GUT MICROBES
2025, VOL. 17, NO. 1, 2516703
https://doi.org/10.1080/19490976.2025.2516703

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

De novo clustering of long-read amplicons improves phylogenetic insight into microbiome data

Yan Hui [a*], Dennis Sandris Nielsen [b], and Lukasz Krych [b*]

[a]Department of Preventive Medicine, School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China; [b]Department of Food Science, Faculty of Science, University of Copenhagen, Frederiksberg C, Denmark

ABSTRACT

Long-read amplicon profiling through read classification limits phylogenetic analysis of amplicons while community analysis of multicopy genes, relying on unique molecular identifier (UMI) corrections, often demands deep sequencing. To address this, we present a long amplicon consensus analysis (LACA) workflow employing multiple *de novo* clustering approaches based on sequence dissimilarity. LACA controls the average error rate of corrected sequences below 1% for the Oxford Nanopore Technologies (ONT) R9.4.1 and ONT R10.3 data, 0.2% for ONT R10.4.1, and 0.1% for high-accuracy ONT Duplex and Pacific Biosciences (PacBio) circular consensus sequencing (CCS) data in both simulated 16S rRNA and real 16-23S rRNA amplicon datasets. In high-accuracy PacBio CCS data, the clustering-based correction matched UMI correction, while outperforming 4× UMI correction in noisy ONT R10.3 and R9.4.1 data. Notably, LACA preserved phylogenetic fidelity in long operational taxonomic units and enhanced microbiome-wide phenotype characterization for synthetic mock communities and human vaginal samples.

ARTICLE HISTORY
Received 17 March 2025
Revised 13 May 2025
Accepted 2 June 2025

KEYWORDS
Amplicon; denoise; long read sequencing; workflow; UniFrac

Introduction

High-throughput amplicon sequencing remains a cost-effective strategy to analyze genetic polymorphisms in DNA specimens, also in situations where the quantity of the sample DNA is low,[1] rendering it a widely employed technique in microbial[2,3] profiling. Until now, short-read technologies like Illumina dominated amplicon sequencing. However, the maximum insert size of ~ 500 bp[4] necessitates either fragmenting amplicons or limiting the length of an amplified region. The emergence of long-read sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), has revolutionized amplicon sequencing, enabling the complete sequencing of amplicons up to 10 kb.[5] However, these technologies exhibit variable sequencing error rates, ranging from 1% to 15%, depending on the used platforms.[6–8] To address this, many long-read amplicon analysis approaches often resort to closed-reference classification. For instance, the official ONT EPI2ME analysis toolbox utilizes this strategy, employing read-by-read classification against a known reference database via Centrifuge.[9] More recently, Emu[10] introduced an expectation-maximization algorithm to achieve species-level microbial profiling through full-length 16S rRNA gene amplicon sequencing. Nevertheless, these methods come with limitations, potentially overlooking species not included in trusted reference databases and failing to retain the evolutionary links of taxonomic traits in downstream analysis.

The variable sequencing error rate challenges *de novo* clustering of long noisy reads into operational taxonomic units (OTUs). IsONclust[11] is an innovative method that leverages base quality values in *de novo* clustering of such reads. Based on this, NGSpeciesID achieved remarkable accuracy, surpassing 99.3% when compared to Sanger

CONTACT Yan Hui huiyan@hznu.edu.cn Department of Preventive Medicine, School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China; Lukasz Krych krych@food.ku.dk Department of Food Science, Faculty of Science, University of Copenhagen, Frederiksberg C DK-1958, Denmark

*These authors jointly supervised this work.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19490976.2025.2516703.

sequencing, in retrieving ribosomal operon sequences from distant fungal strains.[12] Using an assembly-free strategy of read binning with 5-mer frequency similarity,[13] NanoCLUST pioneered a reference-free workflow for OTU clustering of noisy ONT amplicons of full-length 16S rRNA gene.[14] NanoCLUST attained species-level resolution and effectively profiled an eight-strain mock community.[14] In high-accuracy PacBio circular consensus sequencing (CCS) data, IsoCon has demonstrated nucleotide-level precision in deciphering highly similar multigene families.[15] Notably, with noisy ONT R9.4.1 transcriptome data, isONcorrect obtained a comparable median accuracy of 98.9–99.6% to PacBio CCS without reference reliance.[16] This suggests the potential to expand the use of IsoCon on noisy ONT data. Besides, the implementation of unique molecular identifiers (UMIs) in library preparation has been well-documented as an effective means to reduce the error rate in ONT and PacBio CCS sequencing to below 0.01%, but unfortunately the UMI-approach also requires rather deep sequencing depth.[5]

Until now, integrated frameworks such as QIIME 2[17] offer limited support for long-read amplicon analysis. Further, a systematic assessment of the precision and recall of these reference-free methods is lacking. To bridge this gap, this study presents a scalable and reproducible analysis workflow for *de novo* long amplicon consensus analysis (LACA). LACA incorporates multiple sequence dissimilarity clustering approaches to identify true amplification templates amidst the noise. Utilizing *in silico* benchmarking, public UMI-tagged sequencing data, and full-length 16S rRNA gene amplicon ONT sequencing data of synthetic mock and human vaginal microbiomes, our study offers a comprehensive overview of the reference-free solutions in long-read amplicon analysis.

Materials and methods

A reproducible and scalable workflow for long-read amplicon consensus analysis

LACA is designed for *de novo* OTU picking from noisy long-read amplicon sequencing data, regardless of whether UMIs are included in library preparation (Figure 1(a)). It is modularized with Snakemake[18] sub-workflows and offers user-friendly controls for demultiplexing, quality control, OTU picking, and downstream community analysis, i.e., quantification, taxonomy assignments, and phylogenetic reconstruction (Figure 1(b,c)). Taking advantage of Snakemake,[18] LACA assures reproducible analysis with a pre-defined configuration file and is scalable to support large-scale sample processing in cluster systems. Unlike short-read technologies, long-read sequencing preserves the integrity of amplicons. This enables LACA to identify the amplified regions by searching for the linked primer patterns and re-orient the sequence if a reverse strand is sequenced (Figure 1(b)). The remaining reads are pooled or processed independently for clustering and consensus calling after chimera and quality filtering. To avoid the heavy computing cost of pairwise global alignments, LACA borrows multiple alignment-free approaches, e.g., HDBSCAN clustering with k-mer frequency[13,14] (referred to as UMAPclust across the study) or error-aware clustering with minizers[11] (referred to as isONclust across the study), and consolidates the clusters with Meshclust.[19] Relying on the mean shift algorithm[20] and alignment-free identity scores,[21] LACA identifies the centroids and members within the cluster, thus generating the consensus sequences ("kmerCon"). As shown in Figure 3(a), LACA supports refining these clusters according to alignment overlap[13] ("miniCon"), identifying highly similar haplotypes in the cluster with isONcorrect[16] and IsoCon[15] ("isoCon") and also independent molecule-level profiling with UMIs[5] ("umiCon"). According to user requirements, the centroid is picked and polished with Racon[22] or Medaka with the supporting reads either in a kmer-based, alignment-based, or UMI cluster. The corrected sequences are then dereplicated by sequence identity to extract representative OTUs, enabling phylogenetic inference and taxonomic assignment in community analysis (Figure 1(c)). The OTU count matrix can be created by the uniquely assigned FASTQ sequence identifiers in the clustering and demultiplexing records or re-mapping the

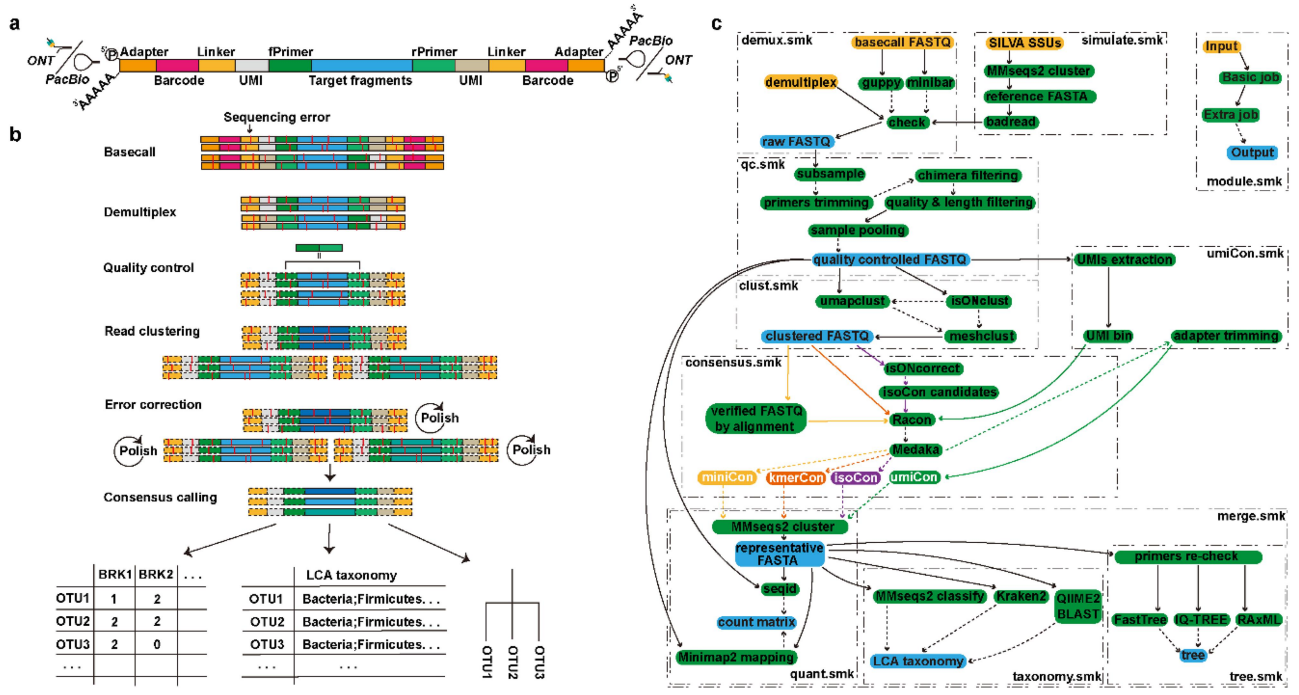

Figure 1. (a) A schematic illustration of the amplicon structure suitable for long amplicon consensus analysis (LACA). (b, c) Graphical representation of the essential steps in the LACA workflow (b) and a bioinformatic flowchart (c). Briefly, LACA enables demultiplexing sequencing reads by Oxford Nanopore technologies with custom barcodes or directly picks external demultiplexed data for downstream analysis. The demultiplexed reads passing through primer trimming and quality control are either pooled together or processed independently to retrieve consensus sequences for OTU picking. The OTUs are used to obtain phylogenetic and taxonomic relationships for community analysis. The OTU count table is generated by incorporating OTU clustering and read demultiplexing information, utilizing either the unique sequence ID (in a clustering-based approach) or read remapping (in a mapping-based approach). The OTUs are re-clustered to select final representatives when merging LACA runs, and the OTU table, taxonomy, and tree file are updated accordingly.

sequencing reads against the generated OTU sequences. Multiple sequencing runs of LACA outputs can be merged for meta-analysis, in which run-specific OTUs are concatenated, re-dereplicated by sequence identity, and the OTU count matrices are merged accordingly. The computational resources required to run LACA are scalable with the input data size. A single MinION run with 96 samples (each with approximately 10,000 reads after quality filtering) performed on an AMD Ryzen 7 processor with 64 GB of RAM would take less than 12 hours. On average, a MinION flow cell generates 3 to 4 million high-quality reads (after demultiplexing and quality control steps), which would yield approximately 30,000 to 40,000 reads per sample if equimolar pooling is achieved (internal data from ~ 20 minION individual flow cells). In practice, this in-house setup ensures that in a 96-sample pool, each sample is represented by no less than 10,000 reads.

Data generation

Long amplicon simulation on SILVA small subunit (SSU) rRNA sequences

The whole simulation process was accomplished with the simulation module in the LACA workflow. In short, the reference sequences were selected according to the sequence divergence requirements (of 5–10%, 3–5%, 2–3%, 1–2%, and 0–1%) through two rounds of MMseqs2[23] clustering of SILVA[24] 138.1 SSU rRNA sequences and were used as reference sequences for read simulation by Badread.[25] For each selected reference set, ten sequences were picked and had the same length of 1522 bp after being clustered by minimal and maximal sequence identity. The relative abundance of each reference sequence was set to equal (even10) or skewed with two abundant sequences 10 times as much as the rest (skew10). The sequence identity of ONT R9.4.1 and ONT R10.4.1 reads followed a beta distribution with

mean, max, and standard deviation of 87.5%, 97.5%, and 5% using the error and quality model of "nanopore2020", and 95%, 99%, and 2.5% for the "nanopore2023" model, respectively. The simulation identity of high-accuracy ONT Duplex and PacBio HiFi reads followed a normal quality score distribution with a mean of 20 and standard deviation of 4 using the error models "nanopore2023" and "pacbio2016", respectively. This led to four sets of long-read sequencing data, namely "ONT2020", "ONT2023", "ONT Duplex", and "PacBio CCS" used throughout the study. By default, the simulated reads included 1% of chimeras and were appended with the ligation adapters of 5'-AATGTACTTCGTTCAGTTACGTATTGCT-3' and 5'-GCAATACGTAACTGAACGAAGT-3' on both ends of either strand. The *in silico* read coverage was set to 40×, 60×, 80×, 100×, 200×, and 400×, respectively. Thus, the simulation process resulted in 120 independent read files (from various sequencing platforms and depth), for the following performance test.

Near full-length 16S rRNA amplicon sequencing on diluted mock serials

DNA source

The ZymoBIOMICS microbial community DNA standard (D6306, Zymo Research) was used as mock DNA for library preparation. This commercial mock contains DNA from ten different microorganisms, including eight bacteria and two yeasts: *Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella enterica*, *Limosilactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, and *Cryptococcus neoformans*. According to the manufacturer's provided reference genomes (https://zymo-files.s3.amazonaws.com/BioPool/ZymoBIOMICS.STD.refseq.v3.zip), the mock product contains forty-nine 16S rRNA operons from the prokaryotes. The original mock DNA was diluted in series with sterile water, resulting in samples with the mock DNA concentrations of 5, 0.5, 0.25, 0.13, 0.05, and 0.03 ng/μL, respectively. The DNA concentration was measured with Qubit 1× dsDNA HS assay kit (Invitrogen, Thermo Fisher Scientific).

Target gene amplification and metabarcoding

The library preparation followed a published protocol[26] for ONT sequencing of near full-length 16S rRNA amplicons with minor modification. In short, near full-length 16S rRNA gene was amplified with a protocol involving multiple forward and reverse primers: 27Fa/b (5'- GTCTCGTGGG CTCGGNNNNN NNNNNNNNNN AGAGTTT GAT YMTGGCTYAG –3', 5'- GTCTCGTGGG CTCGGNNNNN NNNNNNNNNN AGGGTTC GAT TCTGGCTCAG –3'), 338Fa/b (5'- GTCTCGTGGG CTCGGNNNNN NNNNNNNNNN ACWCCTACGG GWGGCAGC AG –3', 5'- GTCTCGTGGG CTCGGNNNNN NNNNNNNNNN GACTCCTAC GGGAGGCWG CAG –3'), 1391 R (5'- GTCTCGTGGG CTCGGN NNNN NNNNNNNNNN GACGGGCGGT GTGTRCA –3') and 1540 R (5'- GTCTCGTGGG CTCGGNNNNN NNNNNNNNNN TACGGYT ACC TTGTTACGACT-3'). First PCR conditions were as follows: 95°C for 5 min, 2 cycles of 95°C for 20 s, 48°C for 30 s, 65°C for 10 s, 72°C for 45 s, and a final extension at 72°C for 4 min. A second PCR step was carried out to barcode the first PCR products with the following conditions: 95°C for 2 min followed by 33 cycles of 95°C for 20 s, 55°C for 20 s, 72°C for 40 s, and a final extension at 72°C for 4 min. PCR products were cleaned up using AMPure XP beads (Beckman Coulter Genomic, CA, USA) after each PCR step. After validation with agarose gel electrophoresis, the final PCR products were pooled for ONT sequencing.

ONT sequencing with barcoded amplicons

The ONT sequencing library was constructed according to the ligation sequencing kit SQK-LSK109 protocol. The library was loaded on ONT R9.4.1 flow cell and sequenced with GridIONX5 platform (Oxford Nanopore Technologies, Oxford, UK). Sequencing data was collected by MinKnow (v21.05) for Guppy (v5.0.11) basecalling in high-accuracy mode.

Long amplicon analysis

Clustering quality comparison of alignment-free approaches on simulated SILVA amplicons

To assess the clustering performance of alignment-free approaches, LACA was adapted to treat the 20 simulated SILVA amplicon read

files (under a fixed sequencing depth of 200×) as independent samples and consensus calling was performed without sample pooling to avoid cross-contamination from the other simulation scenarios e.g., the different sequencing platforms. The ligation adapters were treated as primers by LACA, and the *in silico* reads were checked with Cutadapt[27] for the linked primer pattern of 5'-AATGTACTTCGTTCAGTTACGTATTGCT ... GCAATACGTAACTGAACGAAGT –3') with a maximal error rate of 0.2 and minimal overlaps of 6 bases. The primer sequences were retained in the checked reads and extra bases were trimmed. Chimeric reads were filtered by yacrd[28] with minimal read coverage of 0.4 and the minimal coverage was set to 4 and 3 for ONT and PacBio CCS amplicons, respectively. The read length range was set to 800–2000 bp and low-quality reads were discarded if the average quality score was below 7. The quality-controlled reads were sent for alignment-free clustering with isONclust,[11] UMAPclust and Meshclust.[19]

We benchmarked the effect of various combinations of these alignment-free approaches and the Meshclust[19] thresholds on the clustering performance of the SILVA amplicons from the four types of long-read sequencing data. To determine appropriate Meshclust[19] thresholds for long read data, we used LACA in the "clust" mode only with Meshclust[19] using a set of the identity scores of 0.7, 0.75, 0.8, 0.85 and "auto" for the ONT2020 reads, 0.8, 0.85, 0.9, 0.95 and "auto" for the ONT2023 reads and 0.85, 0.9, 0.95 and "auto" for the ONT Duplex and PacBio CCS reads. In the clustering benchmark of various approaches and combinations, the estimated identity score by Meshclust[19] was applied for all four data types. And LACA was performed in "clust" mode, employing isONclust,[11] UMAPclust, or a sequential combination of both, culminating in the final step of Meshclust[19] clustering. For isONclust,[11] ONT reads were clustered with the recommend parameters of "-k 13 -w 20" and PacBio CCS reads were clustered with the flag "-k 15 -w 50". The UMAPclust clustering stuck to the settings of NanoCLUST[14] with HDBSCAN[29] clustering on the two-dimensional UMAP[30] graph of the canonical 5-mer frequency of these amplicon reads. The UMAP[30] adopted similar parameters with "n_neighbors = 15, min_dist = 0.1, metric=cosine" and the HDSCAN[29] parameters were set as "min_bin_size = 10, min_samples = 10, epsilon = 0.5".

Consensus-calling evaluation on the simulated SILVA amplicons with various sequencing depths

To assess performance of kmerCon, miniCon, isoCon, LACA was adapted to treat the 120 simulated SILVA amplicon read files as independent samples for consensus calling to avoid cross-contamination from the other simulation scenarios. The quality control process stuck to the same procedures as that in the above section "Clustering quality comparison of alignment-free approaches on simulated SILVA amplicons". The passed reads were first clustered by the alignment-free approaches with UMAPclust and Meshclust,[19] and the three types of consensus calling were conducted within the clustered reads. Given increasing $O(n^2)$ of pairwise alignments with cluster size, miniCon was conducted in batches with a maximum size of 5,000 reads per batch. The minimal fraction of maximum score for a miniCon cluster was set to 0.75 and 0.85 for ONT2020 and the rest read types. In the isoCon mode, read correction by isONcorrect[16] was not enabled for high-accuracy ONT Duplex and PacBio CCS data to maximally preserve the variant signals. Each consensus sequence had at least 20 supporting reads and had two rounds of Racon[22] polishing. One extra round of Medaka polishing was introduced to produce the final ONT consensus sequences. The "r941_min_hac_g507" consensus model was used for the ONT2020 amplicons while the "r1041_e82_400bps_hac_v4.2.0" consensus model was adopted for the ONT2023 and ONT Duplex amplicons. For ONT Duplex and PacBio CCS data, we also conducted consensus calling within the clusters generated by isONclust[11] and UMAPclust. This was done under the loose supervision by Meshclust[19] using an identity score of 0.5, which maximally preserved the clusters in the initial two rounds of clustering.

Consensus calling evaluation on UMI-tagged long amplicons

We utilized the PacBio CCS, ONT R9.4.1, and R10.3 UMI-tagged amplicon datasets[5] of the ZymoBIOMICS mock and subsampled each to 100,000 and 1,000,000 reads. The consensus

sequences were extracted with LACA according to the sequencing platform and clustering approaches. For kmerCon, miniCon and isoCon, we followed similar quality control and consensus calling procedures in the section "Clustering quality comparison of alignment-free approaches on simulated SILVA amplicons". The forward (5'-AGRGTTYGATYMTGGCTCAG-3') and reverse primer (5'-CGACATCGAGGTGCCAAAC-3') were checked with outside bases trimmed, and chimeric reads were excluded. To keep consistent with original publication,[5] the allowed read length range was set from 3500 to 6000 bp and the minimal average quality score was 7. Then UMAPclust and Meshclust[19] were used for alignment-free clustering of ONT reads, the "min_bin_size = 50, min_samples = 50" was used for HDBSCAN in UMAPclust and the estimated identity score was used for Meshclust.[19] The PacBio CCS reads were clustered by isONclust,[11] UMAPclust and Meshclust[19] in a sequential order with a Meshclust[19] identity score of 0.5. Given increasing $O(n^2)$ of pairwise alignments with cluster size, miniCon was conducted in batches with a maximum size of 5,000 reads per batch. And the minimal fraction of the maximum score for a miniCon cluster was set to 0.85. For isoCon, read correction by isONcorrect[16] was enabled for ONT data, and consensus calling was performed in batches (with a maximum size of 20,000 reads per batch). Each clustering-based consensus sequence had at least 3 supporting reads. The polishing phase by Racon[22] or Medaka stuck to the same settings as that in UMI-based corrections.[5] For umiCon, the consensus sequences were extracted using the 36-bp UMIs (5'-NNNYRNNNYR NNNYRNNNNN NYRNNNYRNN NYRNNN-3') following the published procedures[5] after read length check. The existence of primers was checked and retained on the corrected sequences and extra bases were trimmed.

Full-length 16S rRNA gene amplicon analysis of a diluted mock series

Demultiplexing was accomplished using Guppy (implemented in LACA) with custom barcodes. To evaluate the LACA performance on the mock samples, the demultiplexed reads were subsampled to 10,000 (subsample 10,000) and 100,000 (subsample 100,000) per sample if possible. Similarly, we followed the same consensus calling procedures in the section "Consensus calling evaluation on UMI-tagged long amplicons" unless otherwise specified. The spurious and chimeric amplicon reads were filtered if the sequence length exceeded the range of 800 to 1600 bp. Given the multiple-primer strategies in amplification, reads were retained if they contained any possible combinations of linked patterns of the forward and reverse primers. The remaining reads were first pooled together for alignment-free clustering of UMAPclust and Meshclust[19] (with a minimal identity score of 0.9), followed by consensus calling with kmerCon, miniCon, and isoCon. Each consensus sequence had at least 10 supporting reads and two rounds of Racon[22] polishing followed by one round of Medaka polishing using the "r941_min_hac_g507" consensus model. The OTUs were the representative sequences picked through MMseqs2[23] clustering with an alignment coverage above 0.9 and a sequence identity above 0.99. The OTU matrix was constructed according to the preserved sequence ID information in the OTU clustering and demultiplexing. The phylogenetic tree was built with FastTree[31] implemented in QIIME 2[17] and the OTU taxonomy was determined by QIIME 2[17] classify-consensus-blast against the SILVA[24] 138.1 SSUs. By default, the local common ancestor (LCA) taxonomy was taken for each OTU sequence based on the agreement among ten BLAST hits.

Full-length 16S rRNA gene amplicon analysis of human vaginal microbiomes

To evaluate LACA performance in real-world applications, we re-utilized the public human vaginal dataset in the Emu[10] paper. The LACA settings were as stated above (section "Full-length 16S rRNA gene amplicon analysis of a diluted mock series") unless otherwise specified. According to the chosen primer sets (27F: AGAGTTTGATCMTGGCTCAG and 1492 R: CGGTTACCTTGTTACGACTT), the amplificon length range was set 1300 to 1700 bp for primer check. We used UMAPclust and Meshclust[19] (with the estimated identity scores) for initial alignment-free clustering, followed by consensus calling with kmerCon, miniCon, and isoCon. And each OTU sets was picked from the respective corrected sequence sets with an alignment

coverage above 0.99 and a sequence identity above 0.99. To maintain consistency with the Emu profile, the respective OTU sets were picked without sample pooling. To compare with the OTU matrix based on clustering information, we mapped quality-controlled reads back to the OTU sequences, generating an alignment-based count matrix. The Emu profiles of each sample were processed with our open-source workflow NART (https://github.com/yanhui09/nart) and merged into one matrix by the detected taxonomic features. Similarly, the primer sequences were retained for Emu alignment and the sequencing data were quality controlled as that in the LACA process. The SILVA 138.1 database was adopted to determine the taxonomic assignments in all profiling approaches.

Data analysis

Clustering evaluation on in silico data

The clustering quality was assessed by examining the number of clusters and two external metrics: purity and normalized mutual information (NMI). The external metrics were calculated using the true class labels of the simulated data as a reference. The clustering purity measures how homogeneous a cluster is with respect to a single class,[32] and is defined in Equation (1) as below:

$$purity(M,D) = \frac{1}{N}\sum\nolimits_{m\in M} max_{d\in D}\left|m \cap^{d}\right| \quad (1)$$

where M is the set of clusters m, D is the set of classes d and N is the number of clustered reads. NMI measures the similarity between the clustering results and the ground truth and trades off the clustering quality against the number of clusters,[32] and is defined in Equation (2) as below:

$$NMI(M,D) = \frac{2 \times I(M;D)}{H(M) + H(D)} \quad (2)$$

where M is the set of the clusters, D is the set of classes, I is the mutual information as defined in Equation (3), and H is entropy as defined in Equation (4).

Given $p(m)$, $p(d)$ and $p(m \cap d)$ are the probabilities of a read being in cluster m of the set M, class d of the set D, in the intersection of m and d, $I(M;D)$ is defined as below:

$$I(M;D) = \sum\nolimits_{m\in M}\sum\nolimits_{d\in D} p\left(m \cap^{d}\right) log_2 \frac{p\left(m \cap^{d}\right)}{p(m)p(d)} \quad (3)$$

Given p_i is the probability of label i in the set M, $H(M)$ is defined as below:

$$H(M) = -\sum\nolimits_{i} p_i log_2(p_i) \quad (4)$$

The clustering purity was calculated by custom R code and NMI by R package aricode[33] as stated in the "Code Availability" section.

Alignment validation with simulation reference sequences

The consensus sequences were split by sequencing depth, simulation reference set, and consensus-calling methods after the adapter sequences were trimmed from both sides with Cutadapt.[27] The split sequences were aligned against the respective reference sequences at the base level with Minimap2[34] under the preset of asm5, asm10, and asm20, and only primary alignments were kept. Given the sequence redundancy of SILVA SSUs, frequent minimizers occurring more than 10,000 times were ignored in Minimap2[34] mapping. The alignment identity was calculated by dividing the number of total bases, including gaps in the mapping, by the number of matching bases. The query (consensus) sequence cover was calculated by dividing the number of matches by the length of the query sequence. The target (reference) sequence cover was calculated by dividing the number of matches by the length of the target sequence.

Error profiling of consensus sequences from UMI-tagged data

The derived consensus sequences were validated using the script "longread qc_pipeline" with the curated 16S-23S rRNA operon ZymoBIOMICS reference and SILVA[24] 132 SSURef Nr99 reference.[5] The R script "validation functions.R" (provided in "Code Availability" section and revised from https://github.com/SorenKarst/longread_umi) was used to compile the generated results for error profiling and quality evaluation. Briefly, the error types (mismatch, deletion, insertion) and the relative positions of the errors were determined with respect to the reference sequence. These errors were then categorized as

being within homopolymer regions (hp+) or not (hp-). Accordingly,[5] the error information combined with other data (such as the number and length of consensus sequences, cluster contamination, ZymoBIOMICS reference-based taxonomy, SILVA taxonomy, and chimera detection) was calculated before and after the PCR artifacts were removed. The consensus sequences were flagged as PCR artifacts if the errors are evenly distributed and if the operon SSU part has a better match with SILVA SSU than the ZymoBIOMICS reference. The remaining PacBio homopolymer artifacts were manually detected and flagged in the corrected sequences. The function "lu_artifact_plot" in the R script was used to identify artifacts, using cluster size intervals of 3 (when possible) from 1 to 60 and larger than 60 for UMI clusters. The intervals for other types of clustering methods were set using quantiles of 0%, 25%, 50%, and 75% and greater than 75% due to the significant decrease in the number of clusters. The cluster contamination was estimated by the percentage of the clustered reads assigned to different ZymoBIOMICS reference-based taxa except for the most common taxonomic class. The chimeras were identified if called by UCHIME2[35] with the curated ZymoBIOMICS reference and if errors tended to occur more frequently on either end of a sequence.

Community analysis of the diluted mock serials and human vaginal microbiomes

The clustering-based count matrices were used for community analysis unless otherwise clarified. We integrated the count matrix, phylogenetic tree file, and LCA taxonomy file for community analysis with R package phyloseq.[36] The result generated by each approach was analyzed independently, and high-quality OTUs were retained for community analysis with at least one BLAST hit showing alignment identity above 97% to the SILVA SSUs. Hellinger transformation was applied on the raw count matrix for dissimilarity metrics computation in the vaginal dataset while the raw count matrix was used for the subsampled synthetic mock dataset. Bray Curtis and weighted UniFrac dissimilarity metrics were calculated in comparison for PERMANOVA test using R package vegan. The BLAST hits for LCA taxonomy were collected to evaluate the alignment identity and cover. The script is provided in the "Code Availability" section.

Intersection visualization and correlation analysis of taxonomic features

The taxonomic features from various profiles were extracted at species and genus level and the intersections between matrix were visualized in UpSet[37] plot. Spearman's rank and Pearson correlation analysis of the relative abundance of shared taxonomic features were conducted between various LACA profiles and Emu output, and the rare features present among less than 30% of samples were excluded. The same procedures were applied for the correlation analysis between the count matrices constructed by the clustering-based and mapping-based approach. The analysis script is provided in the "Code Availability" section.

Results

In silico benchmark with SILVA amplicons

De novo OTU picking relies on sequence dissimilarity information to cluster amplicons from the same template. However, erroneous base calls can lead to misclassification of similar sequences. Given the varied sequencing accuracy among long-read platforms, we posited that clustering performance on long amplicons is influenced by sequencing depth and target sequence divergence. To test this hypothesis, we selected five sets of ten SSU rRNA gene sequences from the SILVA database. These sequence sets had sequence divergences of 5–10%, 3–5%, 2–3%, 1–2%, and 0–1%. They served as reference sequences to simulate long amplicons from the early ONT R9.4.1 (with a mean read identity of 87.5%, referred to as ONT2020 throughout the study), ONT R10.4.1 (with a mean read identity of 95%, referred to as ONT2023 throughout the study), ONT Duplex (with a mean read identity of 99%), and PacBio CCS platforms (with a mean read identity of 99%).[25] We initially assessed the clustering quality of alignment-free approaches on the simulated reads at a sequencing coverage of 200×. We recorded the number of clusters and subsequently calculated purity and NMI scores using the read simulation source for external cluster validation. As expected, clustering performance improved with sequencing accuracy and target sequence divergence, although various approaches exhibited different clustering

sensitivity and specificity. Across all data types and simulation profiles, UMAPclust consistently generated clusters with the highest levels of purity and NMI scores compared to isONclust and Meshclust (Figure 2). UMAPclust retained over 75% of the original sequencing reads except for an approximate 20% decline in identifying ONT2020 reads from highly similar reference sequences with sequence divergence below 1% (Supplementary Figure S1 and S2). However, due to the combination of forward and reverse complement motifs in the 5-mer frequency calculation, UMAPclust lost its ability to identify sequence notation, resulting in clusters of around ten sequences rather than twenty (Supplementary Figure S3). When combined with isONclust or Meshclust, UMAPclust split sequences based on strand orientation, doubling the clustering purity (Figure 2). For high-accuracy ONT Duplex and PacBio CCS reads, the combination of isONclust and UMAPclust successfully generated 20 clusters with NMI scores equal to 1 in both even and skewed simulation scenarios when the target sequence divergence was above 1%. However, when clustering reads from highly similar targets (sequence divergence < 1%), the cluster purity decreased by 0.25 and 0.5 according to target sequences in cases of even or skewed abundance (Figure 2). Using Meshclust to consolidate UMAPclust clusters exhibited comparable NMI scores in the two high-accuracy datasets, while this combination outperformed others in clustering noisy ONT reads. Following the identity scores estimated by Meshclust, UMAPclust precisely clustered ONT2020 and ONT2023 reads from target sequences with sequence divergences above 5% and 2%, and it maintained clustering purity above 0.75 when the target sequence divergence decreased to 3% and 1%, respectively. The sequential use of isONclust, UMAPclust, and Meshclust did not significantly improve cluster purity. Instead, it fragmented the true clusters, though without severely hampering the NMI score. The selection of identity scores influenced both clustering accuracy and the yield of Meshclust. When employing the estimated

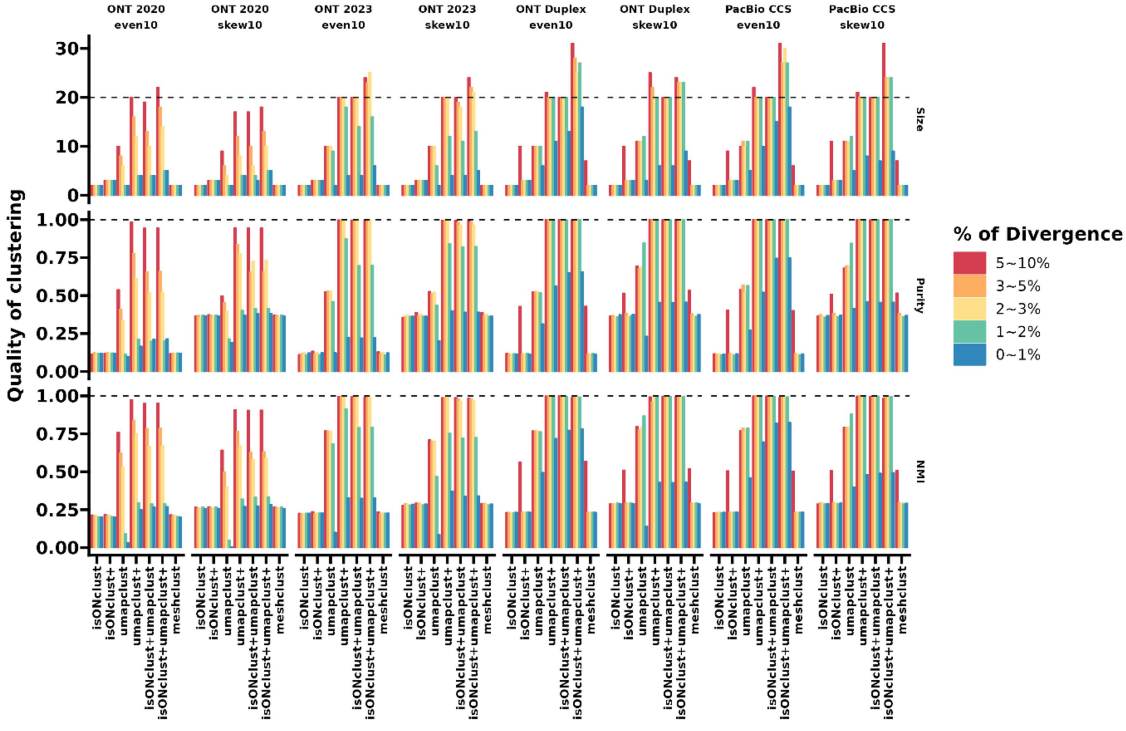

Figure 2. The clustering quality of various alignment-free approaches on the SILVA amplicons using an *in silico* sequencing depth of 200×. The clustering quality is assessed by size (the number of clusters), purity, and normalized mutual information (NMI). The black dashed lines suggest the theoretically optimal values. The sequence orientation (strand) information is considered in the clustering evaluation, resulting in the theoretical number of reference sequences of 20. The simulation template references are chosen from SILVA small subunits sequences with the divergence of 5–10%, 3–5%, 2–3% 1–2%, and 0–1%, and respective results are depicted by color. Even10 refers to a set of ten reference sequences with equal abundance; skew10 refers to ten reference sequences with two sequences 10 times as abundant as the rest. The plus sign suggests the clusters are validated with Meshclust.

identity score derived from the dataset, Meshclust retained approximately 75% of all types of long reads in the test. Clustering purity improved with the defined Meshclust identity score (Supplementary Figure S4), while a corresponding reduction in the yield of retained reads was observed (Supplementary Figure S5).

Utilizing the optimal alignment-free clustering strategy of UMAPclust and Meshclust with estimated identity scores, we benchmarked consensus calling of kmerCon, miniCon, and isoCon across a range of sequencing depths, spanning from 40× to 400× coverage. Except minCon, which exhibited a high rejection rate with noisy ONT2020 reads, LACA utilized over 75% of the reads for consensus calling, rarely incorporating chimeric sequences (Supplementary Figure S6). The sequencing identity of retained reads closely followed the defined distribution (Supplementary Figure S7). Recall of target sequences generally correlated with the clustering ability of various approaches. When the composition of target sequences was skewed, it resulted in a retardant recall of the complete set (Figure 3). In the kmerCon mode, consensus sequences were generated based on alignment-free read clusters. With 100× coverage, this approach could distinguish target sequences with divergences greater than 5%, 3%, 1%, and 1% for

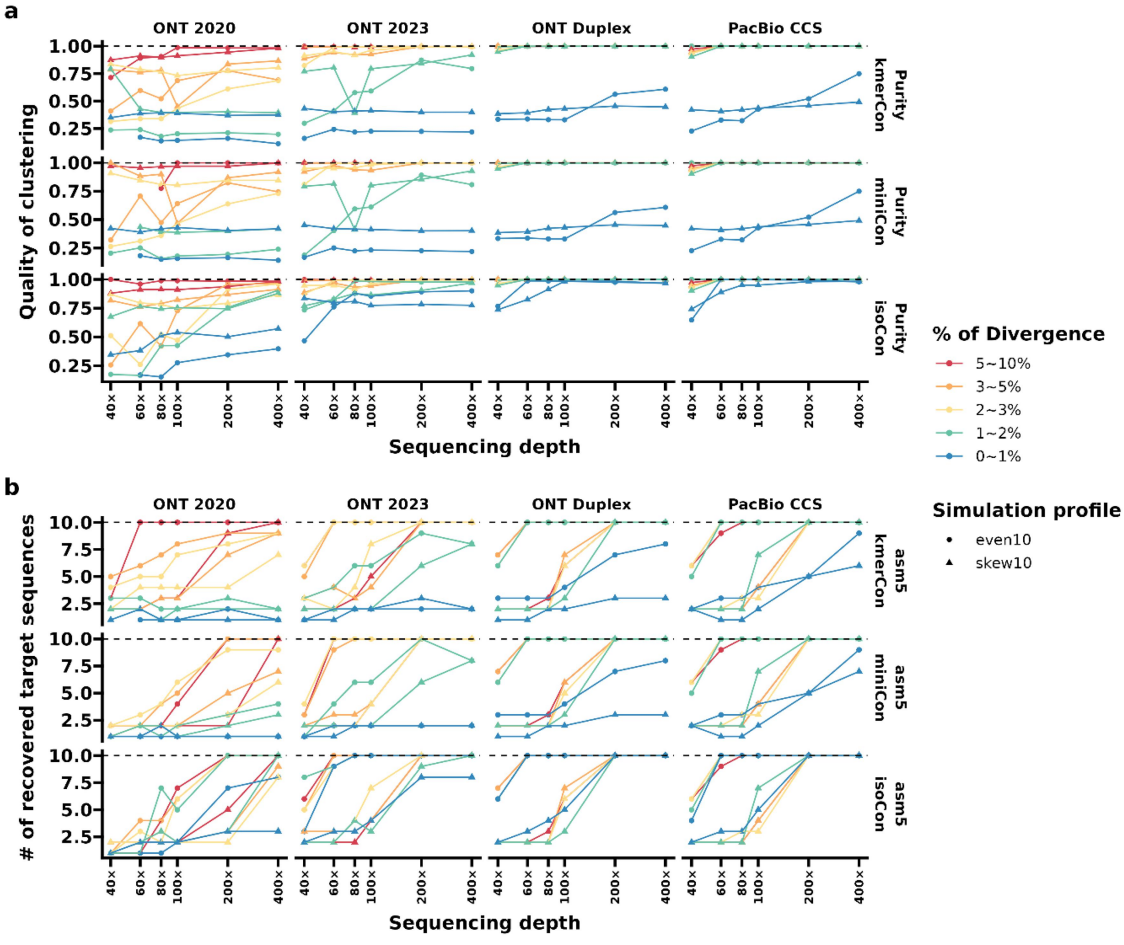

Figure 3. Clustering purity (a) and number of recovered target sequences (b) of the three different consensus calling approaches on the SILVA amplicons with an *in silico* sequencing depth from 40× to 400×. The black dashed lines suggest the theoretically optimal values, and the sequence orientation (strand) information is considered in the clustering evaluation, resulting in the theoretical number of reference sequences of 20. The simulation template references are chosen from SILVA small subunits sequences with the divergence of 5–10%, 3–5%, 2–3% 1–2%, and 0–1%, and respective results are depicted by color. The point shapes indicate the simulation scenarios of reference sequences in even (dots) and skewed (triangles) composition. kmercon, consensus calling on clustered sequences with UMAPclust and Meshclust; minicon, consensus calling on clustered sequences refined by overlap check; isocon, consensus calling of detected isoforms by IsoCon on the clustered sequences. Even10 refers to a set of ten reference sequences with equal abundance; skew10 refers to ten reference sequences with two sequences 10 times as abundant as the rest.

ONT2020, ONT2023, ONT Duplex, and PacBio CCS data, respectively (Figure 3(a)). Clustering purity and recall improved with sequencing depth if the target sequence divergence remained within the discernible range. The miniCon mode generated consensus sequences using refined clusters through alignment overlap. However, we did not observe significant improvements relative to the kmerCon mode. The isoCon mode proved most effective in retrieving the target sequence set. When sequencing depth exceeded 80× coverage, it successfully recovered all ten highly similar sequences (with sequence divergence less than 1%) in all "even10" simulated datasets except the noisy ONT2020 dataset (Figure 3(b) and Supplementary Figure S8). However, the clustering process might fragment the true clusters, and the number of clusters steadily increased with sequencing depth, exceeding 20 clusters at 400× coverage. This resulted in an approximate 10–30% decline in NMI scores across various benchmark profiles and data types (Supplementary Figure S9). The split reads by isoCon demonstrated strong concordance with the simulation source, with cluster purity plateauing at 1 for any type of simulated ONT Duplex and PacBio CCS data and showing a slight drop in the ONT2023 data (Figure 3(a)). In the ONT 2020 dataset, isoCon exhibited a cluster purity above 0.85 when identifying simulated reads from target sequences with sequence divergence above 1% at 400× coverage.

The quality of corrected sequences was primarily influenced by sequencing accuracy rather than sequencing depth. All corrected sequences exhibited significant fidelity to the reference sequence, showing a minimal alignment identity of 99% and an unalignment rate of 2% in most corrected ONT2020 sequences, and 99.9% and 1% for those derived from the remaining three datasets, respectively (Supplementary Figs. S10, S11, and S12). At 200× coverage, LACA controlled the average error rate (sum of variation rate in alignments and unalignment rate) of corrected sequences to 1%, 0.2% and 0.05% for ONT R9.4.1, ONT R10.4.1 and high accuracy ONT Duplex and PacBio data, respectively (Supplementary Table S1). Notably, due to the utilization of Medaka, consensus-based corrections demonstrated the most robust performance in ONT Duplex data. The alignment error rate and the rate of unaligned regions in corrected and reference sequences were consistently below 0.1% (Supplementary Figure S11, S12, and S13). Most corrected sequences covered over 99% of the reference sequences, except for a decline of 1–5% observed in most corrected ONT2020 sequences (Supplementary Figure S13). In our alignment comparisons, we observed that the overlap-refining strategy employed by miniCon improved the quality of consensus sequences, particularly in the case of noisy ONT2020 data. However, its impact diminished as sequencing accuracy improved. The implementation of Meshclust consistently ensured the quality of consensus calling. All corrected sequences obtained from ONT Duplex and PacBio CCS data demonstrated remarkably high alignment identity, surpassing 99% when compared to the source sequences (Supplementary Figure S11 and Supplementary Table 2). Under a more lenient control of Meshclust, supervised clusters by isONclust and UMAPclust led to a decline in identity between consensus sequences and source sequences at shallow depth, even when high-accuracy ONT Duplex and PacBio CCS reads were used (Supplementary Table S3).

Error profiling of clustering-based and UMI-binned consensus sequences

We extended our support for consensus calling by incorporating user-defined UMI patterns and compared these *de novo* clustering approaches with molecule-level UMI corrections in a public long-read UMI amplicon dataset of 16S-23S rRNA operons sequenced via PacBio CCS, ONT R9.4.1, and R10.3 platforms.[5] We subsampled 100,000 and 1,000,000 reads for each type of long-read amplicon data. Consistent with our *in silico* tests, the alignment-free clusters were further refined into micro-clusters by isoCon in the longer amplicon dataset (~4300 bp) (Figure 4(a)). In the case of shallow sequencing with 100,000 reads, the 16× isoCon sequences exhibited equivalent or even superior recall of mock operons compared to 4× UMI consensus sequences, while maintaining comparable accuracy (Figure 4). With sequencing depth ten times deeper, both methods successfully identified all 43 mock 16S-23S rRNA operons in

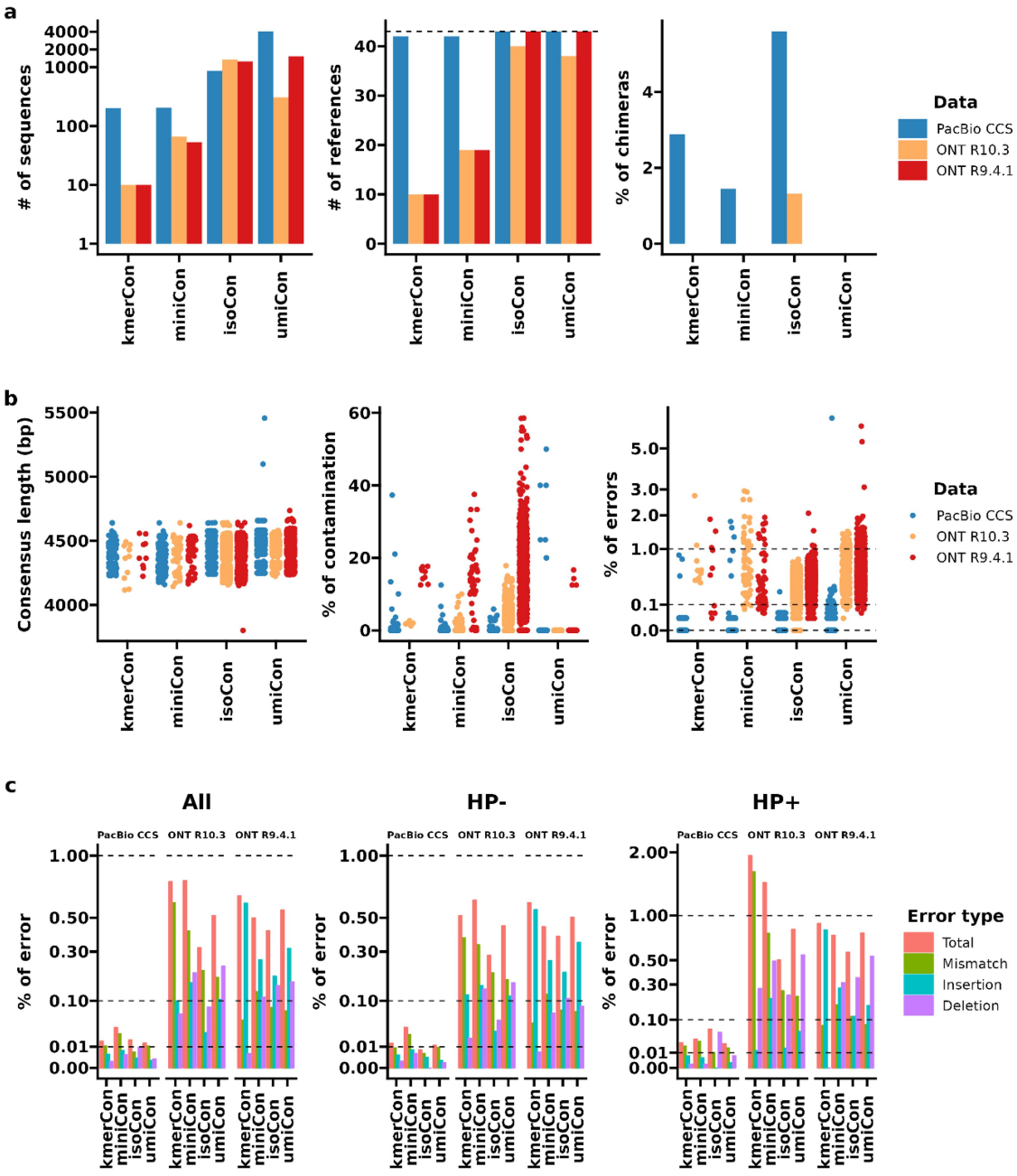

Figure 4. Statistics of quality-controlled amplicon consensus sequences retrieved from the subsampled 100,000 reads tagged with unique molecular identifiers (UMIs). (a) The number of consensus sequences and the identified reference sequences, and the chimera rate; (b) The sequence length, the percentage of reads aligned to a reference sequence other than the primary class (contamination), and the error rate of each consensus sequence; (c) The average error rate for mismatches, insertions, deletions, and total errors of retrieved consensus sequences, split by whether the error occurred inside (hp+) or outside (hp-) a homopolymer (hp) region. The corrected sequences are quality controlled with PCR artifacts removed and have recommended the minimum read coverage of 3×, 5×, 5× and 15× for the corrected sequences from umicon, kmercon, miniCon and isocon, respectively. To enhance visualization, we applied a base-10 logarithmic scaling to the number of consensus sequences in sub-figure (a) and employed square root scaling for the error rates in sub-figures (b) and (c). The three datasets are colored by the sequencing platforms in sub-figure (a) and (b), and the bar colors in sub-figure (c) display different error categories. kmercon, consensus calling on clustered sequences with Umapclust and Meshclust; minicon, consensus calling on clustered sequences refined by overlap check; isocon, consensus calling of detected isoforms by IsoCon on the clustered sequences; umicon, consensuse calling on UMI bins.

the three types of sequencing data (Supplementary Figure S14a). Notably, the number of detected mock operons by kmerCon and miniCon was predominantly determined by sequencing accuracy. Both methods retrieved over 90% of operon copies in the PacBio CCS dataset, while less than half were identified in the two ONT datasets (Figure 4(a) and Supplementary Figure S14a). The UMI correction method demonstrated superior control of chimeras, whereas corrected sequences based on sequence dissimilarity clusters contained 2–6% chimera, particularly in the PacBio CCS data.

All corrected sequences exhibited a consistent length distribution, approximately 4500 bp in size. UMI clusters displayed relatively low contamination, especially in the noisy ONT R9.4.1 data, while contamination rates were comparable across all consensus-calling approaches in the high-accuracy PacBio CCS data (Figure 4(b) and Supplementary Figure S14b). All error-correction strategies demonstrated remarkably low error rates below 0.1% in the accurate PacBio CCS reads after excluding a few small clusters (Supplementary Figure S15 and S16). Applying an acceptable cutoff of 5× for kmerCon and miniCon, and 15× for isoCon, these consensus correction approaches achieved overall error rates similar to those of 4× UMI corrections in the noisy ONT dataset (Figure 4(b)). IsoCon consistently maintained error rates below 1%, which were 3–5 times higher than those of UMI corrections at the recommended coverage cutoff (Supplementary Figure S14b). The error rates of isoCon and umiCon sequences decreased with read coverage, both converging to below 1% (Supplementary Figure S15 and 16). With deeper sequencing, the error rate of umiCon fell below 0.1%, while the error rate of isoCon remained stable (Supplementary Figure S16). At a shallow depth of 100,000 reads, the average overall error rate for 4× umiCon and 15× isoCon consensus sequences was 0.014% and 0.017% on PacBio CCS data, 0.55% and 0.42% on ONT R9.4.1 data, and 0.52% and 0.32% on ONT R10.3 data (see Figure 4(c) and Supplementary Table S4), respectively. When the sequencing went ten times deeper, umiCon and isoCon sequences exhibited average error rates of 0.011% and 0.027% on PacBio CCS data, 0.071% and 0.401% on ONT R9.4.1 data, and 0.028% and 0.31% on ONT R10.3 data (Supplementary Figure S14 and Supplementary Table S5), respectively.

The error types, including deletions, insertions, and mismatches, exhibited patterns specific to their data sources, with slight variations based on the consensus-calling approaches employed. The consensus sequences derived from PacBio CCS data displayed a distinct low error rate, showing no obvious differences in error types based on the presence of homopolymers. Mismatches predominantly contributed to errors, except for a notable presence of remaining deletion errors in homopolymer regions of the corrected isoCon sequences (Figure 4(c) and Supplementary Figure S14). Insertions in non-homopolymer regions were the most prevalent errors in ONT R9.4.1 consensus sequences, whereas mismatch errors were more common in ONT R10.3 corrected sequences (Supplementary Figure S14). For ONT UMI consensus sequences, deletions in homopolymer regions and insertions in non-homopolymer regions were the major error sources (Figure 4(c)). Similarly as in the PacBio CCS data, isoCon exhibited limited control over deletion errors in the ONT data (Supplementary Figure S14 and Supplementary Table S5). The miniCon and kmerCon sequences however exhibited significantly higher levels of insertion errors in homopolymer regions when contrasted with other methods.

ONT 16S rRNA gene amplicon sequencing of a synthetic mock microbiome

We followed a two-step PCR approach to prepare the ONT-based 16S rRNA gene amplicon library for microbiome profiling using metabarcoding.[26] We conducted a series of dilutions of ZymoBIOMICs mock DNA to test the multi-primer amplification library preparation and sequencing. At both sequencing depths of 10,000 and 100,000 reads per sample, all three consensus-calling methods effectively identified three primary OTU clusters, characterized by a length range from 900 to 1400 bp (Figure 5(a)). These OTUs corresponded to the potential combinations of amplified hypervariable regions within the 16S rRNA operons, spanning from V1 and V3 to V8 and V9. Approximately half of the reads from each mock

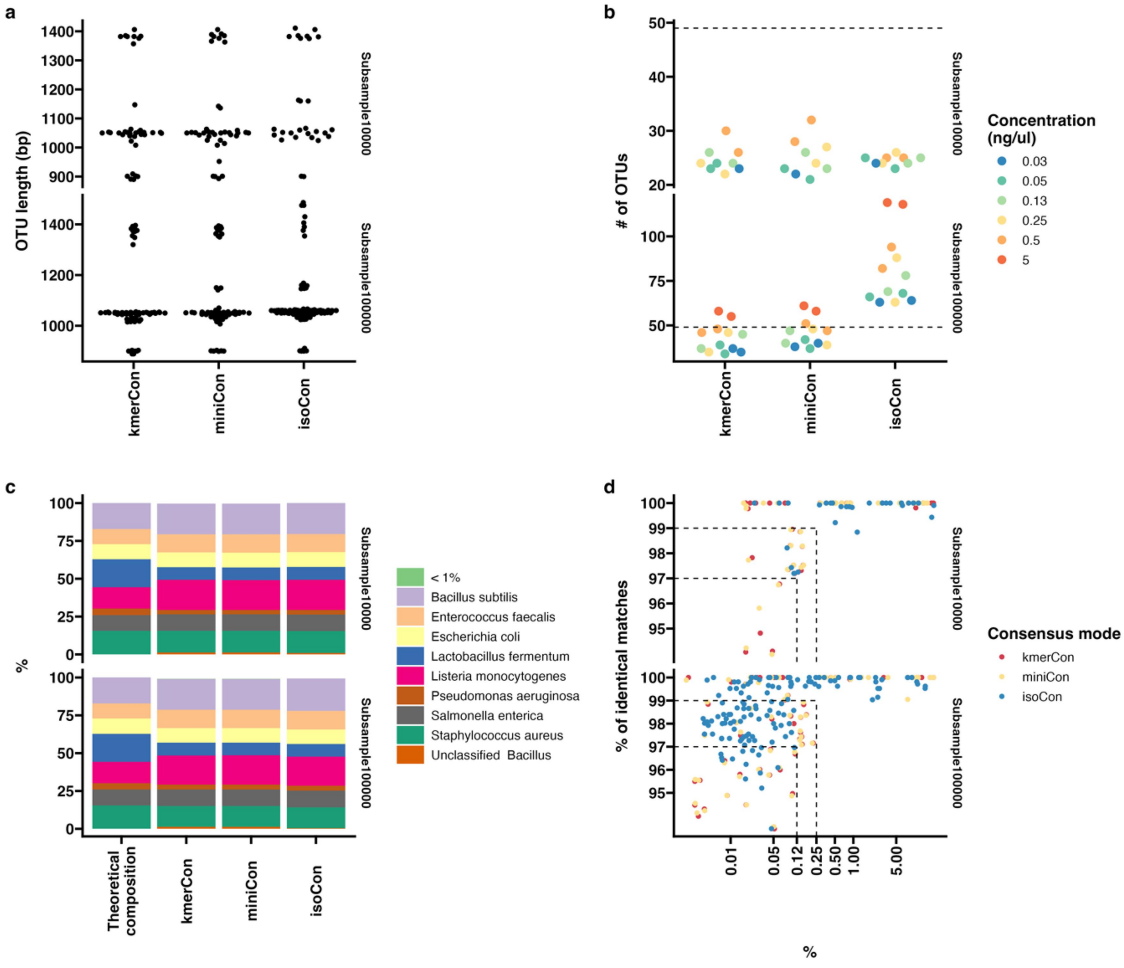

Figure 5. Near full-length 16S rRNA gene profiling of a mock dilution series through Oxford Nanopore sequencing. a, b, the length (a) and the number (b) of recovered OTUs by various OTU-picking approaches on mocks with subsampled reads of 10,000 and 100,000 per sample; (c) species-level community profile and theoretical composition; (d) BLAST identity and relative abundance of picked OTUs. To enhance visualization, we applied a base-10 logarithmic scaling to the relative abundance of OTUs in sub-figure (d). The OTUs are colored by the mock concentration in the dilution series in the sub-figure (b) while the color in sub-figure (d) indicates the consensus calling modes. Taxa with a mean abundance of less than 1% are marked in sub-figure (c) as "< 1%". kmercon, community profile based on clustered sequences with UMAPclust and Meshclust; minicon, community profile based on clustered sequences refined with overlap check; isocon, community profile based on detected isoforms by IsoCon on the clustered sequences.

sample passed quality control for community profiling (Supplementary Figure S17), and almost no OTUs were found in the negative controls (Figure 5(b)). At the shallow sequencing depth of 10,000 reads per sample, the number of OTUs consistently remained below the total number of 49 mock 16S rRNA operons reported by the manufacturer. Nevertheless, with a tenfold increase in sequencing depth, kmerCon and miniCon generated 61 and 64 high-quality OTUs, respectively. Consistent with the results of the *in silico* test, isoCon split the 5-mer clusters and nearly tripled the theoretical number of OTUs (Supplementary Table S6). All eight mock strains were identified by all approaches at the species level (Figure 5(c)). Some variations in relative abundance were observed, possibly due to PCR primer preferences, leading to an increase in *Limosilactobacillus fermentum* and a decrease in *Listeria monocytogenes* relative abundance in comparison to the theoretical composition. The retrieved OTUs demonstrated a high concordance with SILVA SSUs (Supplementary Figure S18). The dominant OTUs, i.e., the OTUs with minimal relative abundances of 0.25%, exhibited a species-level BLAST alignment identity of over 99% to the SILVA SSUs (Figure 5(d)). These OTUs collectively accounted for over 97% of the total read counts

(Supplementary Table 6). As sequencing depth grew, the profiling resolution of different mock DNA groups was improved with reduced inter-sample variance between technical replicates in the Bray Curtis distance metrices (Supplementary Figure S19). The inter-group difference was eliminated when the phylogenetic relationship is integrated into beta-diversity analysis using the weighted UniFrac dissimilarity metric (Supplementary Figure S19).

Real-world application in human vaginal microbiomes

To assess LACA's performance in a real-world application, we utilized a public ONT 16S rRNA gene amplicon dataset[10] of 12 human vaginal microbiomes composed of six controls and six patients with diagnosed bacterial vaginosis, and employed the read-classification approach Emu[10] as a comparison. These approaches consistently identified the dominant bacteria on the vaginal swaps at genus (Supplementary Figure S20) and species level, except that some species-level signals were conservatively annotated to the genus level by LACA given the use of LCA algorithm (Supplementary Figure S21). In comparison to vaginal microbiomes from the healthy controls, all methods revealed a significant increase of non-lactobacilli anaerobes, including *Aerococcus*, *Prevotella*, and *Megasphaera*, in samples from patients with bacterial vaginosis (Supplementary Figure S21 and S22). The healthy vaginal microbiome exhibited distinct compositional differences compared to the vaginosis microbiome, with the former predominantly dominated by lactobacilli species, with the exception of one atypical vaginosis sample dominated by *L. crispatus* (Supplementary Figure S21). Owing to the phylogenetic information preserved in OTUs, the distinction between these two sample categories was more pronounced when evaluated using weighted UniFrac metrics, both with (Figure 6) and without including the atypical vaginosis sample (Supplementary Figure S23). UniFrac metrics demonstrated superior robustness compared to exclusively abundance-based Bray Curtis metrics in discriminating between sample types, regardless of whether atypical samples were included or not.

Emu identified approximately 3 times more taxonomic features at the species level and 0.4 times more at the genus level compared to LACA (Supplementary Figure S24). KmerCon was superior in identifying taxonomic traits at the genus and species levels among the LACA profiles, followed by miniCon and isoCon. The shared features in the LACA and Emu profiles showed a strong positive correlation in relative abundance, with an average Pearson's correlation coefficient above 0.95 ($p < 0.05$, Supplementary Figure S25). Although isoCon identified the fewest taxonomic features, the genus- and species-level abundances were most correlated with the estimated abundance through re-mapping reads against the isoCon-derived OTU sequences (Supplementary Figure S26). The Pearson's correlation between clustering-based and alignment-based abundance declined at the OTU level while the Spearman rank test still indicated a strong correlation ($r > 0.5$, Supplementary Figure S26). Since LACA OTUs were annotated by the BLAST hits-based LCA algorithm, we searched the hit table for the uncovered taxonomic features by LACA. Among all genus- and species-level features identified by Emu, over 70% and 65% were found in the BLAST hit table against SILVA SSUs with a minimum alignment identity above 97% and 99%, respectively (Supplementary Table S7). The mean relative abundance of the uncovered genera and species was below 0.1%, except for *Criibacterium bergeronii* and the *Rikenellaceae* RC9 gut group, where it was 1.43% and 0.24%, respectively (Supplementary Table S8). In BLAST search against the curated EzBioCloud database for 16S rRNA gene sequences, we found that the OTUs assigned to the *Peptostreptococcaceae* family showed over 99.1% identity to the only *Criibacterium bergeronii* strain (CCRI-22567) in the database (Supplementary Table S9) and the identity of the remaining *Peptostreptococcaceae* matches were below 90%. LACA identified the *Rikenellaceae* RC9 gut group, but the relevant OTUs were excluded in analysis due to a low alignment identity of around 90% among all hits against SILVA SSUs. The LCA assignment of one OTU sequence was influenced by the

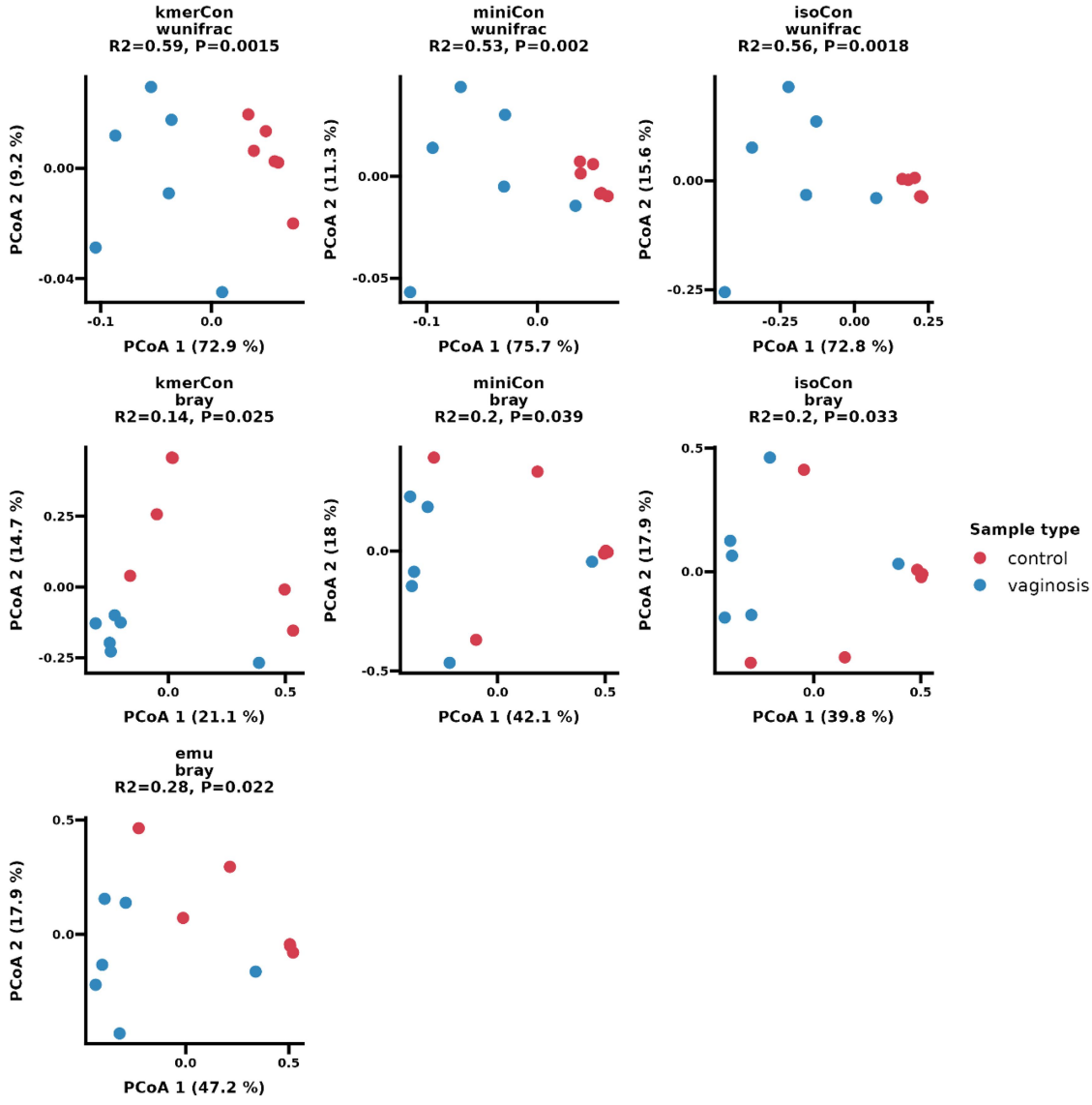

Figure 6. Principal coordinate analysis (PCoA) plots of weighted UniFrac and Bray-Curtis dissimilarity metrics using various profiling approaches on the human vaginal samples. The points are colored by the sample phenotype. Emu, read classification profile with Emu; kmercon, community profile based on clustered sequences with UMAPclust and Meshclust; minicon, community profile based on clustered sequences refined with overlap check; isocon, community profile based on detected isoforms by IsoCon on the clustered sequences.

redundant and un-curated species-level taxonomy in the SILVA database and the homogeneity of 16S rRNA gene copies among closely related species. By default, LACA employs LCA taxonomic assignment on ten BLAST hits to maintain conservative inference. Consequently, despite species-level BLAST hits exhibiting high sequence similarity (>99% identity, Supplementary Table S10), the lactobacilli OTUs from the control vaginal samples (SRR14307925 and SRR14307926, Supplementary Figure S21) were not annotated as *Lactobacillus iners*, owing to the inclusion of genus-level hits for taxonomic determination. Meanwhile, Emu missed some common vaginal member such as *Metamycoplasma hominis*, *Prevotella melaninogenica*, and *Megasphaera elsdenii*, relative to the LACA approach (Supplementary Table S11). The relative abundances of the uncovered features were below 0.1%, except for *Lacticaseibacillus rhamnosus* and *Limosilactobacillus fermentum*, which accounted for 0.2% and 0.7%, respectively.

Discussion

Utilizing a *de novo* OTU picking strategy, LACA offers a full microbiome analysis suite tailored for long-read amplicon analysis, spanning from demultiplexing to taxonomic quantification. Unlike read classification approaches, *de novo* OTU picking enables novel species detection without reference bias, and phylogenetic information is retained, offering feasibility in microbiome meta-analysis.

The variable sequencing error rate poses a challenge for OTU picking from long-read amplicons. While these sequencing errors are generally randomly distributed,[38,39] they can obscure true variants in highly similar regions. To address this, LACA benchmarked multiple alignment-free and alignment-based methods, including the use of short k-mers[13,14] or error-aware clustering with long minizers,[11] or iterative clustering through alignment.[15] In our test, we found the clustering choice was determined by the sequence divergence of the target region and the basecalling accuracy of the applied technologies. To avoid the unnecessary computational cost of pairwise alignment, LACA first clusters sequences with alignment-free approaches. Among these, UMAPclust, incorporating 5-mer frequency information, displayed superior efficiency in clustering long 16S rRNA gene amplicons with few false positives (Figure 1). In combination with Meshclust,[19] UMAPclust accurately identified ONT R9.4.1 16S rRNA gene amplicons with minimal sequence divergence of 5%. The resolution improved to 2% and 1% for ONT R10.4.1 and high-accuracy ONT Duplex and PacBio CCS data, respectively (Figure 2). UMAPclust relies on the use of short 5-mers, which are highly sensitive to point variations, making it effective in many cases. However, it is prone to producing distorted distributions when insertions and deletions are introduced.[40] It is a significant concern in long-read technologies since indels serve as common sources of errors.[41] Thereby, its performance significantly improved in the latest ONT R10.4.1 flow cells using dual-reader nanopores to reduce indel calling errors in the homopolymer region. The alignment-free clusters could be further refined through alignment-based approaches. Combining isONcorrect and isoCon improved cluster purity and extended identifiable sequence divergence to 1% for ONT R9.4.1 data and below 1% for ONT R10.4.1, high-accuracy Duplex and PacBio CCS data. With this refinement, the error rate of consensus sequences was controlled below 1% for ONT R9.4.1 data, 0.2% for ONT R10.4.1 and 0.1% for high-accuracy long-read data (Supplementary Figure S11). This in combination with the preserved phylogenetic information makes the approach well-suited for precise microbiome analysis.

While full-length 16S rRNA gene amplicon sequencing offers higher taxonomic resolution than short-read approaches, particularly for distinguishing closely related species, it is important to interpret its capabilities with appropriate caution. Although *in silico* studies suggest that long-read sequencing may allow classification at the subspecies or strain level, this remains challenging due to the inherently conserved nature of the 16S rRNA gene.[4] The decline in LACA's performance with decreasing sequence divergence observed in our study highlights a key limitation of 16S-based taxonomic classification: its restricted ability to resolve very minor differences (often just a few nucleotides) required for resolution below the species level. In practical applications, sequencing accuracy and depth emerge as the two critical factors determining LACA's performance. Sequence error typically constitutes the performance bottleneck of cluster-based variant analysis, including LACA. Reduced error rates enable more effective clustering with shallower sequencing. In our benchmark, 5-mer UMAP clustering failed to discriminate between closely related sequences with divergence below 5% at 200× coverage using error-prone ONT R9.4.1 data, yet succeeded with more accurate R10.4.1 data at just 60× coverage (Figure 2). When target sequence divergence approaches the identifiable threshold, deeper sequencing becomes essential to ensure clustering accuracy. For studies conducted at fixed sequencing depths, clustering performance may deteriorate for certain target sequences when sequencing coverage is insufficient. Therefore, carefully balancing target sequence divergence against the technical limitations of sequencing technologies and LACA's capabilities is crucial for ensuring result quality. For closely related target regions (~1%

divergence), highly accurate sequencing approaches such as ONT Duplex or PacBio CCS are always recommended. Furthermore, our findings indicate that LACA's classification accuracy varies according to community composition, as demonstrated by performance differences between even and skewed simulation scenarios. These suggest that real-world performance may vary across different microbial environments (e.g., diverse gut communities versus lactobacilli-dominated vaginal microbiota) or health states. For a simple evenly distributed eight-strain mock community, a subsampling depth of 10,000 reads per sample proved sufficient for accurate community profiling (Figure 5c). In skewed communities, deeper sequencing necessarily ensures adequate sampling of rare species (Figure 2). The vaginal microbiome study[10] utilized a single MinION R9 flow cell to sequence 12 samples of 16S rRNA amplicons (approximately 0.8 G bases on average), achieving detailed profiling but with substantial sequencing costs. These considerations should be carefully evaluated when applying LACA in specific research contexts to ensure reliable taxonomic classification and accurate ecological interpretation.

To mitigate long-read sequencing errors, specialized library preparation methods such as UMI[5] and rolling circle amplification[42] have been implemented to ensure molecular-level precision. Nonetheless, molecular-level correction requires sufficient sequencing depth for each individual molecule. In the case of sequencing multi-copy gene families like rRNA genes, the use of UMI tags can result in redundant tagging of identical copies originating from a single template. This leads to a significant increase in sequencing costs, typically requiring optimal dilution strategies to strike a balance between adequate sampling and efficient utilization of sequencing throughput. Determining the appropriate dilution level is challenging, particularly when dealing with unfamiliar ecosystems like soil samples.[43] In shallow sequencing, corrections through sequence similarity clustering demonstrated comparable or even superior sensitivity and error control compared to 4× UMI corrections to identify highly similar 16-23S rRNA gene operons (Figure 3(a)). Despite no further error reduction with increased sequencing depth, most corrected sequences maintained acceptable error rates below 1% for species-level alignment. On high-fidelity PacBio CCS data, the UMI correction and clustering-based approaches exhibited a similar ability to reduce errors to below 0.1% with adequate coverage cutoffs. In real use of Nanopore sequencing, the sequencing hardware and chemistry affect the identification resolution of *de novo* clustering approaches, given the variance of sequencing accuracy. Using both high accuracy (HAC) and super accuracy basecalling (SUP) with the latest ONT basecaller Dorado, the R10.4.1 flow cell achieved > 99% single-read accuracy, which corresponded to our simulated quality of high-fidelity raw reads. Given the limited accuracy improvement (~0.5%) and computationally intense requirement of SUP, HAC is officially recommended for high-throughput projects focusing on variant analysis, including amplicon sequencing (https://nanoporetech.com/platform/accuracy). Besides, the ONT Duplex kit is expected to offer single-molecule resolution with an average error rate of Q30. This shall further broaden the utility of clustering-based corrections for long amplicon analysis.

The OTUs generated through *de novo* clustering faithfully retain the phylogenetic information originating from the profiled samples. Based on the weighted UniFrac distance metric, we found that these long OTUs effectively contribute to characterizing sample phenotypes, even when working with the ONT 9.4.1 flow cell where the error rate can range from 5–15%. In a beta-diversity analysis of a synthetic mock community, significant differences in Bray Curtis dissimilarity were observed due to variations in PCR template concentrations. These differences became more pronounced with deeper sequencing, while disparities arising from technical replicates diminished. Stochastic fluctuations in priming efficiency, a consequence of varying template concentrations, likely contributed to this phenomenon.[44,45] However, the incorporation of evolutionary information through UniFrac distance metrics effectively mitigated these differences. Another noteworthy example is the characterization of vaginal microbiomes. Healthy vaginal microbiomes are typically dominated by

lactobacilli.[46] The phylogenetic resemblance in the amplification region led to improved clustering of these samples when evaluated using the weighted UniFrac distance metric. This phylogenetic weighted clustering facilitated a clear separation between control samples and those associated with vaginosis, surpassing even the Emu profile, which incorporated more taxonomic features based on read classification.

While clustering-based approaches exhibited decreased sensitivity in detecting rare species compared to read classification tools such as Emu, LACA effectively identified all dominant microorganisms (>0.1%) at both the genus and species levels, demonstrating strong agreement between the two methods. *De novo* picked OTUs require a minimum cluster size to distinguish the true biological unit from spurious sequences, rendering them predisposed to potentially overlooking subtle signals from rare species. On the other hand, augmented sensitivity of profiling with read classification carries a higher risk of generating false positives. The original study[10] notes escalated false-positive counts of Emu in characterizing a synthetic community relative to NanoCLUST,[14] despite Emu using an expectation-maximization algorithm to iteratively refine species-level abundance. This discrepancy could be attributed to mapping errors introduced during the initial alignment of long, noisy reads against highly similar but redundant gene sets, such as rRNA operons. In our analysis, we encountered situations where the LCA taxonomy for certain OTUs could not be definitively resolved at the species level, even though they exhibited over 99% BLAST identity to reference sequences (as observed with *Lactobacillus iners* and *Criibacterium bergeronii* in vaginal microbiomes). The underlying causes of this issue are related to the sequence homogeneity of closely related species in the amplification region and a substantial portion (~72%) of the undetermined species labels in the SILVA database.[47] The LCA assignment is determined by the lowest taxonomic categories shared by the hit matches, wherein genus-level hits may compromise species-level resolution (Supplementary Table S10). Consequently, both the reference database and the criteria established for accepting meaningful hit matches significantly influence the sensitivity and accuracy of taxonomic classification. Employing a less redundant database with updated species-level information[48] or a habitat-specific curated database[49] can enhance species-level resolution. Conversely, utilizing databases with ambiguous or inaccurate species-level information may lead to false taxonomic inference. Analogous findings have been reported in efforts to utilize the complete 16S-23S rRNA gene operon for species-level assignments,[5] although improved species-level resolution has been observed with the increase in amplicon length. The database challenges are likely to persist for the foreseeable future. Like other *de novo* clustering approaches, LACA preserves the OTU sequences, allowing subsequent reannotation with updated databases or interrogation with expert knowledge.

Conclusions

LACA provides a versatile, reproducible, and scalable workflow for *de novo* OTU analysis of lengthy noisy amplicon data. In our assessment, the clustering-based OTUs effectively preserved the essential phylogenetic information for long amplicon microbiome analysis, with an acceptable trade-off in accuracy. Although we benchmarked LACA with rRNA operons for microbiome analysis, these clustering-based corrections can also be applied to other types of amplicon data within a discernible range. Furthermore, LACA's flexibility extends to its configuration file, making it easy to adapt to novel models and flow cells as sequencing technologies continue to evolve.

Acknowledgments

We acknowledge the high-performance computing resources provided by Danish national supercomputer for life sciences (Computerome, https://computerome.dk).

Author contributions

CRediT: **Yan Hui:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Dennis Sandris Nielsen:** Conceptualization,

Resources, Writing – review & editing; **Lukasz Krych:** Conceptualization, Data curation, Investigation, Resources, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work is supported by the Starting Research Fund from Hangzhou Normal University (4285C50225204011).

ORCID

Yan Hui http://orcid.org/0000-0002-2388-131X
Dennis Sandris Nielsen http://orcid.org/0000-0001-8121-1114
Lukasz Krych http://orcid.org/0000-0003-3224-1346

Data availability statement

The near full-length amplicon Nanopore sequencing results of the ZymoBIOMICS microbial community standard D6306 were deposited in the NCBI Sequence Read Archive as Bioproject PRJNA1036680. Public data used in this investigation include the ZymoBIOMICS microbial community standard D6300 UMI-tagged sequencing data from the European Nucleotide Archive under the project PRJEB32674, the 12 vaginal samples in real-world application from Sequence Read Achieve under the project PRJNA723982 and the SILVA 138.1 SSURef Nr99 database (https://www.arb-silva.de).

LACA is open-source and available at https://github.com/yanhui09/laca. The LACA v0.3.0 was used to process the long amplicon reads in this paper. NART is open-source and available at https://github.com/yanhui09/nart. The data analysis scripts can be accessed at https://github.com/yanhui09/laca_archive.

References

1. Flaherty BR, Barratt J, Lane M, Talundzic E, Bradbury RS. Sensitive universal detection of blood parasites by selective pathogen-DNA enrichment and deep amplicon sequencing. Microbiome. 2021;9(1):1. doi:10.1186/s40168-020-00939-1.
2. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Abnet CC, et al. Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (MBQC) project consortium. Nat Biotechnol. 2017;35(11):1077–1086. doi: 10.1038/nbt.3981.
3. Adriaenssens EM, Cowan DA, Wood TK. Using signature genes as tools to assess environmental viral ecology and diversity. Appl Environ Microbiol. 2014;80(15):4470–4480. doi: 10.1128/AEM.00878-14.
4. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10(1):1–11. doi: 10.1038/s41467-019-13036-1.
5. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods. 2021;18(2):165–169. doi: 10.1038/s41592-020-01041-y.
6. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–1365. doi:10.1038/s41587-021-01108-x.
7. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics, Proteomics & Bioinf. 2015;13(5):278–289. doi:10.1016/j.gpb.2015.08.002.
8. Zhang T, Li H, Ma S, Cao J, Liao H, Huang Q, Chen W. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. Appl Environ Microb. 2023;89(10):e00605–23. doi: 10.1128/aem.00605-23.
9. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721–1729. doi: 10.1101/gr.210641.116.
10. Curry KD, Wang Q, Nute MG, Tyshaieva A, Reeves E, Soriano S, Wu Q, Graeber E, Finzer P, Mendling W, et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat Methods. 2022;19(7):845–853. doi: 10.1038/s41592-022-01520-4.
11. Sahlin K, Medvedev P. De Novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm. J Comput Biol. 2020;27(4):472–484. doi:10.1089/cmb.2019.0299.
12. Sahlin K, Lim MCW, Prost S. NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. Ecol Evol. 2021;11(3):1392–1398. doi: 10.1002/ece3.7146.
13. Beaulaurier J, Luo E, Eppley JM, Uyl PD, Dai X, Burger A, Turner DJ, Pendelton M, Juul S, Harrington E, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res. 2020;30(3):437–446. doi: 10.1101/gr.251686.119.
14. Rodríguez-Pérez H, Ciuffreda L, Flores C, Inanc B. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. Bioinformatics. 2021;37(11):1600–1601. doi: 10.1093/bioinformatics/btaa900.

15. Sahlin K, Tomaszkiewicz M, Makova KD, Medvedev P. Deciphering highly similar multigene family transcripts from iso-seq data with IsoCon. Nat Commun. 2018;9(1):1–12. doi: 10.1038/s41467-018-06910-x.
16. Sahlin K, Sipos B, James PL, Medvedev P, McNeil K. Anomalous collapses of Nares Strait ice arches leads to enhanced export of Arctic sea ice. Nat Commun. 2021;12(1):1–13. doi: 10.1038/s41467-020-20340-8.
17. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–857. doi: 10.1038/s41587-019-0209-9.
18. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. Sustainable data analysis with Snakemake. F1000 Res. 2021;10:33. doi: 10.12688/f1000research.29032.2.
19. Girgis HZ. MeShClust v3.0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. BMC Genomics. 2022;23(1):423. doi: 10.1186/s12864-022-08619-0.
20. James BT, Luczak BB, Girgis HZ. MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Res. 2018;46(14):e83. doi: 10.1093/nar/gky315.
21. Girgis HZ, James BT, Luczak BB. Identity: rapid alignment-free prediction of sequence alignment identity scores using self-supervised general linear models. NAR Genomics Bioinf. 2021;3(1):lqab001. doi: 10.1093/nargab/lqab001.
22. Vaser R, Sović I, Nagarajan N, M Š. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–746. doi: 10.1101/gr.214270.116.
23. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–1028. doi:10.1038/nbt.3988.
24. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41(D1):D590–D596. doi: 10.1093/nar/gks1219.
25. Wick RR. Badread: simulation of error-prone long reads. J Open Source Softw. 2019;4(36):1316. doi: 10.21105/joss.01316.
26. Hui Y, Tamez-Hidalgo P, Cieplak T, Satessa GD, Kot W, Kjærulff S, Nielsen MO, Nielsen DS, Krych L. Supplementation of a lacto-fermented rapeseed-seaweed blend promotes gut microbial-and gut immune-modulation in weaner piglets. J Anim Sci Biotechnol. 2021;12(1):85. doi: 10.1186/s40104-021-00601-2.
27. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–12. doi: 10.14806/ej.17.1.200.
28. Marijon P, Chikhi R, Varré JS, Birol I. Yacrd and fpa: upstream tools for long-read genome assembly. Bioinformatics. 2020;36(12):3894–3896. doi: 10.1093/bioinformatics/btaa262.
29. McInnes L, Healy J, Astels S. Hdbscan: hierarchical density based clustering. J Open Source Softw. 2017;2(11):205. doi: 10.21105/joss.00205.
30. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. J Open Source Softw. 2018;3(29):861. doi: 10.21105/joss.00861.
31. Price MN, Dehal PS, Arkin AP, Poon AFY. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLOS ONE. 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490.
32. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. England: Cambridge University Press; 2008.
33. Sundqvist M, Chiquet J, Rigaill G. Adjusting the adjusted rand index – a multinomial story (United States: Cornell Tech). 2020. doi: 10.48550/arXiv.2011.08708.
34. Li H, Birol I. Minimap2: pairwise alignment for nucleotide sequences. Birol I, ed. Bioinformatics. 2018;34(18):3094–3100. doi: 10.1093/bioinformatics/bty191.
35. Edgar R. UCHIME2: improved chimera prediction for amplicon sequencing. Biorxiv. 2016; doi: 10.1101/074252.
36. McMurdie PJ, Holmes S, Watson M. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLOS ONE. 2013;8(4):e61217. doi: 10.1371/journal.pone.0061217.
37. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graphics. 2014;20(12):1983–1992. doi: 10.1109/TVCG.2014.2346248.
38. Magi A, Giusti B, Tattini L. Characterization of MinION nanopore data for resequencing analyses. Briefings Bioinf. 2017;18(6):940–953. doi: 10.1093/bib/bbw077.
39. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–569. doi: 10.1038/nmeth.2474.
40. Blanca A, Harris RS, Koslicki D, Medvedev P. The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches. J Comput Biol: J Comput Mol Cell Biol. 2022;29(2):155–168. doi: 10.1089/cmb.2021.0431.
41. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. Oxford nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19(7):823–826. doi: 10.1038/s41592-022-01539-7.

42. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc Natl Acad Sci USA. 2018;115(39):9726–9731. doi: 10.1073/pnas.1806447115.
43. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol. 2017;15(10):579–590. doi:10.1038/nrmicro.2017.87.
44. Chandler DP, Fredrickson JK, Brockman FJ. Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. Mol Ecol. 1997;6(5):475–482. doi: 10.1046/j.1365-294x.1997.00205.x.
45. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 2015;43(21):e143. doi:10.1093/nar/gkv717.
46. Gajer P, Brotman RM, Bai G, Sakamoto J, Schütte UME, Zhong X, Koenig SSK, Fu L, Ma Z(, Zhou X, et al. Temporal dynamics of the human vaginal microbiota. Sci Transl Med. 2012;4(132):132ra52. doi: 10.1126/scitranslmed.3003605.
47. Robeson MS, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, Bokulich NA. Rescript: reproducible sequence taxonomy reference database management. PLOS Comput Biol. 2021;17(11):e1009581. doi: 10.1371/journal.pcbi.1009581.
48. Cabezas MP, Fonseca NA, Muñoz-Mérida A. Mimt: a curated 16S rRNA reference database with less redundancy and higher accuracy at species-level identification. Environ Microbiome. 2024;19(1):88. doi: 10.1186/s40793-024-00634-w.
49. Escapa I F, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, Lemon KP. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. 2020;8(1):65. doi: 10.1186/s40168-020-00841-w.