a 3.5 Kb and a 6 Kb region, covering multiple rRNA genes and internal transcribed spacer regions, to classify fungi in clinical samples. [2] described a protocol to characterize the bacterial communities associated with seaweeds by sequencing the near full-length 16S rRNA gene.

Despite continuous improvements in sequencing chemistry, flow cells and base calling algorithms, the quality of sequencing data is still markedly lower than that generated by second-generation short-read sequencing platforms. Due to the relatively high number of erroneous bases (modal read accuracy for the R9.4.1 and R10.4 flow cells of >96% and > 99% [3];, the increased amplicon length obtained with ONT sequencing is a challenge in itself and does not necessarily translate in finer phylogenetic resolution. For instance, direct taxonomic assignment of single full-length 16S rRNA sequences at the subgenus level is problematic using established tools such as the EPI2ME platform [4]. One way to increase read accuracy is to ensure that the same DNA strand and its complement or copies, are sequenced multiple times. This can be achieved with unique molecular identifiers that link daughter sequences to the original template DNA [5] or rolling circle amplification [6, 7]. The downside of such approaches is that they significantly complicate workflows, increase costs, and reduce the number of unique reads that can be sequenced within a run.

Given these complications, researchers have focused on developing post-sequencing solutions to achieve species-level resolution with ONT amplicon data. Emu [8] employs an expectation–maximization algorithm to provide species-level taxonomic assignments from full-length 16S rRNA reads. NanoCLUST classifies consensus sequences rather than individual reads [9]. These tools, as well as other commonly used tools for taxonomic assignment of ONT amplicon sequencing data, including Kraken2 [10], require a reference database such as SILVA [11], UNITE [12], or PR2 [13] to function. However, the databases available to date are far

Received: June 28, 2024. **Revised:** October 30, 2024. **Accepted:** December 1, 2024

from exhaustive and often lack strain or even species-level information, hampering finer taxonomic assignment of long amplicon ONT sequencing with the available tools. This is, for instance, the case for marine bacterioplankton, where many oligotrophic bacteria are challenging to isolate [14, 15]. Moreover, databases are only available for a few loci (ITS, rRNA genes), and might not be suitable when working on an understudied group or in unexplored habitats.

As with other sequencing technologies, ONT-generated sequences can be chimeric. Chimeras in amplicon sequencing data can easily make up one third of the data [16] although the frequency of chimeric reads depends on the sequencing protocol (i.e., primers used, PCR conditions) as well as the sample composition itself [17, 18]. Chimeras are formed during the PCR as well as post-amplification, during the ligation [19]. The amplification of a longer region generally produces more by-products [20] as template switching during DNA synthesis is more likely to occur. Mapping of reads with chimeric sequences still present is likely to result in false positive detection of species and inflate diversity estimates. Several tools have been developed to detect and remove chimeric reads in ONT sequencing data, e.g., Liger2LiGer [21], https://github.com/rlorigro/Liger2LiGer, but these are originally not designed to handle amplicon data and are computationally demanding as they require all-against-all mapping of reads.

Here, we present the tool CONCOMPRA, CONsensus approach for COMmunity Profiling with nanopoRe Amplicon sequencing data. This tool creates a consensus sequence database followed by abundance profiling of amplicon ONT sequencing data. Our workflow is designed to be fast, customizable (implemented in bash and python) and flexible (working from one-to-many samples). We show that this workflow, using full-length 16S rRNA gene data, allows for the discrimination of closely related species and works with long amplicons for which no suitable reference database is available.

Methods

CONCOMPRA workflow

The input data consists of basecalled ONT sequencing reads in fastq format. CONCOMPRA (Fig. 1) processes all fastq files present in the appointed working directory. On a file-per-file basis, a list of consensus sequences is generated from the reads according to the following steps. Reads outside of the user-defined length window (based on the expected length distribution of the amplicon) are discarded. Forward reads with primers at the expected positions are identified using primer-chop (https://gitlab.com/mcfrith/primer-chop), which also trims the primers and bases preceding the forward primer and trailing the reverse primer. Simultaneously, the mapping of the primers to the sequences is used to estimate the rates (probabilities) of insertion, deletion, and substitution in the sequencing data [22], which are required for the LAST alignment algorithm [23] implemented in primer-chop and our consensus generation approach. To prevent highly erroneous reads from influencing the downstream processing, only the top 80% of the forward, trimmed reads, with the fewest expected errors are retained using Filtlong (https://github.com/rrwick/Filtlong). These sequences are used for unsupervised clustering, similar to the NanoCLUST approach. The reads are clustered with UMAP-OPTICS [24, 25] based on their 3mer (substrings of length 3) composition. A consensus sequence is produced from a subset of the reads (40 by default, user defined parameter) within each cluster using lamassemble [26] with the previously estimated insertion, deletion, and substitution rates. Potentially chimeric sequences are flagged using the vsearch uchime_denovo algorithm, taking into account the number of reads assigned to the cluster from which the consensus sequence was generated. Since this chimera detection step is performed for each sample, these chimeric consensus sequences are referred to as 'local chimeric sequences'.

Vsearch [27] is used to deduplicate sequences and, optionally, remove highly similar consensus sequences coming from the different samples. An abundance table of the consensus sequences across samples is created by mapping the reads from the original fastq files to the consensus sequences with minimap2 [28]. Reads that fail to map to the consensus sequences are written to a separate 'unmapped' fastq file. Consensus sequences that are deemed chimeric based on the number of reads mapped to them across samples (vsearch uchime_denovo) are considered 'global chimeric sequences'. The consensus sequences that have been flagged as local chimeric sequences or global chimeric sequences are removed in order to obtain a final, chimera-free consensus sequence table.

Datasets

CONCOMPRA was compared to alternative tools on three datasets: (1) the 16S rRNA gene sequences from a synthetic bacterial community composed of 20 equally abundant bacteria (16S MOCK), (2) the previously published 16S rRNA gene sequences from the Gut Microbiome Standard, a microbial community mimicking the human gut microbiome (Gut Microbiome Standard), and (3) low-quality 16S rRNA gene sequences from natural planktonic bacterial communities (16S natural).

16S MOCK: The full-length 16S rRNA genes from 1 μl of a synthetic bacterial community, consisting of evenly mixed genomic DNA from 20 bacterial strains, (ATCC MSA-1002), was amplified with the primer set 27F_Bctail-FW (TTTCTGTTGGTGCTGATATTGC_AGAGTTTGATCMTGGCTCAG) and 1492R_Bctail-RV (ACTTGCCTGTCGCTCTATCTTC_CGGTTACCTTGTTACGACTT) using the Phire Tissue direct PCR Master Mix (Thermo Fisher) as previously described [2]. Approximately 100 ng DNA (Qubit, ThermoFisher, Massachusetts, United States) of each PCR product was barcoded using the Oxford Nanopore PCR Barcoding Expansion Pack 1–96 (EXP-PBC096) and pooled equimolarly with various other samples. The pooled DNA was purified using the Agencourt AMPure XP system. Amplification and barcoding of the synthetic bacterial community DNA extract was performed four times independently, resulting in four different barcoded PCR products. The library was prepared for two different nanopore flow cell chemistries (two barcoded PCR products each), R9.4.1 and R10.4.1 flow cells, using the ONT SQK-LSK109 and SQK-LSK114 (V14) ligations kits respectively, with the 'Amplicon by Ligation' protocol. Approximately 1 μg of pooled and purified DNA was used to prepare the library. 560 ng of the library were loaded on a flow cell. Sequencing on the R9.4.1 flow cell was halted after 48 h and on the R10.4.1 flow cell after 72 h. The neural network-based tool Guppy [29] was used for super-accurate basecalling (v6.3.8 for the R9.4.1 flow cell data; v6.5.7 for the R10.4.1 flow cell). The minimum data quality threshold was set to Q10. Data quality was assessed with NanoPlot v1.42.0 [30]. The first 10 000 reads of each sequence file were blasted (BLAST 2.2.29+) against the reference sequences (retrieved from the genomes provided by ATCC) to estimate sequence similarity of the data to the ground truth. An Illumina MiSeq system (300 bp paired end) sequenced dataset, restricted to the V1-V3 region of the same mock community, was generated using the pA

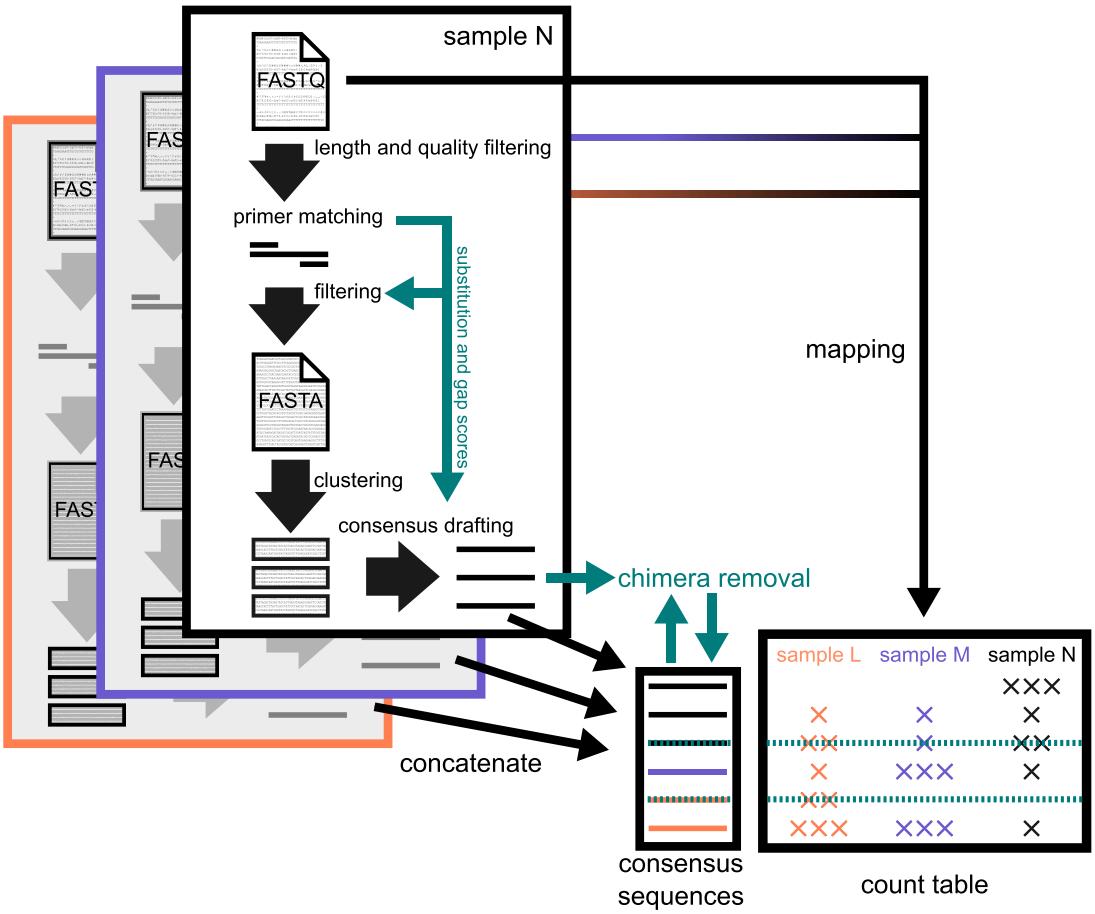

Figure 1. CONCOMPRA workflow. Consensus sequences are generated per sample and then concatenated across samples. Sequences are mapped to the consensus sequence table to generate a count table with the number of sequences mapped to each consensus sequence per sample. Chimeric sequences are detected and removed.

(5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGTTTGATCCTGGCTCAG-3′) and BKL1 primer pair (5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTATTACCGCGGCTGCTGGCA-3′) as in [NO_PRINTED_FORM] [31].

Gut Microbiome Standard: A microbial mock community, produced by Zymo Research (California, United States). The nearly complete 16S rRNA gene sequences (Sequence Read Archive: ERR10318834) were previously generated and published by [32] The community is comprised of 21 different strains to mimic the human gut microbiome, including three non-bacteria. The theoretical relative abundance of the members in the mock ranges from 14% to 0.0001%. Five different strains of *Escherichia coli* are present within the mock, each at a theoretical relative abundance of 2.8%. There were minor differences in the forward primer (5′-TTTCTGTTGGTGCTGATATTGC-AG**R**GTT**Y**GAT**Y**MTGGCTCAG-3′, differences in bold) as well as the PCR mix and protocol used by [32] compared to the ONT 16S rRNA gene data generated in this study. The library containing the 16S rRNA gene PCR product from the mock and various other samples, was prepared using the Nanopore Ligation Sequencing Kit SQK-LSK112 sequenced on an R9.4.1 flow cell.

16S natural: Three independent water samples (50-100 ml), collected in different months at different locations in the Belgian part of the Schelde river were filtered through a 0.22 μm MF-Millipore Membrane filter (Merck Millipore). DNA was extracted using the Dneasy PowerLyzer Microbial Kit (Qiagen). For Illumina sequencing, the V1-V3 region of the 16S rRNA gene was amplified and sequenced as described above. The full-length 16S rRNA gene for the three communities was amplified and sequenced on a R9.4.1 flow cell as described for the synthetic bacterial community. Base calling was performed using Guppy v 4.3.4. Since the resulting basecalled data was of low quality (less than 3% of the data was above Q10 quality threshold that was used for the other datasets; Table 1), no quality threshold was set for the 16S natural sequencing data prior to using it for community profiling.

Data analysis

CONCOMPRA was run on the datasets with sequence length window set to 1400–1700 bp for the inhouse sequenced samples and 1500–1800 bp for the Gut Microbiome Standard sample, based on the mean expected sequence length of the amplicons (including flanking sequences). The mock samples were analyzed individually, and the natural samples were analyzed in batch. Taxonomic annotations of the CONCOMPRA-generated consensus sequences for the 16S MOCK and 16S natural datasets were obtained using the IdTaxa function from the DECIPHER package v2.28.0 [33]with the SILVA_SSU_r138 reference database [34]. NGSpeciesID v0.3.0 [35]and amplicon_sorter (version Feb 20, 2024) [36] were used to compare the quality of the obtained consensus sequences obtained for both mock datasets. NGSpeciesID was run with an abundance ratio threshold of 0.002 and Racon as polisher with three iterations. Chopper v0.7.0 [30] was used to retain only ≥ Q12 reads prior to using amplicon_sorter setting the same size window as above. Consensus sequences generated by

Table 1. Summary of the ONT 16S rRNA gene data used in this study.

sample ID	sequencing chemistry	read count	mean read quality	Mean read length
16S_R9_1	R9	41,195	13	1544.5
16S_R9_2	R9	119,342	12.8	1576.5
16S_R10_43	R10	51,157	15.4	1556.4
16S_R10_60	R10	171,686	15.3	1538.7
Gut Microbiome Standard	R9	181,006	16.2	1584.2
natural sample (id4)	R9	25,441	7.8	1885.9
natural sample (id26)	R9	48,129	7.8	1996.6
natural sample (id47)	R9	33,823	7.8	1949.3

CONCOMPRA, NGSpeciesID, and amplicon_sorter were compared to the reference sequences using BLAST 2.2.29+. Only reference sequences with a > 95% match to a consensus sequence were considered to be represented by the consensus sequences. The F1 score [37] was calculated to compare the performance of the three consensus sequence building approaches, where true positives (TP) were defined as the number of consensus sequence that uniquely matched a reference sequence with >99% sequence identity. False negatives were defined as reference sequences that did not have a uniquely matching (>99% sequence identity) consensus sequence, and false positives as the total number of consensus sequences minus TP. The F1 scores were compared between consensus sequence building approaches across the four 16S mock communities using a Repeated Measures ANOVA (ezANOVA, ez v4.4) followed by a pairwise t-test with Holm-Bonferroni p-value correction for multiple testing.

The 16S MOCK data and 16S natural data were also analyzed with the reference-based tools Emu v3.4.5 and Kraken2 v2.1.2. SILVA_SSU_r138 was set as reference database. The Gut Microbiome Standard was only analyzed with Emu. The 16S MOCK data was analyzed with both of these tools (1) as is, (2) after length (1400–1700 bp) filtering with Chopper, and (3) after removal of putative chimeric sequenced regions (scrubbing) with yacrd v1 [21] on an all-to-all file generated by mm2-fast v2.24 [38]. As scrubbing failed for one of the R10 datasets, this single sample was excluded from the calculation of filtering and scrubbing statistics.

The Illumina data was processed with DADA2 (v1.28.0) [39]. The primer regions as well as the trailing ends of low quality were trimmed prior to applying the default DADA2 workflow (v1.16). The IdTaxa function from the DECIPHER package v2.28.0 with the SILVA_SSU_r138 reference database was used for taxonomic annotation of the sequence variants (SVs).

To compare the taxonomic annotations between approaches, taxonomic tables were aggregated at the genus level using Phyloseq (1.44.0; R 4.3.1) [40]. Only bacterial reads were retained in the samples by removing reads assigned to mitochondria, chloroplasts, or not assigned to the domain Bacteria. Reads that could not be annotated at the phylum level were also removed. The accuracy of the different methods was quantified by summing the absolute values of the observed relative abundances minus the expected relative abundances across genera.

To compare taxonomic resolution in the absence of a predefined taxonomic framework, the OTUs/SV assigned to the understudied NS11–12 marine group (Sphingobacteriales) within the natural 16S dataset were compared. An approximately-maximum-likelihood phylogenetic tree (FastTree 2.1.11) [41] was constructed from the OTU/SV sequences, which were aligned with SSU-ALIGN 0.1.1 [42] *Sphingobacterium multivorum* OM-A8 (ENA accession: AB020205) was used as outgroup for rooting.

Results

16S mock community

The reads, generated by ONT sequencing (Table 1), had a median sequence similarity to the reference sequences of 92.0% and 94.4% for the R9 and R10 datasets respectively. The number of non-chimeric consensus sequences produced by CONCOMPRA was close to the number of bacteria present in the mock (19.5 ± 2.4 sd) and the average sequence similarity between the consensus sequences and the best-matching genome-derived 16S rRNA gene sequences (https://www.atcc.org/products/msa-1002) was ≥99.9% for the four samples (Fig. 2). Over half of the reads were assigned to chimeric OTUs (59.3% ±4.5% sd), which is in line with the relatively low similarity of the reads to the actual reference sequences (Fig. 2).

The consensus building tools NGSpeciesID and amplicon_sorter yielded 77.3 (±12.1 sd) and 36.5 (±8.3 sd) consensus sequences, respectively. Both tools were consistently less accurate than CONCOMPRA, obtaining an average sequence similarity of 99.1% (±0.3 sd) and 99.5% (±0.3 sd; Fig. 2). The F1 score was significantly higher for CONCOMPRA (0.53 ± 0.3 sd) than the other two methods (amplicon_sorter: 0.35 ± 0.03 sd; NGSpeciesID: 0.25 ± 0.03 sd) (ezANOVA & paired t-tests: $P < 0.01$; Supplementary Fig. 1), mainly due to a higher precision (fewer false positives). CONCOMPRA was also the only tool that obtained consensus sequences that were identical to the reference sequences. This was the case for over half the matching sequences (Fig. 3).

No consensus sequences with more than 95% similarity were recovered for *Bifidobacterium adolescentis* or *Schaalia odontolytica* in any of the tests, suggesting that these 16S rRNA gene sequences were not efficiently amplified (Fig. 3). In the R10 datasets, *Cutibacterium acnes* was missing once in a CONCOMPRA consensus sequence set and *Bacillus pacificus* was missing in an NGSpeciesID consensus sequence set. More bacteria remained undetected in the R9 datasets: *Cutibacterium acnes* was absent in both CONCOMPRA sets and *E. coli*, *Enterococcus faecalis,* and *Streptococcus agalactiae* were each absent once in the CONCOMPRA sets. *Bacillus pacificus* was missing in both amplicon_sorter sets, and *Cereibacter sphaeroides* in one NGSpeciesID set. The two *Streptococcus* strains were distinguished by all tools. CONCOMPRA was the only tool that consistently distinguished the *Staphylococcus* strains, although the consensus sequence for the *S. agalactiae* in one of the R9 sets was only 95% identical to the reference sequence.

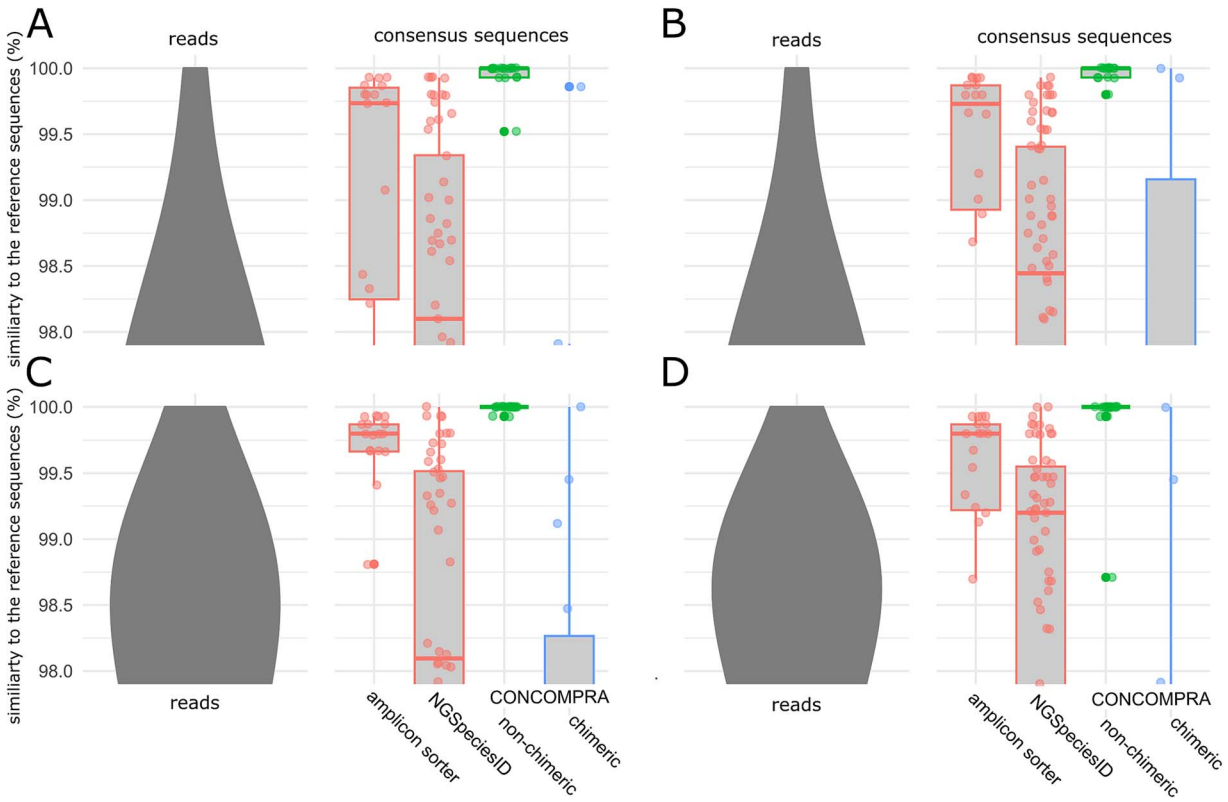

Figure 2. Similarity of the sequencing data and the consensus sequences to the reference 16S mock community. The blast-based percentage similarity to the references is shown for the reads themselves (violin plots) and generated consensus sequences (box plots). The similarity to the references is shown for each 16S mock sample in a different panel: A = 16S_R9_1; B = 16S_R9_2; C = 16S_R10_43; D = 16S_R10_60. The reads (dark grey; left side of each panel) from the R10 sets (bottom row, C & D) were markedly more similar to the reference sequences (>20% of the reads were at least 98% similar to the references) than the reads from the R9 sets (>5% of the reads were at least 98% similar to the references; top row, A & B). The alternative consensus building tools (red) generated more consensus sequences, but these were less similar than the non-chimeric consensus sequences generated by CONCOMPRA (green). The chimeric consensus sequences from CONCOMPRA were the least similar (97% on average) to the reference sequences.

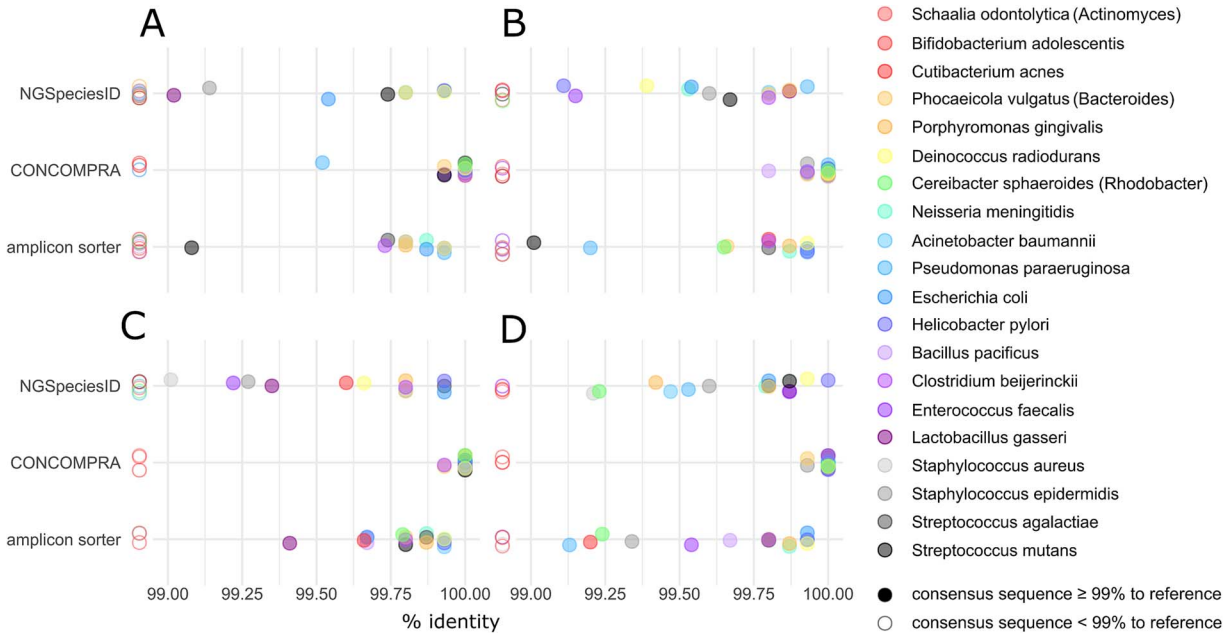

Figure 3. Detection of the 20 different species in the bacterial mock community. The highest similarity to any of the 16S rRNA gene copy sequences to each species (colors) is shown for the different consensus building tools (rows) for the different 16S mock datasets. Only non-chimeric sequences are considered for CONCOMPRA. Species for which no consensus sequences with an identity above 99% was present are shown as hollow circles to the left of the axis. The similarity to the references is shown for each 16S mock sample in a different panel: A = 16S_R9_1; B = 16S_R9_2; C = 16S_R10_43; D = 16S_R10_60.

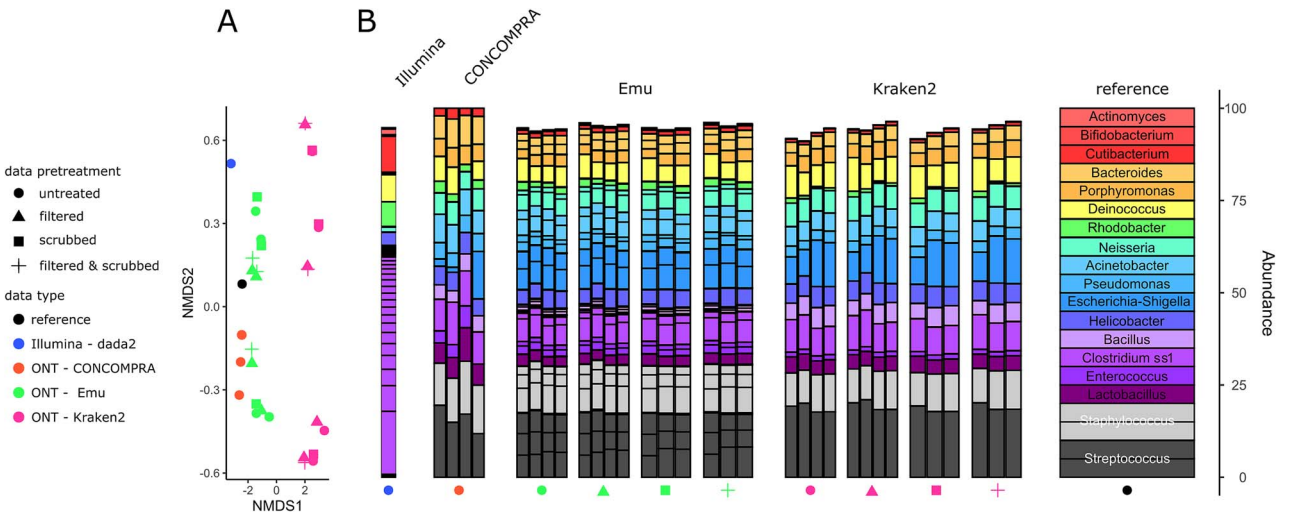

Figure 4. Inferred composition of the mock communities. The beta diversity is visualized through non-metric multidimensional scaling (NMDS; A) using Bray–Curtis dissimilarity on the relative abundances of the genera (2D stress = 0.037). Colors are used to indicate the data types and pipelines used and the shape reflects the data treatment. The community composition (B) differed most between the Illumina data and the ONT data. Only the genera present in the reference community are shown in the bar plots. Nomenclature from the SILVA_SSU_r138 reference database is used.

The two alternative, reference-based community profiling tools returned a higher number of OTUs for the mock communities than CONCOMPRA. Kraken2 predicted 500 (±78.7 sd) OTUs, and Emu predicted 80.7 (±14.2 sd) OTUs. Filtering out the sequences outside the size window removed ∼ 16% of the sequences and reduced the predicted number of OTUs (Kraken2: 339.3 ± 48.3 sd; Emu: 68 ± 11.4 sd). Scrubbing the reads removed 0.2% of the reads, but did not reduce the number of predicted OTUs (Kraken2: 505 ± 76.6 sd; Emu: 83.3 ± 13.4 sd).

Community profiles remained highly similar, regardless of data treatment (Fig. 4). The Illumina data, processed with the DADA2 pipeline, yielded 185 sequence variants (SVs) and was distinct from any of the ONT datasets (Fig. 4). The Illumina data were highly skewed, *Clostridium* making up over half the community, while *Streptococcus* and *Staphyloccoccus* were present at less than 1%. CONCOMPRA best approximated the true abundance profile (summed deviation CONCOMPRA = 46.9 ± 6.2% sd; Kraken2 = 52.7 ± 1.6% sd; Emu = 55.7 ± 1.7% sd; Illumina = 139.1%). Filtering out the sequences outside of the set size window did not reduce the deviation of the two alternative community profiling tools (Kraken2 = 53.6 ± 2.5% sd; Emu = 55.7 ± 1.4% sd).

Gut microbiome standard

CONCOMPRA returned 21 non-chimeric consensus sequences for the Gut Microbiome Standard. Twenty were highly similar (99.9 ± 0.1% sd; Supplementary Fig. 2) to the reference sequences. Each of these uniquely matched a single reference sequence. Amplicon_sorter and NGSpeciesID yielded 11 consensus sequences. All of the amplicon_sorter consensus sequences were highly similar (99.6 ± 0.2% sd; Supplementary Fig. 2) to the reference sequences. Nine of the NGSpeciesID consensus sequences were highly similar (99.1 ± 1.0% sd; Supplementary Fig. 2) to the reference sequences. Among the highly abundant taxa (theoretical abundance >1.5%), CONCOMPRA failed to detect *Bacteroides fragilis* and *Faecalibacterium prausnitzii*. Both were detected by the other tools. While all consensus sequence building approaches detected *E. coli*, only CONCOMPRA detected the presence of multiple *E. coli* strains within the mock community (three out of the four detected). Overall, CONCOMPRA had the highest F1 score of the three approaches (0.36 versus 0.14 for NGSpeciesID and 0.22 for amplicon_sorter).

CONCOMPRA and amplicon_sorter generated good consensus sequences for the two taxa with a theoretical abundance of 1.5% (100% and 99.7% similarity of the consensus sequences to the reference sequence for *Akkermansia muciniphila* & 99.9% and 99.8% for *Clostridioides difficille*, respectively; Fig. 5).

All three consensus sequence building approaches failed to generate a sequence for the three taxa with a theoretical abundance below 0.1% in the Gut Microbiome Standard (Fig. 5). Emu detected *Salmonella enterica* (observed rel. Abundance of 0.0002%; theoretical abundance of 0.01%) but also failed to detect *E. faecalis* (theoretical abundance of 0.001%) and *Clostridium perfringens* (theoretical abundance of 0.0001%).

Natural bacterial samples

CONCOMPRA identified 64 nonchimeric OTUs across the three samples, assigned to 15 different bacterial genera. Kraken2 assigned the ONT reads to 1894 different bacterial genera (2585 taxa). This was substantially less with Emu, namely 265 different bacterial genera (506 taxa). Using the paired-end Illumina data, DADA2 identified 948 SVs assigned to 81 different bacterial genera. Although CONCOMPRA identified the fewest genera out of the three ONT-based workflows, 81.7 (± 4.7 sd) % of the assigned reads in CONCOMPRA were assigned to genera that were also found in the Illumina-based DADA2 dataset. This was much lower for the reads analyzed with Kraken2 and Emu (Kraken2: 16.6 ± 2.6% sd; Emu: 40.6 ± 7.3% sd). The reverse is also true, as the majority (67.4 ± 10.7% sd) of the reads from the DADA2 datasets were assigned to the genera found by CONCOMPRA.

The Illumina-DADA2 and CONCOMPRA analyses captured a similar trend in community composition (Fig. 6), although differences between communities were more pronounced for Illumina-DADA2 datasets. The accumulated compositional difference between Illumina data and ONT-based approaches was largest for Kraken2 (Kraken2: 116.7 ± 7.6% sd; Emu: 83.0 ± 6.9% sd; CONCOMPRA 81.7 ± 17.6% sd).

The NS11–12 marine group (Sphingobacteriales) was detected in all three samples, with all methods (Fig. 7), although their relative abundance differed between the methods. Kraken2 and

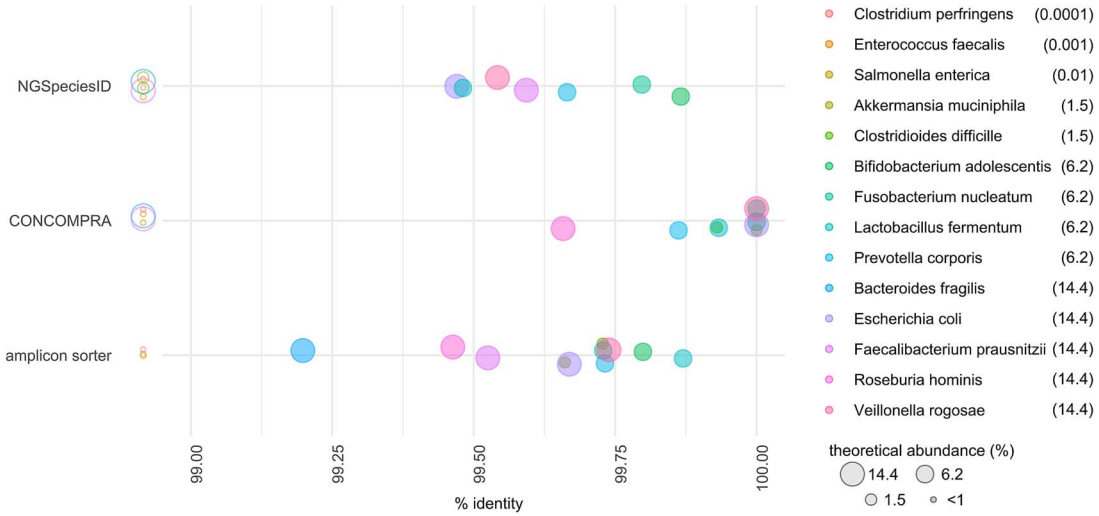

Figure 5. Sensitivity of the different consensus sequence building approaches. The highest similarity to any of the 16S rRNA gene copy sequences to each species (colors) is shown for the different consensus building tools (rows) for the gut microbiome standard. The theoretical abundance of each species (in %) in the community is shown in brackets, next to the species' name and as the size of the circles. Only non-chimeric sequences are considered for CONCOMPRA. Species for which no consensus sequences with an identity above 99% was present are shown as hollow circles to the left of the axis.

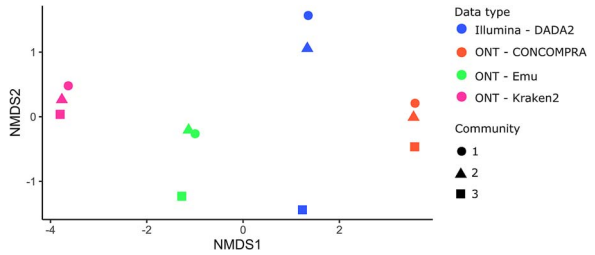

Figure 6. The differences in community composition between natural samples are most pronounced based on Illumina sequencing but the overall trends are conserved across sequencing approaches and analyses. Beta diversity is visualized by non-metric multidimensional scaling (NMDS) using Bray–Curtis dissimilarity. Symbols are used to indicate the different natural bacterial communities and colors are used to represent the different types of analyses. Stress = 0.04.

Emu datasets only had a single taxonomic ID associated with the NS11–12 marine group, whereas CONCOMPRA and DADA2 drafted four and six different SVs assigned to this group, respectively. These different sequences could be assigned to two different clades (Fig. 7). CONCOMPRA identified predominantly members of the first NS11–12 clade in the first two samples and of the second clade in the third sample. This is in line with the results of DADA2.

Discussion

We have presented a novel tool that, starting from ONT amplicon sequencing data, generates a *de novo*, consensus-based sequence database which is then used for abundance profiling of microbial communities. This tool, CONCOMPRA, proved to be more sensitive and generally more precise than the other consensus sequence generation tools tested here. CONCOMPRA performed well in samples with low diversity (20 strains) and was able to capture the diversity of abundant taxa in natural samples.

In natural, more diverse samples, the bacterial diversity reported by CONCOMPRA was lower than the output returned by Emu and substantially lower than the diversity results from Kraken2. Based on the diversity reported by DADA2, the real bacterial diversity within those samples, at least in terms of number of genera present, is likely in between the CONCOMPRA and Emu results. As both CONCOMPRA and Emu resulted in datasets that were compositionally similar to the DADA2 datasets, the differences between CONCOMPRA and Emu reflect the trade-off between sensitivity and specificity.

Reference databases, even of established marker genes such as the 16S rRNA gene that was used in this study, are incomplete. This is the case for the NS11–12 marine group (Bacteroidota), which is important for structuring planktonic bacterial communities in coastal and estuarine waters [43, 44]. As this group is only represented by a single taxon ID in the SILVA_SSU_r138 database, Kraken2 and Emu could not distinguish the different clades present in the natural samples. CONCOMPRA on the other hand, was able to assign the reads to two clades within the NS11–12 marine group, in line with the output from the DADA2 workflow. This example illustrates the benefit of being independent of an existing reference database for assigning reads to OTUs

Our results showed that the bulk of the ONT sequencing data was less similar to references than would be expected based on reported base calling error rates [3]. This is in line with the expected high frequency of chimeric reads in the data. The prevalence of chimeras made it difficult for tools such as NGSpeciesID to generate accurate consensus sequences and inflated the estimated richness by the Emu and Kraken2. This warrants the inclusion of extensive chimera detection and removal steps in long-read amplicon sequencing data workflows. Yacrd was developed for chimera removal during genome assembly rather than handling amplicon sequencing data [21], and it failed to remove most chimeras, while the vsearch uchime_denovo algorithm, as implemented in CONCOMPRA, removed a substantial proportion of likely chimeric reads. By removing chimeric consensus sequences only after mapping the reads in our workflow, we aimed to increase mapping accuracy. This approach is only feasible in a workflow that relies on drafting *de novo* consensus sequences. The use of a *de novo* drafted reference database, as is also the case in most Illumina processing workflows [27, 39, 45] can therefore be considered an advantage over mapping to an existing reference database, as is now implemented in most Kraken2 or Emu workflows. Modifications to the sequencing protocol such as

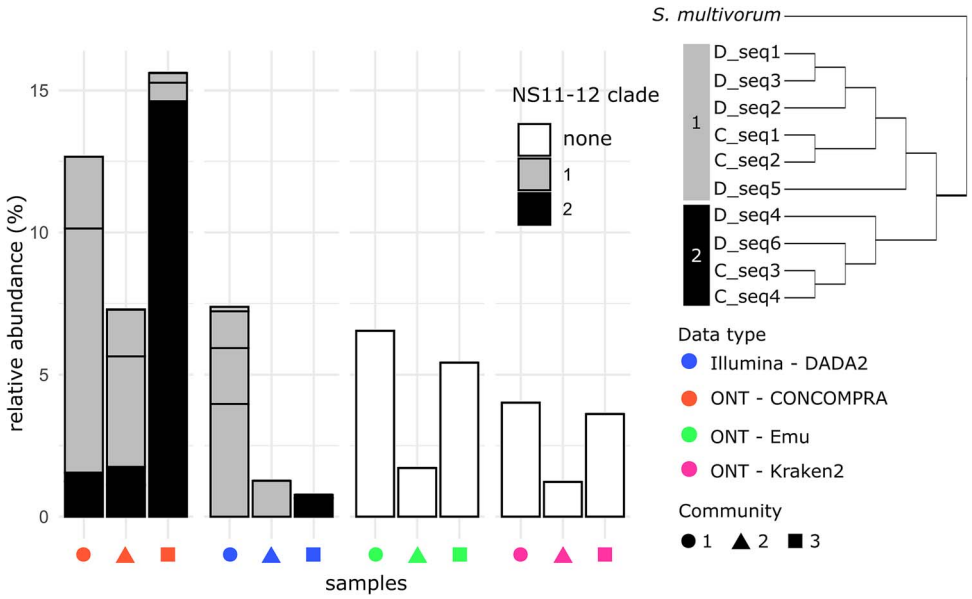

Figure 7. The relative abundance of the NS11–12 marine group in the natural 16S samples. The different clades within the NS11–12 marine group are indicated in the bar plots and cladogram by grey and black. There is no fill (white) if no distinction can be made. The community and data type are indicated in the bar plot by the color and shape of the symbol. The sequences from the DADA2 and CONCOMPRA dataset are indicated in the phylogenetic tree with a 'D' and a 'C' in the tip labels, respectively.

the optimization of PCR conditions [46], are likely to reduce the number of chimeric reads and are important to consider as well.

Many workflows have been developed to process ONT-sequenced (long) amplicon data, most of which rely on mapping reads to existing databases [8, 10, 47, 48]. Here, we argued that there are important benefits to breaking free from reference databases and using a de novo approach instead. We introduced and validated a novel tool, CONCOMPRA, that drafts and maps to consensus sequences, thus allowing the distinction of closely related strains without any external reference database. Although we demonstrated CONCOMPRA on various 16S rRNA gene datasets, its independence of a database makes it particularly suitable for less commonly used amplicons.

Continued improvements in base calling models (e.g., the Dorado v0.7.3 sup model) and error correction tools [49] have improved the accuracy of nanopore sequencing data to the point where nearly flawless bacterial genome assemblies can be obtained [50, 51]. We can expect that similar improvements will soon enable us to detect single base pair differences between closely related strains in ONT-sequenced amplicon data without relying on consensus sequences. This will be an essential step in taking advantage of the long amplicon capability specific to the ONT platforms.

Key Points

- We introduce a reference-free tool for processing (long) amplicon Oxford Nanopore sequencing data.
- We show that it is capable of achieving a higher accuracy than existing tools.
- Chimera detection and removal is an essential step in the processing of long amplicon Oxford Nanopore sequencing data.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the Research Foundation Flanders (FWO) [1252821 N to W.S. and 3F020119 to L.M.L.] and Biodiversa+ RESTORESEAS (G0G3921N), with research infrastructure provided by EMBRC Belgium - FWO project [GOH3817N, I001621N].

Data availability

The ONT sequencing data generated within this study is available in the NCBI BioProject database at https://www.ncbi.nlm.nih.gov/bioproject/, and can be accessed under the accession number PRJNA1129458. The consensus sequences generated by CONCOMPRA, NGSpeciesID, and amplicon_sorter for the synthetic communities are available on Figshare (https://doi.org/10.6084/m9.figshare.27083161.v1). CONCOMPRA is available on GitHub (https://github.com/willem-stock/CONCOMPRA).

Ethics statement

This research has been conducted in a fair and ethical manner, according to international and local guidelines. No human or animal derived data or test subjects were involved.

References

1. D'Andreano S, Cuscó A, Francino O. Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples. *Biol Methods Protoc* 2021;**6**:6. https://doi.org.10.1093/biomethods/bpaa026.

2. van der Loos LM, *et al.* Characterizing algal microbiomes using long-read nanopore sequencing. *Algal Res* 2021;**59**:102456. https://doi.org.10.1016/j.algal.2021.102456.
3. Luo J, Meng Z, Xu X, *et al.* Systematic benchmarking of nanopore Q20+ kit in SARS-CoV-2 whole genome sequencing. *Front Microbiol* 2022;**13**:973367. https://doi.org.10.3389/fmicb.2022.998647.
4. Winand R, Bogaerts B, Hoffman S, *et al.* Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (Illumina) and third (Oxford Nanopore technologies) generation sequencing technologies. *Int J Mol Sci* 2020;**21**:298. https://doi.org/10.3390/biomimetics9030142.
5. Karst SM, Ziels RM, Kirkegaard RH, *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods* 2021;**18**:165–9. https://doi.org.10.1016/j.bbr.2024.115233.
6. Baloğlu B, Chen Z, Elbrecht V, *et al.* A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods Ecol Evol* 2021;**12**:794–804. https://doi.org.10.1111/2041-210X.13561.
7. Calus ST, Ijaz UZ, Pinto AJ. NanoAmpli-Seq: A workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience* 2018;**7**:1–16. https://doi.org.10.1093/gigascience/giy140.
8. Curry KD, Wang Q, Nute MG, *et al.* Emu: Species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nat Methods* 2022;**19**:845–53. https://doi.org.10.3310/JYTR6938.
9. Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: A species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* 2021;**37**:1600–1. https://doi.org.10.1093/bioinformatics/btaa900.
10. Lu J, Salzberg SL. Ultrafast and accurate 16S rRNA microbial community analysis using kraken 2. *Microbiome* 2020;**8**:1–11. https://doi.org.10.1186/s40168-020-00900-2.
11. Yilmaz P, Parfrey LW, Yarza P, *et al.* The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 2014;**42**:D643–8. https://doi.org.10.1093/nar/gkt1209.
12. Abarenkov K, Nilsson RH, Larsson KH, *et al.* The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: Sequences, taxa and classifications reconsidered. *Nucleic Acids Res* 2024;**52**:D791–7. https://doi.org/10.1093/nar/gkad1039.
13. Guillou L, Bachar D, Audic S, *et al.* The protist ribosomal reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 2013;**41**:D597–604. https://doi.org.10.1093/nar/gks1160.
14. Henson MW, Pitre DM, Weckhorst JL, *et al.* Artificial seawater media facilitate cultivating members of the microbial majority from the Gulf of Mexico. *mSphere* 2016;**1**:10–1128. https://doi.org.10.1002/hbm.70043.
15. Katayama T, Nobu MK, Imachi H, *et al.* A marine group a isolate relies on other growing bacteria for cell wall formation. *Nat Microbiol* 2024;**9**:1954–63. https://doi.org.10.1038/s41564-024-01717-7.
16. C-y Wang G, Wang Y. Frequency of formation of chimeric molecules as a consequence of PCR Coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microbiol* 1997;**63**:4645–50. https://doi.org.10.1128/aem.63.12.4645-4650.1997.
17. Ho M, Moon D, Pires-Alves M, *et al.* Recovery of microbial community profile information hidden in chimeric sequence reads. *Comput Struct Biotechnol J* 2021;**19**:5126–39. https://doi.org.10.1016/j.csbj.2021.08.050.
18. Qin Y, Wu L, Zhang Q, *et al.* Effects of error, chimera, bias, and GC content on the accuracy of amplicon sequencing. *mSystems* 2023;**8**:8. https://doi.org.10.1128/msystems.01025-23.
19. Eccles D, White R, Pellefigues C, *et al.* Investigation of chimeric reads using the MinION. *F1000Research* 2017;**6**:631. https://doi.org.10.1080/13811118.2024.2405737.
20. Laver TW, Caswell RC, Moore KA, *et al.* Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci Rep* 2016;**6**:1–6. https://doi.org.10.1002/ksa.12499.
21. Marijon P, Chikhi R, Varré JS. Yacrd and fpa: Upstream tools for long-read genome assembly. *Bioinformatics* 2020;**36**:3894–6. https://doi.org.10.1093/bioinformatics/btaa262.
22. Hamada M, Ono Y, Asai K, *et al.* Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics* 2017;**33**:926–8. https://doi.org.10.1093/bioinformatics/btw742.
23. Kiełbasa SM, Wan R, Sato K, *et al.* Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011;**21**:487–93. https://doi.org.10.1101/gr.113985.110.
24. McInnes L, Healy J, Saul N., Großberger L.. *Melville J*. Uniform Manifold Approximation and Projection for Dimension Reduction: UMAP, arXiv preprint 2018, **3**, 10.21105/joss.00861.
25. Malzer C, Baum M. A hybrid approach to hierarchical density-based cluster selection. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems 2020*. Karlsruhe, Germany: IEEE Xplore; 2020-September:223–8, https://doi.org/10.3390/ma14143884, **14**.
26. Frith MC, Mitsuhashi S, Katoh K. Lamassemble: Multiple alignment and consensus sequence of long reads. *Methods Mol Biol* 2021;**2231**:135–45. https://doi.org.10.1007/978-1-0716-1036-7_9.
27. Rognes T, Flouri T, Nichols B, *et al.* VSEARCH: A versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584. https://doi.org.10.7717/peerj.2584.
28. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org.10.1093/bioinformatics/bty191.
29. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford nanopore sequencing. *Genome Biol* 2019;**20**:1–10. https://doi.org.10.1186/s13059-019-1727-y.
30. De Coster W, Rademakers R. NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics* 2023;**39**:btad311. https://doi.org.10.1093/bioinformatics/btad311.
31. D'Hondt AS, Stock W, Blommaert L, *et al.* Nematodes stimulate biomass accumulation in a multispecies diatom biofilm. *Mar Environ Res* 2018;**140**:78–89. https://doi.org.10.1016/j.marenvres.2018.06.005.
32. Szoboszlay M, Schramm L, Pinzauti D, *et al.* Nanopore is preferable over Illumina for 16S amplicon sequencing of the gut microbiota when species-level taxonomic classification, accurate estimation of richness, or focus on rare taxa is required. *Microorganisms* 2023;**11**:804. https://doi.org.10.3390/microorganisms11030804.
33. Murali A, Bhargava A, Wright ES. IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 2018;**6**:1–14. https://doi.org.10.1186/s40168-018-0521-5.
34. Quast C, Pruesse E, Yilmaz P, *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6. https://doi.org/10.1093/nar/gks1219.
35. Sahlin K, Lim MCW, Prost S. NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. *Ecol Evol* 2021;**11**:1392–8. https://doi.org.10.1002/ece3.7146.

36. Vierstraete AR, Braeckman BP. Amplicon_sorter: A tool for reference-free amplicon sorting based on sequence similarity and for building consensus sequences. *Ecol Evol* 2022;**12**:e8603. https://doi.org.10.1002/ece3.8603.
37. Hossin M, Sulaiman M. A review on evaluation metrics for data classification evaluations. *Zeitschrift für die gesamte Anatomie 1 Abt* 2015;**5**:01–11. https://doi.org.10.5121/ijdkp.2015.5201.
38. Kalikar S, Jain C, Vasimuddin M, *et al.* Accelerating minimap2 for long-read sequencing applications on modern CPUs. *Computational science* 2022;**2**:78–83. https://doi.org/10.1038/s43588-022-00201-8.
39. Callahan BJ, McMurdie PJ, Rosen MJ, *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3. https://doi.org.10.1101/2024.11.03.619087.
40. McMurdie PJ, Holmes S. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One* 2013;**8**:e61217. https://doi.org.10.1371/journal.pone.0061217.
41. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PloS One* 2010;**5**:e9490. https://doi.org.10.1371/journal.pone.0009490.
42. Nawrocki E, Structural RNA *Homology Search and Alignment Using Covariance Models*. Ann Arbor, Michigan, USA: UMI Dissertation Publishing, 2009.
43. Wang H, Chen F, Zhang C, *et al.* Estuarine gradients dictate spatiotemporal variations of microbiome networks in the Chesapeake Bay. *Environmental Microbiomes* 2021;**16**:1–18. https://doi.org.10.1186/s40793-021-00392-z.
44. Korlević M, Markovski M, Herndl GJ, *et al.* Temporal variation in the prokaryotic community of a nearshore marine environment. *Sci Rep* 2022;**12**:1–9. https://doi.org.10.1038/s41598-022-20954-6.
45. Estaki M, Jiang L, Bokulich NA, *et al.* QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Curr Protoc Bioinformatics* 2020;**70**:e100. https://doi.org.10.1002/cpbi.100.
46. Fujiyoshi S, Muto-Fujita A, Maruyama F. Evaluation of PCR conditions for characterizing bacterial communities with full-length 16S rRNA genes using a portable nanopore sequencer. *Sci Rep* 2020;**10**:1–10. https://doi.org.10.1093/dote/doae097.
47. Ammer-Herrmenau C, Pfisterer N, van den Berg T, *et al.* Comprehensive wet-bench and bioinformatics workflow for complex microbiota using Oxford Nanopore technologies. *mSystems* 2021;**6**:10.1128/msystems.00750-21. https://doi.org.10.1136/gutjnl-2024-332236.
48. Bertolo A, Valido E, Stoyanov J. Optimized bacterial community characterization through full-length 16S rRNA gene sequencing utilizing MinION nanopore technology. *BMC Microbiol* 2024;**24**:1–12. https://doi.org.10.1186/s12866-024-03208-5.
49. Stanojević D, Lin D, Sessions PF de, *et al.* Telomere-to-telomere phased genome assembly using error-corrected simplex nanopore reads. *bioRxiv*. **2024**. 594796, https://doi.org.10.1183/13993003.01675-2024.
50. Sanderson ND, Hopkins KMV, Colpus M, *et al.* Evaluation of the accuracy of bacterial genome reconstruction with Oxford nanopore R10.4.1 long-read-only sequencing. *Microb Genom* 2024;**10**:001246. https://doi.org.10.1099/mgen.0.001246.
51. Biggel M, Cernela N, Horlbog JA, *et al.* Oxford Nanopore's 2024 sequencing technology for listeria monocytogenes outbreak detection and source attribution: Progress and clone-specific challenges. *bioRxiv* 2024.07.12.603236. https://doi.org.10.1128/jcm.01083-24.

Briefings in Bioinformatics, 2025, **26(1)**, bbae642
https://doi.org/10.1093/bib/bbae642
Problem Solving Protocol