


# High accuracy meets high throughput for near full-length 16S ribosomal RNA amplicon sequencing on the Nanopore platform

Xuan Lin <sup>a</sup>, Katherine Waring <sup>a</sup>, Hans Ghezzi <sup>b</sup>, Carolina Tropini <sup>b,c,d,e</sup>, John Tyson <sup>f,g</sup> and Ryan M. Ziels <sup>a,\*</sup>

<sup>a</sup>Civil Engineering, The University of British Columbia, 6250 Applied Science Ln #2002, Vancouver, BC, Canada V6T 1Z4

<sup>b</sup>Graduate Program in Bioinformatics, The University of British Columbia, Vancouver, BC, Canada V5Z 4S6

<sup>c</sup>Department of Microbiology and Immunology, The University of British Columbia, Vancouver, BC, Canada V6T 1Z3

<sup>d</sup>School of Biomedical Engineering, The University of British Columbia, Vancouver, BC, Canada V6T 2B9

<sup>e</sup>Humans and the Microbiome Program, Canadian Institute for Advanced Research (CIFAR), Toronto, ON, Canada M5G 1M1

<sup>f</sup>British Columbia Center for Disease Control Public Health Laboratory, Vancouver, BC, Canada V5Z 4R4

<sup>g</sup>Pathology and Laboratory Medicine, The University of British Columbia, Vancouver, BC, Canada V6T 1Z7

\*To whom correspondence should be addressed: Email: [ziels@mail.ubc.ca](mailto:ziels@mail.ubc.ca)

Edited By Panayiotis Benos

## Abstract

Small subunit (SSU) ribosomal RNA (rRNA) gene amplicon sequencing is a foundational method in microbial ecology. Currently, short-read platforms are commonly employed for high-throughput applications of SSU rRNA amplicon sequencing, but at the cost of poor taxonomic classification due to limited fragment lengths. The Oxford Nanopore Technologies (ONT) platform can sequence full-length SSU rRNA genes, but its lower raw-read accuracy has so-far limited accurate taxonomic classification and de novo feature generation. Here, we present a sequencing workflow, termed ssUMI, that combines unique molecular identifier (UMI)-based error correction with newer (R10.4+) ONT chemistry and sample barcoding to enable high throughput near full-length SSU rRNA (e.g. 16S rRNA) amplicon sequencing. The ssUMI workflow generated near full-length 16S rRNA consensus sequences with 99.99% mean accuracy using a minimum subread coverage of 3x, surpassing the accuracy of Illumina short reads. The consensus sequences generated with ssUMI were used to produce error-free de novo sequence features with no false positives with two microbial community standards. In contrast, Nanopore raw reads produced erroneous de novo sequence features, indicating that UMI-based error correction is currently necessary for high-accuracy microbial profiling with R10.4+ ONT sequencing chemistries. We showcase the cost-competitive scalability of the ssUMI workflow by sequencing 87 time-series wastewater samples and 27 human gut samples, obtaining quantitative ecological insights that were missed by short-read amplicon sequencing. ssUMI, therefore, enables accurate and low-cost full-length 16S rRNA amplicon sequencing on Nanopore, improving accessibility to high-resolution microbiome science.

**Keywords:** microbiome, 16S rRNA, amplicon sequencing, Nanopore, long reads

## Significance Statement

The ability to generate accurate full-length 16S ribosomal RNA (rRNA) gene sequences in a high-throughput manner can advance microbiome science by both improving taxonomic classification of reads as well as propagating public gene databases used for taxonomic classification and oligonucleotide probe design. While recent chemistry changes on the Oxford Nanopore Technologies (ONT) platform have increased its raw-read accuracies, we found that these improvements are not sufficient to overcome erroneous 16S rRNA gene sequence feature generation, thus warranting error correction in this application. To address this, we introduce a workflow for full-length 16S rRNA sequencing on ONT that exceeds the accuracy of other current sequencing platforms, and generates perfect sequence features. The competitive cost and ease of use open doors for high-resolution microbiome science.

## Introduction

The amplification and sequencing of small subunit (SSU) ribosomal RNA (rRNA) genes (e.g. 16S and 18S rRNAs) is a widely used method to study the diversity and taxonomic composition of microbial communities within a variety of environments. The foundational work of Woese and Fox (1) utilized the conserved function of

rRNA across all self-replicating cells to establish the first phylogenetic description of the domains of life, and provided a basis for taxonomically classifying microorganisms based on their evolutionary divergence. Since then, comparative analysis of SSU rRNA gene sequences has enabled the discovery of new uncultivated microbial lineages (2, 3), surveys of microbial community composition in

**Competing Interest :** The authors declare no competing interests.

**Received:** November 21, 2023. **Accepted:** September 5, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of National Academy of Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact [reprints@oup.com](mailto:reprints@oup.com) for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com).

host-associated (4, 5) and natural environments (6–8), and the design of oligonucleotide hybridization probes for environmental monitoring of select taxa (9, 10). Within the past decade, the throughput of SSU rRNA sequence generation has been enhanced by so-called “next-generation” sequencing platforms, such as Roche 454 (8, 11) and Illumina (12) platforms, which are capable of sequencing millions of amplicons generated over hundreds of samples (13, 14).

While next-generation sequencing platforms have provided an affordable and high-throughput approach for generating SSU rRNA gene sequences from multiplexed environmental samples, the taxonomic resolution of amplicon sequences generated from such technologies is limited by their short-read lengths (e.g. up to ~500 bp in paired-end mode (15)). To circumvent this limitation, high-throughput amplicon sequencing of environmental SSU rRNA genes typically relies on the use of conserved primers to amplify discrete variable regions (e.g. V1–V9 regions in 16S rRNA). The selection of the variable region for amplification of 16S rRNA gene fragments can introduce biases into diversity estimates, as the 16S rRNA gene does not evolve evenly along its length (16). The accuracy of taxonomic assignment of 16S rRNA sequences has also been shown to increase with amplicon sequence length, with full-length 16S rRNA sequences required to capture most taxonomic ranks (17, 18). Moreover, the taxonomic characterization of SSU rRNA gene fragments from unknown microorganisms relies on their comparison to reference databases of full-length sequences. The widespread application of short-read sequencing platforms for SSU rRNA gene profiling has resulted in a decreased rate of full-length sequence generation, which is needed for phylogenetic analysis of novel lineages as well as for the development of new oligonucleotide hybridization probes (19–21). Thus, there is a need for new high-throughput approaches capable of sequencing full-length SSU rRNA gene fragments to improve taxonomic classification of microbiomes, as well as to increase the number of full-length SSU rRNA sequences in public databases.

Recently, there have been advancements in single-molecule sequencing technologies capable of generating long reads, such as the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms (15). While these long-read sequencing platforms can alleviate many of the abovementioned problems associated with classifying short reads, their raw sequence data have been limited by high error rates (0.5–2%) compared with second-generation sequencers (< 1%) (22–24). These high error rates of raw long-read platforms can obfuscate SSU amplicon sequence clustering and taxonomic assignments (25, 26), impacting the accuracy of such methods for microbiome profiling. As ONT sequencers (e.g. MinION) have a relatively low capital cost (27) and can be used in field settings (28–31), developing high-throughput and accurate SSU amplicon sequencing with the ONT platform could help to advance microbiome science in diverse applications worldwide.

To circumvent higher error rates in raw long reads, previous strategies have utilized various forms of consensus sequencing approaches for error correction, which involve redundant sequencing of multiple copies of the sample DNA template molecule of interest to obtain a consensus sequence with reduced error (22, 32–34). In particular, it was recently shown that highly accurate (> 99.99%) consensus sequences could be generated on the ONT platform using unique molecular identifiers (UMIs) to tag individual DNA molecules prior to amplification and sequencing (22). In this UMI-based sequencing approach, independent reads sharing the same molecular barcodes are grouped together to enable

consensus sequence generation and error correction. However, a relatively high UMI subread coverage (15–25 $\times$ ) was necessary to reach accuracies above 99.99% (22), thus requiring a high per-sample read depth and limiting the throughput of this method for routine microbiome science involving many samples. Since the development of this UMI-based amplicon sequencing approach, new ONT sequencing chemistries and pores have been developed (e.g.  $\geq$ R10.4) with higher raw-read accuracies (35) that could increase sample throughput by improving UMI detection as well as requiring a lower subread coverage for a desired consensus sequence accuracy. It could also be possible that the higher raw-read accuracies of newer ONT chemistries are sufficient for high-throughput SSU sequencing alone, without the need for read error correction.

Here, we explore the application of ONT sequencing to high-throughput and high-accuracy full-length SSU amplicon sequencing for microbiome profiling. We present a UMI-based full-length SSU amplicon sequencing workflow, termed *ssUMI*, that employs accurate quantification of starting template molecules (near full-length 16S rRNA genes) for library preparation and leverages the higher raw-read accuracy of newer ONT chemistries for stringent UMI detection and binning. We validate the *ssUMI* approach using two synthetic microbial community standards and show that it improves amplicon sequence variant (ASV) and species detection compared with quality-filtered (i.e. nonerror-corrected) Nanopore reads. We also demonstrate its high-throughput scalability by sequencing 87 environmental microbiome samples and 27 human gut samples at a competitive per-sample cost, obtaining strain-resolved ecological insights that were not achievable with short-read sequencing. This approach thus facilitates the use of ONT long-read amplicon sequencing in large-scale microbiome studies.

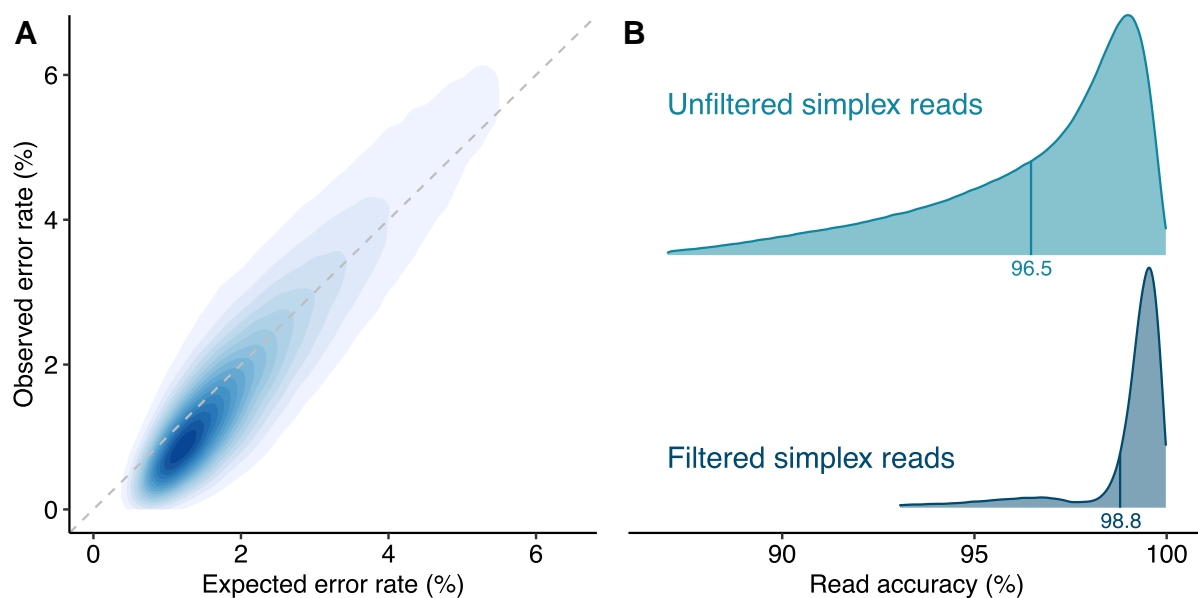
## Results

### Evaluating ONT raw-read accuracy with a mock microbial community standard

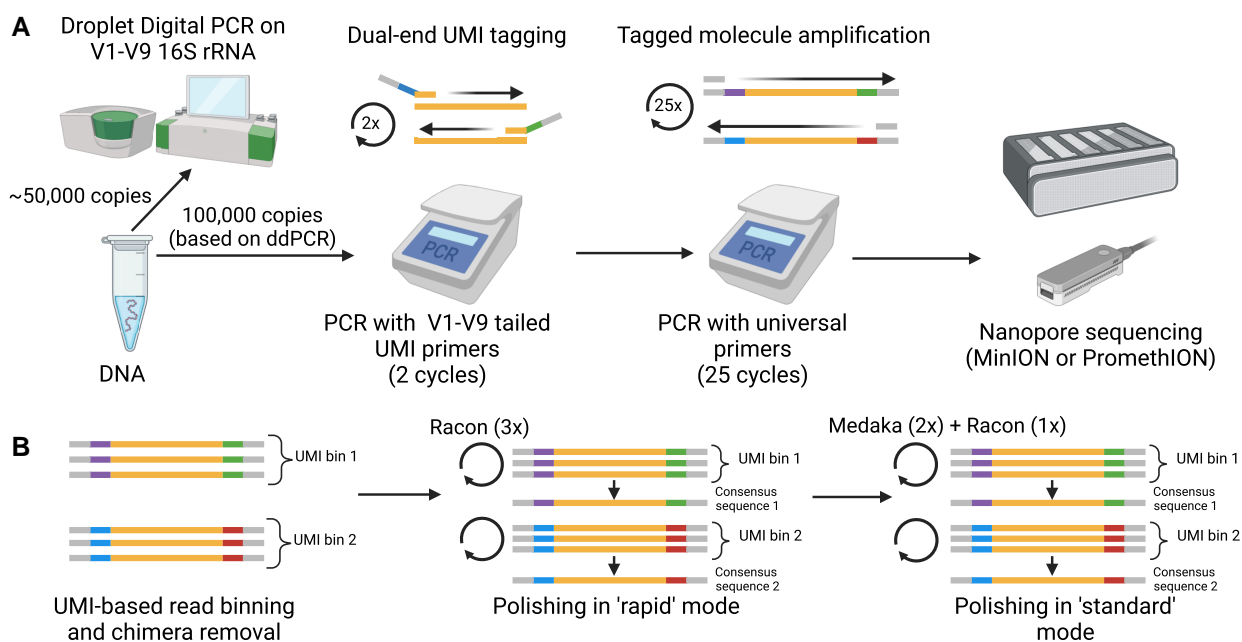
To determine whether raw reads generated with the ONT “Q20+” chemistry (i.e. R10.4+ flowcell) were sufficient for full-length 16S rRNA gene amplicon sequencing analysis, we first assessed the error rate distribution for 4.4 M amplicon reads sequenced from the eight-species ZymoBIOMICS Microbial Community DNA Standard (Fig. 1). Without any quality filtering, the raw reads had a mean accuracy of 96.5% (Fig. 1B), which is insufficient to resolve species or operational taxonomic units (OTUs) at a 97% cluster identity. We found a correspondence between the expected error (EE) rate predicted from the per-base quality (Q) scores and the empirical EE rate of the raw reads (Fig. 1A). We, therefore, implemented an EE-filter threshold of 1%, which filtered 93% of the raw reads and improved the mean read accuracy to 98.8% (Fig. 1B).

### *ssUMI* enables accurate profiling of mock microbial communities

The above analysis of raw reads motivated us to develop a high-throughput long-read sequencing workflow for highly accurate near full-length 16S rRNA genes on the ONT platform. We built upon the dual-UMI-based amplicon sequencing method described by Karst et al. (22), with several key modifications made here to the library preparation and data analysis steps to enable high-throughput 16S rRNA gene sequencing (Fig. 2). Specifically, rather than relying on a trial-and-error approach to determine the number of starting molecules for UMI tagging, we developed a droplet



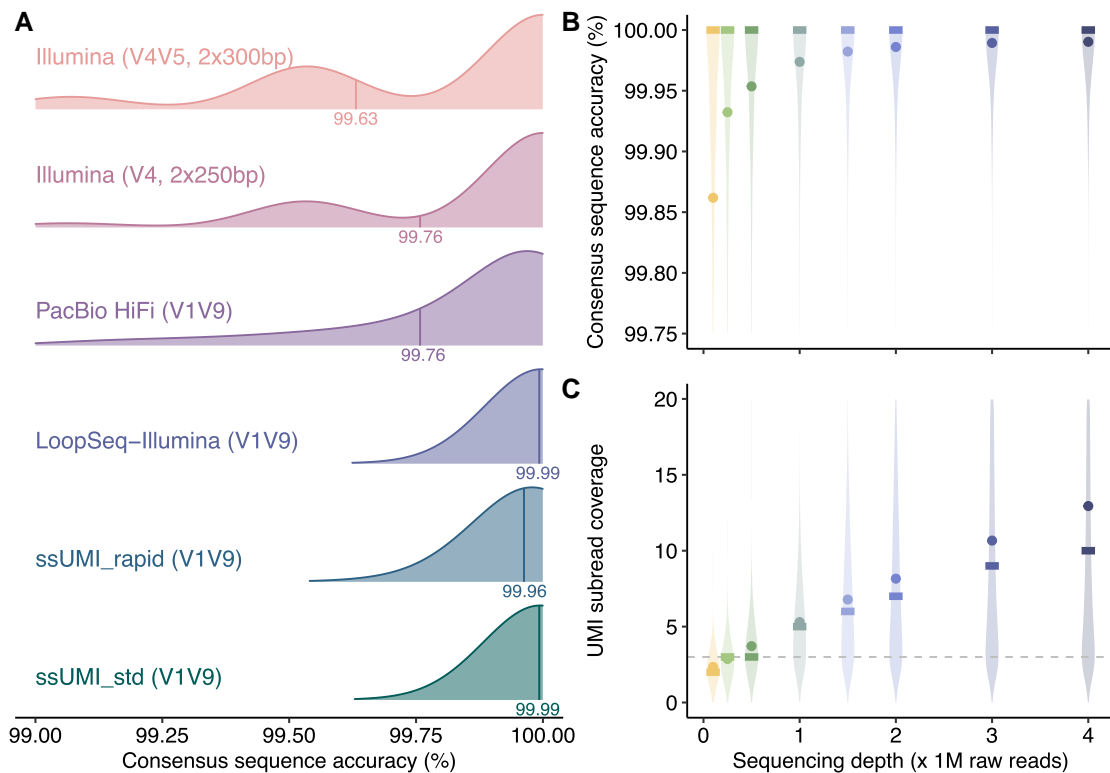
**Fig. 1.** Characterization of raw ONT R.10.4 read quality for 16S rRNA gene amplicons. A) Observed vs. EE rates of length-filtered raw Nanopore reads. The darker shading indicates higher density of reads within that plot region. The dashed gray line represents a 1:1 slope. B) Density plot of read accuracy distribution of unfiltered and length + EE-filtered raw Nanopore reads. Mean accuracy values are indicated with vertical lines and are provided as text below the lines.



**Fig. 2.** Summary of UMI-based SSU rRNA gene sequencing (ssUMI) workflow. A) The wet-laboratory steps, in which DNA templates are first quantified with ddPCR with a near full-length 16S rRNA gene assay. Based on ddPCR quantification, sample DNA containing 100,000 16S rRNA gene copies is added to the first round of ssUMI PCR for UMI tagging. Following two cycles of PCR, the UMI-tagged amplicons are further amplified in a second round of PCR using universal primers that flank the template and UMIs. After PCR amplification, the products are sample-barcoded, pooled, and sequenced on a Nanopore instrument. B) Data analysis workflow following sequencing, in which the reads are analyzed with the ssUMI pipeline for generation of high accuracy near full-length 16S rRNA consensus sequences. Initially, reads are quality filtered, binned based on UMIs from both ends (i.e. UMI pairs), and chimeras are removed. Consensus sequences are polished with Racon (3x) only in the rapid mode of the workflow, or followed by Medaka (2x) and Racon again (1x) in the standard workflow mode.

digital PCR (ddPCR) approach to accurately measure near full-length 16S rRNA gene copies within each sample. Using this approach, we determined an optimum input template molecule number of  $1 \times 10^5$  copies into the UMI-tagging PCR that balances the UMI-tagging efficiency and sequencing throughput (see

Text S1), which is 10-times that used by Karst et al. (22) and is intended to increase detection of rare community members when applied to complex communities. We also developed two data analysis pipelines, termed ssUMI\_rapid (i.e. “rapid mode”) and ssUMI\_std (i.e. “standard mode”), that perform length filtering,



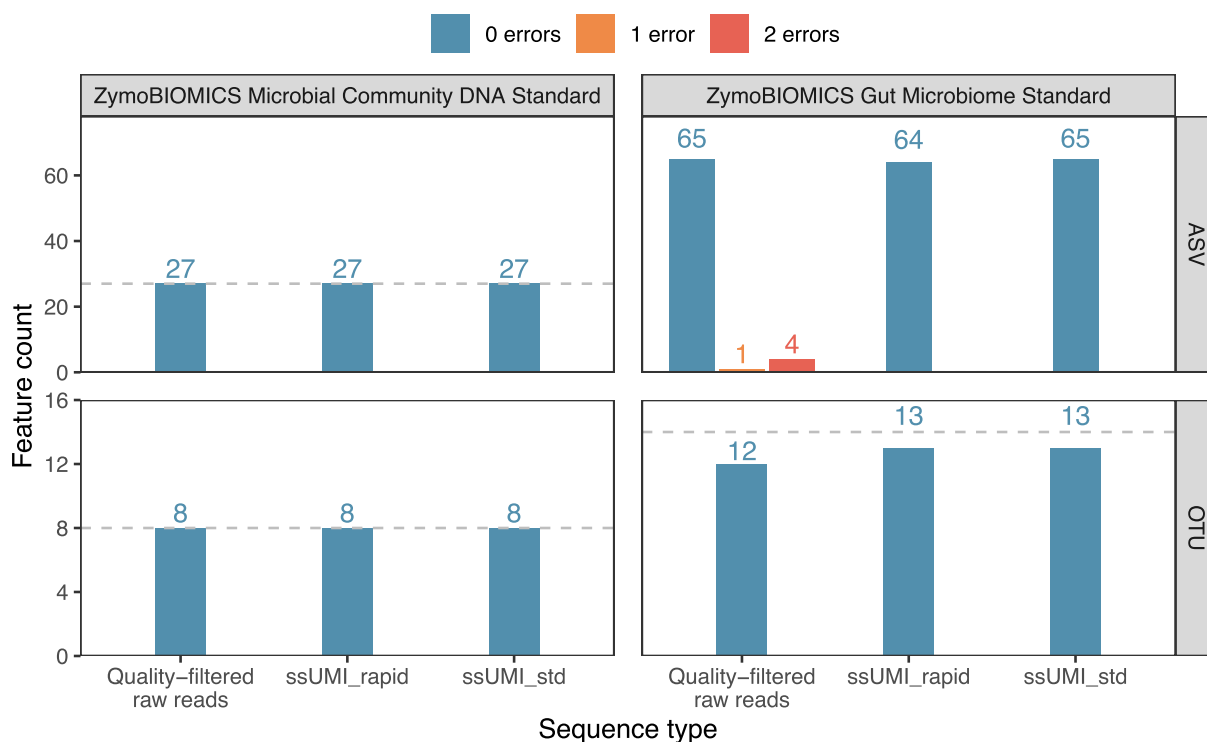
**Fig. 3.** Accuracy and throughput of full-length 16S rRNA gene amplicons with ssUMI workflow. A) Comparison of amplicon sequence accuracies obtained for the ZymoBIOMICS Microbial Community DNA Standard using: Illumina short reads targeting the 16S rRNA gene (V4–V5 region), fully overlapped Illumina short reads ( $2 \times 250$  bp) targeting the 16S rRNA gene (V4 region), PacBio HiFi sequencing targeting the near full-length 16S rRNA gene (V1–V9 regions), LoopSeq (Illumina) synthetic long reads targeting near full-length 16S rRNA gene (V1–V9 regions), as well as UMI-based amplicon sequencing on ONT with ssUMI\_rapid (rapid mode) and ssUMI\_std (standard mode) targeting near full-length 16S rRNA gene (V1–V9 regions). For all sequence data types, amplicons were quality-filtered, primer-trimmed, contaminant sequences were removed, and read counts were normalized to the same depth (18,000 reads) across data types (see the Methods section). Impact of raw-read sequencing depth on distribution of (B) consensus sequence accuracy distribution and (C) UMI subread coverage, for ssUMI\_std applied to full-length 16S rRNA gene amplicon (V1–V9 regions) from the ZymoBIOMICS Microbial Community DNA Standard. Subplots B and C share the x-axis. Different colors represent raw-read sequencing depths, circular points represent mean values and the crossbars represent the median values, while the shaded violin region represents the density distribution of the values at each depth. For subplot (C), the horizontal dashed line represents the minimum UMI subread coverage of 3 $\times$  implemented in this study.

EE filtering, primer trimming, UMI detection, and consensus sequence generation using different modes of consensus read polishing (Fig. 2B). Due to the higher raw-read accuracy of ONT R10.4 chemistry used, it was possible to implement more stringent UMI-based read binning in our analysis workflow by reducing the allowed UMI hamming distance to prevent erroneous read binning. Finally, to improve the throughput of the ssUMI approach and reduce the overall sequencing depth required per sample, we allowed a minimum subread coverage of 3 $\times$  per UMI bin, rather than 15–25 $\times$  used by Karst et al. (22).

For the library generated with the eight-species ZymoBIOMICS Microbial Community DNA Standard,  $8.1 \times 10^4$  UMI-based consensus sequences (with coverage  $\geq 3\times$ ) were generated from a single MinION R10.4 flowcell. Using only three rounds of Racon polishing, hereby termed “rapid mode” in the ssUMI workflow (i.e. ssUMI\_rapid), the mean accuracy of the UMI consensus sequences was 99.96% (Fig. 3A) and 68.1% of sequences were error free. We found a slight increase in consensus sequence accuracy using two rounds of Medaka after the initial Racon polishing (Fig. S1). Applying a final round of Racon polishing after Medaka led to a significant reduction in error rate (Fig. S1), producing a mean sequence accuracy of 99.99% (Fig. 3A). Interestingly, this final round of polishing with Racon after Medaka was more effective than simply applying four sequential rounds of Racon without Medaka (Fig. S2). We term the three-step polishing procedure

using Racon and Medaka “standard mode” in the ssUMI workflow (i.e. ssUMI\_std; Fig. 3A). The greater consensus accuracy achieved by the standard mode was associated with increased computational requirements compared with rapid mode (Table S1). Notably, the mean accuracies of the UMI-based consensus sequences in both rapid and standard modes were higher than that of quality-filtered PacBio HiFi sequences and Illumina short-read sequences from the same microbial community standard (Fig. 3A). Kozich et al. (12) reported a lower error rate ( $\sim 0.06\%$ ) than we observed here ( $\sim 0.26\%$ ) for fully overlapping Illumina  $2 \times 250$  bp paired-end amplicons of the V4 region. Yet, that error rate is still higher than we observed with ssUMI consensus sequences of the full-length 16S rRNA gene. The UMI-based consensus sequences from standard mode had a similar mean accuracy to quality-filtered synthetic long reads generated with LoopSeq, with both strategies able to generate 92.5 and 94.6% (36) error-free sequences, respectively (Fig. 3A).

Because the sequencing depth of the ZymoBIOMICS Microbial Community library was much greater (i.e. one sample per MinION flowcell) than would be typical in a high-throughput application with many samples, we explored the effect of per-sample raw-read depth on UMI consensus sequence generation and accuracy. By randomly subsampling the original sequence library and generating UMI-based consensus sequences, we found a saturation-like behavior in the number of generated UMI



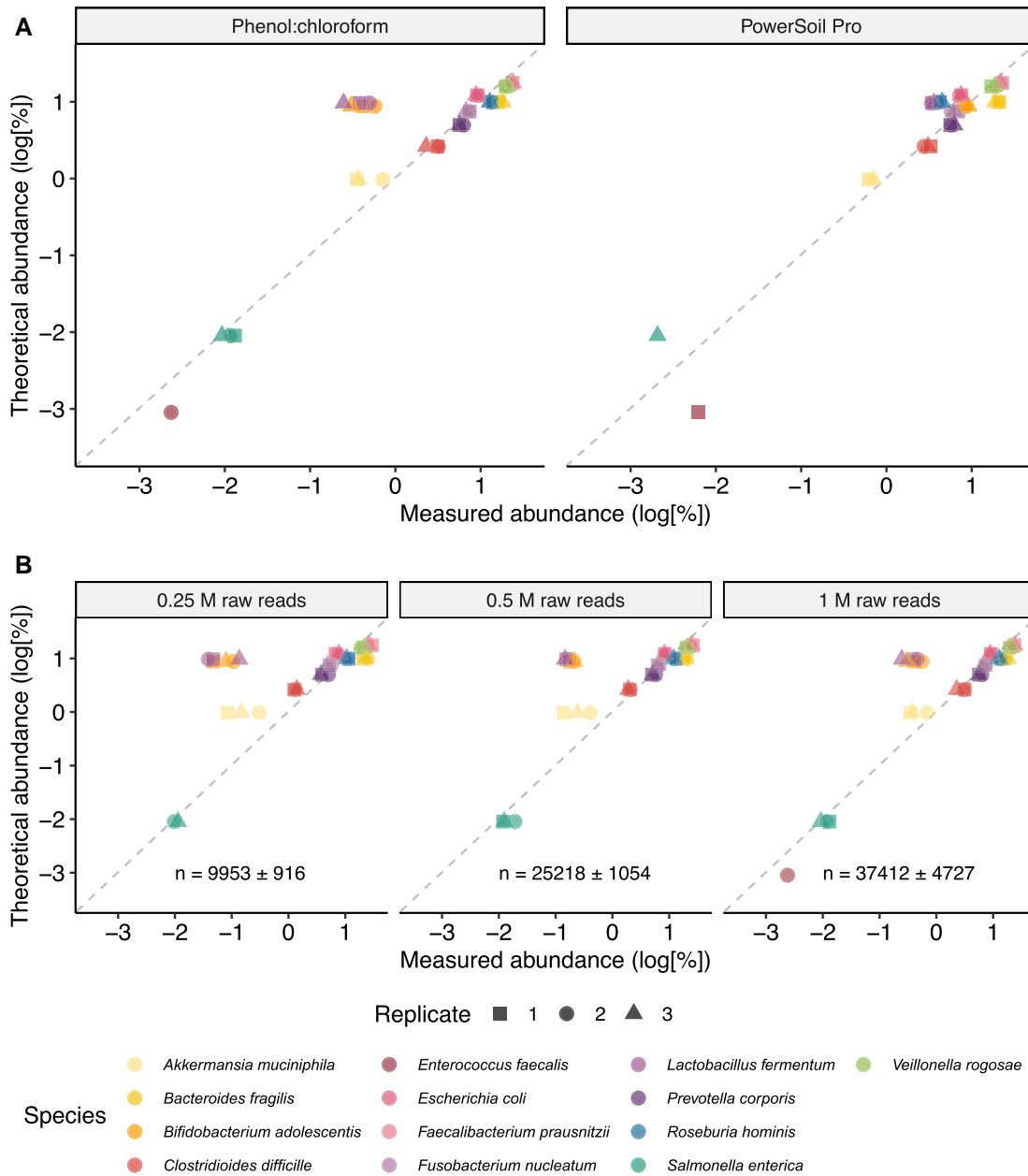
**Fig. 4.** Accuracy of de novo features generated with ssUMI. The number of de novo sequence features generated for ASVs and OTUs at a 97% identity threshold. Sequence features were generated for the 8 bacterial species ZymoBIOMICS Microbial Community DNA Standard and the 14 bacterial species ZymoBIOMICS Gut Microbiome Standard, using either quality-filtered Nanopore raw reads or Nanopore reads processed with the ssUMI pipeline in rapid (i.e. ssUMI\_rapid) and standard (i.e. ssUMI\_std) modes. The number of sequence errors in features are indicated with the fill colors. The dashed gray line indicates the expected number of bacterial full-length 16S rRNA features in the reference community. The lack of a dashed gray line for ASVs in the ZymoBIOMICS Gut Microbiome Standard is due to uncertainty on the true number (see the Methods section). The results of the ZymoBIOMICS Microbial Community DNA Standard were generated with a single sample on a single R10.4 MinION flowcell, and that of the ZymoBIOMICS Gut Microbiome Standard were generated with combined sequences from six technical replicates of two different DNA extractions (see the Methods section) on two R10.4 MinION flowcells. For a given sample type, the same number of reads were used as input for ASV generation with quality-filtered Nanopore reads (e.g. no error correction) and ssUMI workflows.

consensus sequences as a function of sample raw-read depth (Fig. S3). As more UMI-based consensus sequences are recovered with greater raw-read depth (up to the saturation level), users can modify per-sample throughput based on their application and desired sensitivity for detecting rare members. We also observed that greater per-sample sequencing depth increased the UMI-based consensus sequence accuracy, which reached a mean of 99.99% above a per-sample throughput of 2 M reads (Fig. 3B). This trend is largely attributed to higher consensus accuracies achieved at greater UMI-based subread coverage (22), as we also observed that UMI subread coverages increased with raw-read depth (Fig. 3C). Notably, the median UMI-based consensus sequence accuracy remained at 100% and the fraction of error-free reads remained above 50% down to a per-sample raw-read depth of 0.1 M (Fig. 3C). At a per-sample raw-read sequencing depth of 0.25 M reads, the median UMI subread coverage was equal to the minimum threshold of 3x (Fig. 3C). These results indicate that per-sample raw-read depth can be reduced (e.g. for sample multiplexing) while still preserving adequate UMI detection and coverage to generate highly accurate consensus sequences.

We then assessed the accuracy of ASVs and OTUs (at 97% identity) generated for the ZymoBIOMICS Microbial Community DNA Standard using both quality-filtered Nanopore raw reads and UMI-based consensus sequences. Both rapid and standard modes of UMI-based consensus sequences generated 100% accurate ASVs and OTUs that perfectly matched all expected 27 ASVs and 8 OTUs in the reference community (Fig. 4). Surprisingly to

us, quality-filtered Nanopore reads were also capable of generating all ASVs and OTUs perfectly from the mock community, without any false positives (Fig. 4).

To better assess the reproducibility and capability of our ssUMI approach applied to a complex microbiome sample, we sequenced two DNA extracts in (technical) triplicate from the log-distributed ZymoBIOMICS Gut Microbiome Standard cell mixture containing 14 bacterial species. One DNA extract was obtained using the Qiagen MagAttract PowerSoil Pro kit and the other with a phenol:chloroform-based extraction. Each extraction set was sequenced in triplicate on one R10.4 MinION flowcell, yielding  $8.4 \pm 1.7 \times 10^5$  raw reads and  $3.8 \pm 0.6 \times 10^4$  UMI-based consensus sequences per sequencing replicate. The measured relative abundance values in the UMI-based consensus sequences were consistent among technical replicates, and generally matched well with the theoretical abundance for community members that were more abundant than 0.01% (Fig. 5A). An exception was *Lactobacillus fermentum* and *Bifidobacterium adolescentis* in the phenol:chloroform extraction, which showed a clear abundance skew that was consistent within all technical replicates (Fig. 5A). We attribute this aberration to DNA extraction bias, rather than an artifact of the ssUMI pipeline, as this abundance skew was not observed in the replicates extracted with MagAttract PowerSoil Pro (Fig. 5A). Reads from rare community members that were <0.01% abundance were identified sporadically with UMI-based consensus sequences. *Salmonella enterica* (0.009% theoretical abundance) was detected in all replicates of the phenol:



**Fig. 5.** Verifying accuracy and reproducibility of microbial abundance profiles obtained with ssUMI. A) Composition of ZymoBIOMICS Gut Microbiome Standard based on near full-length 16S rRNA (V1–V9 region) consensus sequences processed with the ssUMI\_std pipeline (standard mode), in comparison to the theoretical abundances provided by the vendor. Technical PCR and sequencing replicates are shown for two different DNA extractions of the same cell mixture, phenol:chloroform and MagAttract PowerSoil Pro. B) The impact of sample raw-read depth on the resulting microbial community composition of the ZymoBIOMICS Gut Microbiome Standard (phenol:chloroform DNA extraction) obtained with 16S rRNA consensus sequences processed with the ssUMI\_std pipeline (standard mode). To perform this analysis, raw reads were randomly subsampled from the original sequence libraries to given depths, and UMI-based consensus sequences were generated with the ssUMI pipeline (see the Methods section). The text values shown in the subplots represent the numbers of UMI-based consensus sequences generated (mean  $\pm$  SD of triplicates).

chloroform extraction, but was only detected in the single replicate of the MagAttract PowerSoil Pro extract that had the most raw reads (Fig. 5A). Similarly, *Enterococcus faecalis* (0.0009% theoretical abundance) was only identified in a single technical replicate from both DNA extracts (Fig. 5A).

Increasing sample throughput (i.e. more samples multiplexed in a single run) requires a reduction in the per-sample raw-read count. We therefore investigated the effect of sample read depth on the relative abundance distribution obtained from ssUMI consensus sequences by randomly subsampling the original sequence

libraries from the ZymoBIOMICS Gut Microbiome Standard DNA extracted with phenol:chloroform. Reducing the raw-read depth from 1 to 0.25 M did not greatly impact the observed distribution of taxa in the UMI-based consensus sequences, with the exception that detection of rare species was reduced at lower raw-read depths (Fig. 5B).

For the more complex ZymoBIOMICS Gut Microbiome Standard, we recovered 65, 64, and 65 perfect ASVs from quality-filtered Nanopore raw reads, ssUMI\_rapid consensus sequences, and ssUMI\_std consensus sequences, respectively, by pooling

sequences from all six extraction replicates (Fig. 4). No errors were observed in any ASVs generated from UMI-based consensus sequences, regardless of the data analysis mode (e.g. rapid or standard), while five erroneous ASVs containing one or more errors were generated with quality-filtered Nanopore raw reads (Fig. 4; Tables S2–S4). No sequence type was able to recover ASVs corresponding to *E. faecalis* (0.0009% theoretical abundance; Tables S2–S4). Only ssUMI\_std was able to recover an ASV from *S. enterica* (0.009% theoretical abundance), while this organism was missed with ssUMI\_rapid and quality-filtered Nanopore raw reads (Tables S2–S4). After clustering into 97% OTUs, we recovered 12, 13, and 13 error-free OTUs from quality-filtered Nanopore raw reads, ssUMI\_rapid consensus sequences, and ssUMI\_std consensus sequences, respectively (Fig. 4; Tables S5–S7). For OTUs generated with both ssUMI\_rapid and ssUMI\_std sequences, all bacterial species in the ZymoBIOMICS Gut Microbiome Standard were detected, except for *Clostridium perfringens* (0.0002% theoretical abundance; Tables S5 and S6). For OTUs generated with quality-filtered Nanopore raw reads, 12 bacterial species at or above 0.01% were detected (Table S7). The ssUMI workflow therefore improved de novo feature (e.g. ASV/OTU) detection and accuracy relative to uncorrected Nanopore raw reads.

### Application ssUMI for high-throughput microbiome profiling of human gut and environmental samples

We further demonstrated the scalability of the ssUMI workflow by applying it to 90 wastewater samples collected bi-weekly from a nearby wastewater treatment facility. A total of seven wastewater sample matrices were collected and prepared with the ssUMI workflow, and the products were sequenced on the ONT PromethION platform, generating a total of 103.8 Gb raw-read data (Table S8). On average, each sample yielded  $4.3 \pm 1.6 \times 10^5$  raw reads, with exception of three samples that were not sufficiently barcoded (Fig. S4). The number of UMI-based consensus sequences increased with sequencing depth (Pearson's  $r = 0.73$ ,  $P = 8.5e-16$ ,  $n = 87$ ) for all wastewater sample types (Fig. S5), which is in accordance with the abovementioned observations obtained by in silico subsampling of mock community libraries. Wastewater samples generated significantly more UMI-based consensus sequences than the mock community at the same sequencing depth (Student's  $t$  test,  $P < 0.05$  for all read depths), indicating that the UMI tagging and PCR were not inhibited by the complex wastewater matrices. No significant difference (ANOVA,  $P = 0.135$ ,  $F = 1.792$ ,  $df = 6$ ) was observed for the number of UMI-based consensus sequences generated in standard mode from different sample types, with an average of  $3.6 \pm 1.1 \times 10^4$  consensus sequences per sample (Fig. 6A).

We then generated near full-length 16S rRNA gene ASVs with the UMI-based consensus sequences from all wastewater samples, yielding a total of 8,349 bacterial ASVs from all sequenced samples. Rarefaction analysis showed this average ssUMI sequence depth was capable of capturing the majority of ASV richness in the wastewater communities (Fig. S6). Based on a principal co-ordinate analysis (PCoA) of Bray–Curtis dissimilarity, there was a clear impact of wastewater sample type on the bacterial community structure (Fig. 6B). Influent, trickling filter, mixed sludge, and anaerobic digester samples were each tightly clustered in the PCoA, while activated sludge, waste activated sludge, and effluent samples formed a relatively loose cluster (Fig. 6B). The lowest ASV richness (mean  $1,165 \pm 271$ ) was observed in anaerobic digester samples, while the highest ASV richness

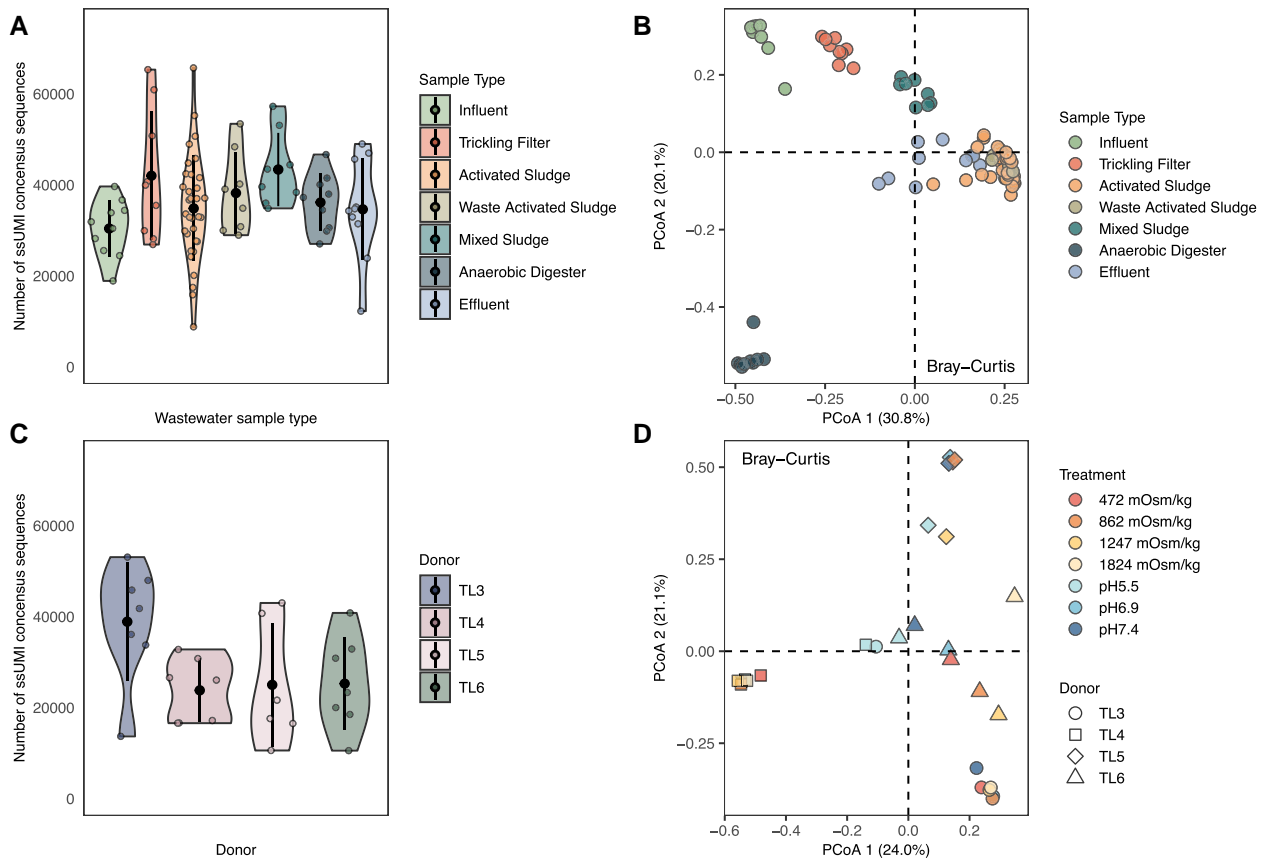
(mean  $3,260 \pm 524$ ) was observed in mixed sludge samples (Fig. S7). Based on ASV abundance profiles, the bacterial community structures of each sample type were relatively stable over the study period (Figs. 6B and S8).

We also explored the applicability of the ssUMI approach for profiling the human gut microbiota by comparing microbial dynamics between the ssUMI workflow and Illumina short-read 16S rRNA sequencing. Human stool samples were collected from four donors and cultured under increasing osmolality and pH levels to replicate environmental stressors (37). Across all donors and conditions, a total of 28 samples were characterized both by 16S rRNA amplicon sequencing targeting the V4–V5 regions with Illumina and the V1–V9 regions with the ssUMI approach, with 27 samples yielding sufficient data on both sequencing platforms (Table S9). Similar to wastewater samples, no significant differences were observed for the number of UMI-based consensus sequences obtained from different donors (ANOVA,  $P = 0.06$ ,  $F = 2.847$ ,  $df = 3$ ), with an average of  $2.8 \pm 1.2 \times 10^4$  consensus sequences per sample generated from  $7.1 \pm 2.4 \times 10^5$  raw reads per sample (Figs. 6C and S4). After rarifying pooled ssUMI sequences and quality-filtered Illumina reads across these 27 samples to  $6.61 \times 10^5$  reads, we generated 867 and 314 ASVs from both datasets, respectively. Rarefaction curves showed this ssUMI sequencing depth was sufficient to capture the majority of ASV richness across the gut samples (Fig. S6). Based on a PCoA of Bray–Curtis dissimilarity of ssUMI ASV abundances, the gut microbiome samples clustered more strongly based on the donor rather than the osmolality and pH treatments, except for culturing under pH 5.5 (Fig. 6D).

Similar relative abundance profiles of prevalent bacterial families were obtained using both Illumina sequencing and the ssUMI pipeline across most treatment conditions (Fig. S9). However, different profiles were observed using absolute abundances inferred from the ssUMI workflow, which were based on combining ddPCR quantification with ssUMI compositional data (see the Methods section). For example, the absolute abundance of the highly prevalent family *Enterococcaceae* was relatively unimpacted by increasing osmolality up to a concentration of 1,824 mOsm/kg (Figs. 7C and S8C). This was in contrast with its relative abundance obtained with both Illumina and ssUMI, which showed *Enterococcaceae* increasing in proportion at higher osmolality concentrations (Figs. 7A, 6B, and S9A, B), as was previously reported (37).

To deconvolute a possible correlation between *Enterococcaceae* abundance and osmolality, we further investigated population dynamics within this family. Only two *Enterococcaceae* ASVs (with  $>0.5\%$  abundance) were recovered across all samples with Illumina V4–V5 16S rRNA amplicons. When sequencing V1–V9 16S rRNA amplicons with the ssUMI workflow, 12 *Enterococcaceae* ASVs ( $>0.5\%$  abundance) were recovered (Fig. 7). Of those, 11 ASVs shared 100% identity with the 2 V4–V5 Illumina *Enterococcaceae* ASVs, and 1 had a single nucleotide difference (Table S10). The two Illumina V4–V5 *Enterococcaceae* ASVs both shared 100% identity with 6–8 species in the NCBI database, while four ssUMI *Enterococcaceae* ASVs could be uniquely mapped to a single species (Table S11). ssUMI, therefore, improved the taxonomic resolution obtained from 16S rRNA amplicons in comparison with short-read sequencing by enabling more unique species-level identifications (4/11 vs. 0/2 for ssUMI and Illumina, respectively).

With both ssUMI and short-read sequencing, it was apparent that different *Enterococcaceae* ASVs dominated different human donors (Fig. 7). Specifically, the Illumina *Enterococcaceae* ASV “Illumina\_ASV\_2” dominated samples from donor TL4, whereas



**Fig. 6.** Application of ssUMI to high-throughput profiling of wastewater and human gut samples. A) Number of UMI-based consensus sequences generated with ssUMI\_std mode for samples representing wastewater matrix types. Each colored point represents one sequenced sample, the black dots represent the mean number of consensus sequences and the bars show SD. The shaded region represents the density distribution of all samples within that matrix type. B) PCoA of ASV Bray-Curtis dissimilarity, showing the clustering of the different wastewater sample types collected over 2 months. Each colored point within the PCoA represents one sequenced sample. C) Number of UMI-based consensus sequences generated with ssUMI\_std mode for human stool samples inoculated with different donors. Each colored point represents one sequenced sample, the black dots represent the mean number of consensus sequences and the bars show SD. The shaded region represents the density distribution of all samples inoculated with that donor. D) PCoA of ASV Bray-Curtis dissimilarity, showing the clustering of the different human stool inocula and treatments. Each colored point within the PCoA represents one sequenced sample.

the other Illumina *Enterococcaceae* ASV “Illumina\_ASV\_1” dominated samples from the remaining donors (Fig. 7A). Such donor partitioning was also observed with ssUMI ASVs, yet more strain-level diversity was observed with the long-read approach (Fig. 7B). Samples from donor TL4 were dominated by three ssUMI V1–V9 ASVs that were not observed in the other donors, rather than the abovementioned single Illumina ASV. Interestingly, the remaining *Enterococcaceae* ssUMI ASVs varied among donors TL3, TL5, and TL6, while the single Illumina ASV Illumina\_ASV\_1 dominated those donors. These observations collectively indicate that accurate long-read amplicon sequencing with ssUMI can provide a finer resolution of intrahost population dynamics relative to short-read amplicon sequencing, likely by resolving intragenomic rRNA copies as well as potential strain variation.

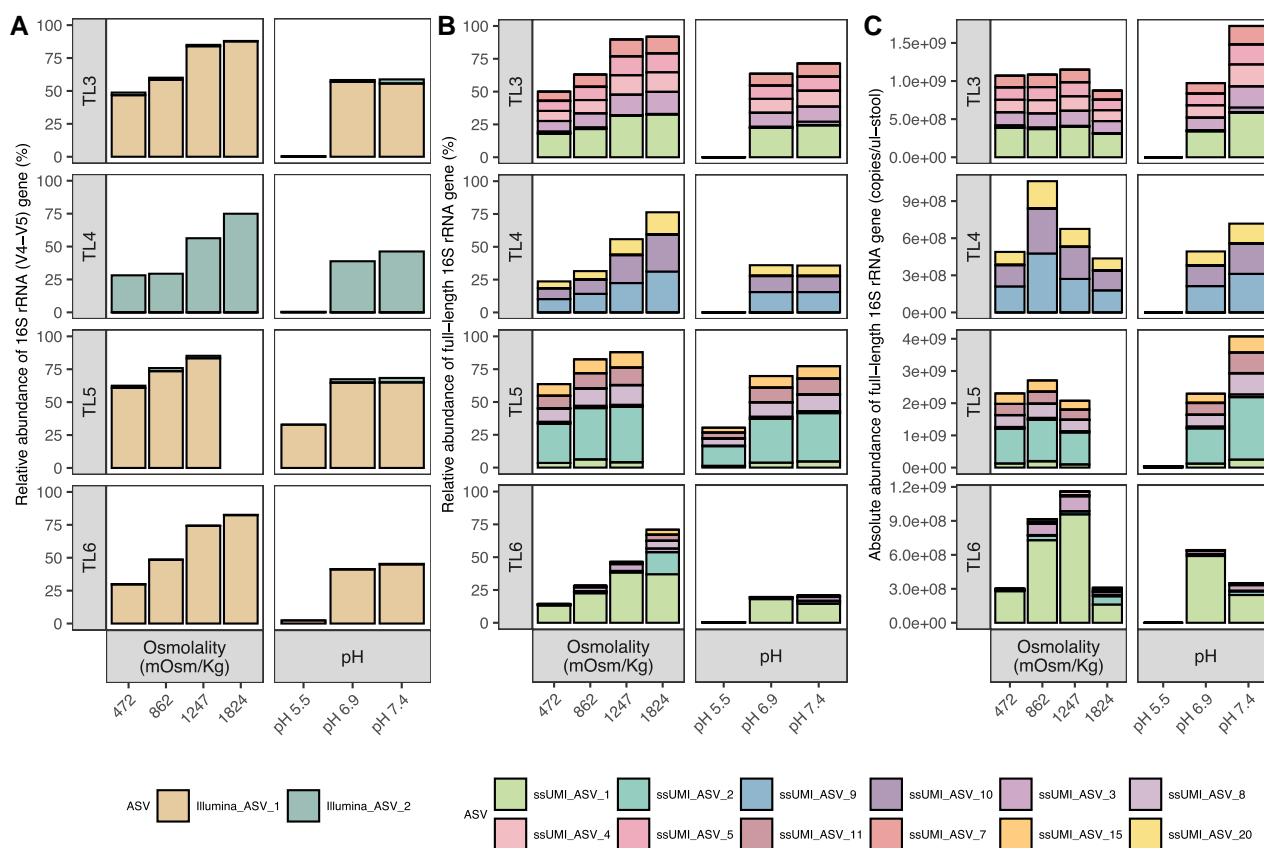
## Discussion

Numerous studies have shown the benefits of full-length 16S rRNA amplicon sequencing for taxonomic classification compared with short-read amplicon sequencing (18, 19, 38–41). However, previous approaches for high-throughput full-length 16S rRNA gene amplicon sequencing with ONT typically relied solely on alignment to reference databases (29, 42), rather than de novo generation of sequence features (e.g. ASVs or OTUs).

Accurate full-length 16S rRNA gene sequence features are critical for developing ecosystem-specific databases (20, 43), designing organism-specific primers or probes (10, 44), and for cross-study analyses (45). In this study, we introduced the ssUMI workflow for high-throughput UMI-based error correction of near full-length 16S rRNA amplicons, which provided higher sequence accuracy than was observed for PacBio HiFi long reads and Illumina short reads, and reached a similar accuracy to LoopSeq synthetic long reads sequenced on Illumina. Using two mock microbial community standards of varying complexity, we found that the de novo sequence features (ASVs and 97% OTUs) generated with ssUMI were error free with no false positives, indicating the approach can be applied for accurate microbiome profiling as well as for SSU sequence database propagation.

Surprisingly, even though the mean accuracy of nonerror-corrected quality-filtered Nanopore raw reads was below 99%, it was still possible to generate perfect ASVs and 97% OTUs for the simpler ZymoBIOMICS Microbial Community DNA Standard. However, when applied to the more complex ZymoBIOMICS Gut Microbiome Standard, the nonerror-corrected Nanopore reads generated false-positive (i.e. erroneous) ASVs, and the generated 97% OTUs had lower sensitivity than ssUMI sequences. The poorer sensitivity of sequence feature detection using Nanopore raw reads is likely caused by the requirement of denoising and





**Fig. 7.** Abundance of *Enterococcaceae* ASVs under different osmolality and pH conditions. A) Relative abundances of *Enterococcaceae* ASVs generated with Illumina 16S rRNA amplicon sequencing targeting V4–V5 region. B) Relative abundances of *Enterococcaceae* ASVs generated with ssUMI\_std mode for near full-length 16S rRNA amplicon sequencing targeting V1–V9 region. C) Absolute abundances of *Enterococcaceae* ASVs obtained by combining ddPCR quantification of total 16S rRNA genes and relative abundances obtained with ssUMI\_std mode (see the Methods section). Only *Enterococcaceae* ASVs with > 0.5% abundance are shown. Note, the sample from 1,824 mOsm/kg in TL5 was not analyzed in this study because it did not yield sufficient data (Table S9).

clustering algorithms to observe a unique sequence multiple times (46), which was limited by the higher average sequence error rate. Sensitivity was also likely impacted by the large fraction (93%) of raw reads falling below the quality filter. Based on these findings, it may be appropriate to use quality-filtered Nanopore raw reads for microbial profiling only if the community is known to be relatively simple with few rare members (e.g. all members >0.1% relative abundance). Future improvements in Nanopore read accuracy, such as via improved base-calling models that convert raw signal into predicted DNA sequence data (47) or by revised nanopore chemistry (48), could soon enable Nanopore reads to be used without read error correction in diverse applications of 16S rRNA gene amplicon sequencing of complex microbiomes.

Based on the above need for error correction of Nanopore reads for 16S rRNA amplicon sequencing, we introduced both a standard and rapid mode of consensus sequence generation in the ssUMI workflow, depending on use cases: standard mode being more appropriate if computational resources are not limiting and/or high sensitivity is required, and rapid mode being more appropriate for analysis where detection of rare species is not a priority and/or when compute resources are limited. The ssUMI workflow in “standard” mode (i.e. ssUMI\_std) achieved the greatest sequence accuracy, and consequently ASVs generated with these sequences had the highest sensitivity for species detection. That is, ASVs produced from ssUMI\_std sequences for the ZymoBIOMICS Gut Microbiome Standard included *S. enterica*

that was present at <0.01% theoretical abundance, whereas ASVs from this species were missed by rapid mode (i.e. ssUMI\_rapid) due to residual errors reducing unique sequence counts. With either approach, the use of UMI tags on both ends of the amplicons enables precise removal of chimeras (22), which are otherwise difficult to detect without molecular identifiers (49). The abundance distributions obtained by UMI-tagged 16S rRNA gene amplicons should also have reduced effects of PCR amplification bias, as UMI-based abundances are based on single-molecule counting (50).

We further explored the capacity of this highly accurate UMI-based sequencing method for high-throughput profiling of microbial communities in 87 complex environmental (e.g. wastewater) samples collected from 7 different sampling locations over 2 months. The abundance profiles of full-length 16S rRNA ASVs generated with the ssUMI workflow were well clustered according to sample type, providing insights into microbial community composition and colonization patterns within the treatment plant. For instance, the “trickling filter” received flow from the treatment plant “influent,” and the bacterial communities in those sample types were closely clustered based on ASV Bray–Curtis dissimilarity. Similarly, “waste activated sludge” samples represent biomass removed from the “activated sludge” bioreactors, and thus, the co-clustering of communities from these sample types was expected. The relatively high similarity between the microbial community in activated sludge and “secondary clarifier effluent” was likely

the result of the physical carryover of microorganisms from the sludge into treatment plant effluent. These findings align with previous efforts using short-read sequencing that observed microbial immigration can impact spatial variation in microbial diversity and community structure across transects of full-scale wastewater treatment plants (51, 52). Overall, these results showcase that ssUMI can be used for multisample experiments across various environmental matrices to gather ecological insights into microbial dynamics.

We also demonstrated that the ssUMI workflow can be applied for quantitative profiling of human gut samples. By applying the ssUMI workflow to 27 stool samples incubated under various pH and osmolality conditions, we obtained a higher taxonomic resolution than was achieved with short-read-based amplicon sequencing along with a finer resolution of population structure. The improved resolution of population ASVs (e.g. from *Enterococcaceae*) observed with ssUMI is likely attributed to the greater probability of sequence divergence across the full-length 16S rRNA gene, rather than within a single subregion, within intragenomic rRNA copies (18, 39, 53). Finally, by using the same full-length 16S rRNA primers for quantification by ddPCR and UMI-based amplicon sequencing, the ssUMI workflow can enable quantitative microbiome profiling using estimates of absolute microbial load, which has been previously shown to capture ecological trends not revealed by relative abundance alone (54, 55). Similarly, we observed contrasting relationships between *Enterococcaceae* abundance with osmolality in the human gut samples when using relative abundance and absolute abundance estimates, highlighting the utility of the ssUMI workflow to combine quantitative microbial load with compositional data.

Successful implementation of the ssUMI workflow depends on several key factors, including sample DNA quality, access to ONT instruments, and available computational resources. We recommend that users extract DNA using protocols that yield high-molecular-weight DNA to obtain a sufficient amount of full-length 16S rRNA genes. For users with limited access to ddPCR and/or quantitative PCR (qPCR), a trial-and-error approach (22) with the ssUMI workflow could be used to determine the optimal input DNA amount. Users could also reduce costs and computational requirements by performing rarefaction analysis to determine an effective minimum sequencing depth required to cover a desired level of species richness, as we found 0.25 M raw reads (producing  $\sim 2.5 \times 10^4$  ssUMI consensus sequences) was sufficient for wastewater and human gut samples in this study. At a target per-sample sequencing depth of 0.5 M raw reads (sufficient to generate  $\sim 4 \times 10^4$  UMI-based consensus sequences), we find that full-length 16S rRNA gene amplicon sequencing with the ssUMI workflow on ONT is cost competitive with other current long-read amplicon sequencing platforms, while remaining more expensive than short-read amplicon sequencing on Illumina (Table S12). It should be noted that this sequencing depth was chosen to maximize ssUMI consensus sequence accuracy and throughput with ONT Q20+ chemistry. However, this cost landscape will inevitably change as long-read platforms continue to progress in terms of accuracy and throughput. Regardless, unlike other existing sequencing platforms, 16S rRNA gene amplicon sequencing on ONT platforms (e.g. PromethION P2 or MinION) can easily be performed in a typical laboratory due to their lower capital costs, thus providing more flexibility and faster data turnaround.

Taken together, this study demonstrates that UMI-based 16S rRNA gene sequencing on the Nanopore platform can be applied in a high-throughput manner to sensitively and accurately measure microbial community structures in complex microbiomes. We

expect the reduced instrumentation and sample processing costs associated with ssUMI will improve accessibility to high-resolution microbiome science and to increase the generation of highly accurate near full-length SSU sequences that can expand our public databases.

## Methods

### Sources of DNA

*Escherichia coli*, Strain 83972 (BEI Resources, Manassas, VA, USA) and two mock community products (Zymo Research, Irvine, CA, USA) were used for validation of the ssUMI pipeline. DNA from *E. coli* was extracted with the MagAttract HMW DNA kit (Qiagen, Hilden, Germany) following the manufacturer's protocol for DNA extraction from gram-negative bacteria. The ZymoBIOMICS Microbial Community DNA Standard (cat. no.: D6306, lot no: 213089; Zymo Research) is a DNA standard consisting of eight evenly distributed bacteria (three gram-negative and five gram-positive) species. The ZymoBIOMICS Gut Microbiome Standard (cat. no: D6331, lot no: ZRC194753; Zymo Research) is a cell standard comprising 18 bacterial strains (14 species), 2 fungal strains, and 1 archaeal strain mixed at log-distributed cell concentrations. DNA was extracted from 125  $\mu$ L of fully resuspended ZymoBIOMICS Gut Microbiome Standard cell mixture with a phenol:chloroform extraction protocol (56, 57) and the MagAttract PowerSoil Pro DNA kit (Qiagen), following the published protocol or manufacturer's instructions, except for the following modifications: (i) MetaPolyzyme treatment (58) was used for cell lysis for the phenol:chloroform extraction; and (ii) the volume of MagAttract Suspension G beads and Buffer QSB1 were doubled for the MagAttract PowerSoil Pro DNA kit.

Wastewater samples were collected from a wastewater treatment facility in the Vancouver area (British Columbia, Canada) from June to August 2022, and included: primary clarifier effluent (referred to herein as influent), trickling filter effluent, activated sludge mixed liquor, waste activated sludge, mixed primary and secondary sludge ("mixed sludge"), anaerobic digester sludge, and secondary clarifier effluent ("effluent"). Wastewater samples were shipped to the University of British Columbia (British Columbia, Canada) on ice, aliquoted, and concentrated via flocculation (influent, trickling filter, activated sludge, waste activated sludge, and effluent) or centrifugation (mixed sludge and anaerobic digester sludge) within 24 h (see Text S2), and stored at  $-20$  °C before extraction. The preserved wastewater samples were thawed at 4 °C, and DNA was extracted with the MagAttract PowerSoil Pro DNA kit (Qiagen) using an Opentrons-2 (Opentrons Labworks, Queens, NY, USA) automated liquid handler (see Text S2). Extracted DNA samples were quantified with Qubit dsDNA HS Assay Kit using a Qubit 4 fluorometer (Invitrogen, Waltham, MA, USA).

Human fecal samples were collected from six healthy individuals (TL1, TL2, TL3, TL4, TL5, and TL6), fermented anaerobically for 24 h, then cultured for 48 h in Mega Medium under various pH and osmolality conditions, as described by Ng et al. (37). Following incubation, DNA from the cultured samples was extracted with the DNeasy PowerSoil Pro kit (Qiagen), following the manufacturer's instructions. DNA samples from donors TL3, TL4, TL5, and TL6 that were cultured under pH 5.5, 6.9, or 7.4 and osmolality of 472, 862, 1,247, or 1,824 mOsm/kg were selected for 16S rRNA gene amplicon sequencing on both Illumina and Nanopore platforms. Human samples were collected in accordance with the University of British Columbia Office of Research Ethics (protocol H21-02464), and samples were de-identified prior to use in this study.

## Molecule tagging and PCR amplification for ssUMI

DNA samples were first quantified with a ddPCR full-length 16S rRNA assay to infer the proper input amount for the UMI-tagging reaction in ssUMI (see [Text S1](#); [Supplementary Figures S10–S12](#)). In addition to developing a near full-length 16S rRNA gene assay with ddPCR, we also found that qPCR targeting the V4–V5 region of the 16S rRNA gene worked reasonably well to determine the required DNA input for ssUMI (see [Text S3](#); [Supplementary Figure S13](#)). The ssUMI PCR consisted of two reactions, UMI tagging and PCR amplification. UMI tagging was conducted using sample DNA containing 100,000 16S rRNA gene copies per reaction (informed by on ddPCR quantification of the sample), using a modified PCR program and conditions from ONT (Custom PCR UMI protocol) (59). In brief, 16S rRNA genes were dual-tagged with UMIs in two cycles of PCR (ssUMI-PCR1), then amplified with two additional PCR runs (ssUMI-EarlyPCR2 and ssUMI-LatePCR2) consisting of 10 and 15 cycles, respectively. Each ssUMI-PCR1 reaction contained 5  $\mu$ L diluted DNA template (100,000 16S rRNA copies by ddPCR), 500 nM UMI-containing forward (8F) and reverse (1391R) primers (Integrated DNA Technologies, Coralville, IA, USA) targeting the 16S rRNA gene (Table S13), and 25  $\mu$ L 2 $\times$  Platinum SuperFi II Green PCR Master Mix (Thermo Fisher Scientific, Waltham, MA, USA) in a 50  $\mu$ L total volume. The two secondary rounds of PCR (i.e. ssUMI-EarlyPCR2 and ssUMI-LatePCR2) were comprised of 18  $\mu$ L cleaned PCR products from the previous step, 100 nM forward and reverse universal primers, 1 mM MgCl<sub>2</sub>, and 25  $\mu$ L 2 $\times$  Platinum SuperFi II Green PCR Master Mix in 50  $\mu$ L total reaction volumes. The ssUMI PCR thermocycling conditions were optimized for the full-length 16S rRNA gene to reduce nonspecific amplification and chimeras by keeping a low PCR cycle number and using a longer extension time (Table S14). All primer sequences and detailed thermocycling conditions for each PCR are summarized in Tables S13 and S14. After each PCR step, PCR products were cleaned with Mag-Bind TotalPure NGS beads (0.6 $\times$  beads/sample ratio; Omega Bio-tek, Norcross, GA, USA) following the manufacturer's instructions. An online interactive protocol for the ssUMI workflow is available at: [dx.doi.org/10.17504/protocols.io.6qpvr3qkpvmk/v1](https://dx.doi.org/10.17504/protocols.io.6qpvr3qkpvmk/v1).

## Sequencing library preparation and sequencing

Nanopore sequencing libraries of UMI-tagged full-length 16S rRNA amplicons from the ZymoBIOMICS Microbial Community DNA Standard, ZymoBIOMICS Gut Microbiome Standard, wastewater, and human gut samples were prepared using the ONT Ligation Sequencing Kit 12 and Native Barcoding Kit (SQK-LSK112, SQK-LSK112.24, and SQK-LSK112.96, respectively) following the manufacturer's instructions, and sequenced in three different runs for 72 h: (i) ZymoBIOMICS Microbial Community DNA Standard was sequenced on a MinION R10.4 flowcell; (ii) the three technical replicates of ZymoBIOMICS Gut Microbiome Standard extracted with phenol:chloroform extraction and MagAttract PowerSoil Pro were barcoded and sequenced on two MinION R10.4 flowcells; (iii) wastewater samples (90 in total), human gut samples (28 in total) and a no-template control were barcoded and sequenced on three PromethION R10.4 flowcells.

The ZymoBIOMICS Microbial Community DNA Standard was also prepared for Illumina sequencing of V4–V5 regions of the 16S rRNA gene (nonUMI tagged) with primers 515F-Y 5'-GTGYC AGCMGCCGCGGTAA-3' and 926R 5'-CCGYCAATYMTTTRAGTT T-3' (60), following the Earth Microbiome Project 16S Illumina Amplicon Protocol (13) using an Illumina MiSeq in 2 $\times$ 300 paired-end mode.

## Nanopore read base calling and processing

Nanopore sequencing raw data were first base called with guppy v6.3.8 (<https://nanoporetech.com/community>) using the super high-accuracy model (dna\_r10.4\_e8.1\_sup.cfg), and then demultiplexed using guppy v6.3.8 with default settings.

For assessment of Nanopore raw reads (i.e. without UMI-based error correction) for microbial profiling of the ZymoBIOMICS Microbial Community DNA Standard and ZymoBIOMICS Gut Microbiome Standard, raw reads were length filtered (1,200–2,000 bp) and quality filtered based on a maximum EE rate of 1% with VSEARCH v11 (61) using the “fastq\_filter” command (61). Nanopore sequencing adapters and 16S rRNA primers (8F/1391R) were removed from the unfiltered and quality-filtered raw reads with Porechop v0.2.4 (<https://github.com/rrwick/Porechop>) and cutadapt v2.7 (62), as previously described (22). Sequences not containing both primers were discarded.

Quality-filtered raw reads were de-replicated using USEARCH v11.0 (63) with the “-fastx\_uniques” command and a minimum number of sequence observations of 2. ASVs were generated with the de-replicated sequences using the UNOISE3 algorithm (64) in USEARCH v11.0, with a minimum unique size of 10. OTUs clustered at 97% identity were generated from size-sorted sequences de-replicated with the “-cluster\_otus” command in USEARCH v11.0.

## ssUMI data processing pipeline

Raw reads from each sample were analyzed with the ssUMI data analysis pipeline for generation of high-accuracy consensus sequences. The ssUMI pipeline was derived from the longread\_umi package developed by Karst et al. (22), which identifies UMI sequences in raw reads, bins raw reads by shared UMI pairs (e.g. UMIs from both ends), and generates consensus sequences for each UMI bin. We made several key modifications to adapt the workflow for the newer ONT sequence chemistry applied to 16S rRNA gene amplicon sequencing. Specifically, a new EE-rate-based quality filtering was applied to the raw reads using VSEARCH v11 with the fastq\_filter command and an EE threshold of 10%. Length filtering for our near full-length 16S rRNA target was modified to 1,200 to 2,000 bp. Due to the lower raw-read error rates for Nanopore reads generated with the R10.4 chemistry, the allowed error rate in a 36-bp UMI pair (e.g. 18 bp UMIs from each end) was reduced from 6 to 4 bp, the maximum mean errors per UMI pair in a bin was set to 2, and the minimum allowed UMI cluster size was reduced to 3. We also implemented two modes of consensus sequence generation, termed “ssUMI\_rapid” for fast analysis and “ssUMI\_std” for high-accuracy analysis. In ssUMI\_rapid mode, sequences were polished with three rounds of Racon v. 1.4.10 (65); while in ssUMI\_std mode, sequences were polished with a three-step method, consisting of three rounds of Racon, followed with two rounds of Medaka v.1.7.2, and a final round of Racon. The “r104\_e81\_sup\_g610” model was used for polishing with Medaka. Minimap2 v2.17 (66) was used for mapping raw reads during polishing steps, with the “-ax map-ont” flag. Scripts associated with the ssUMI data processing pipeline are available at: <https://github.com/ZielsLab/ssUMI>.

ASVs and 97% OTUs were generated with UMI-based consensus sequences following the same procedure described above for quality-filtered raw Nanopore reads.

To investigate the effects of sequencing depth, we randomly subsampled raw reads from mock community samples using seqtk v1.3 (67) to specified read counts. We then processed the subsampled reads with the ssUMI analysis workflows described above.

## PacBio, LoopSeq, and Illumina read processing

PacBio HiFi (CCS) reads of amplicons targeting the full rRNA operon of the ZymoBIOMICS Microbial Community DNA Standard (cat. D6306, lot no: ZRC190811) (22) were downloaded from the European Nucleotide Archive under accession ERR3813246. PacBio HiFi sequences were quality filtered using VSEARCH v11 with the `fastq_filter` command and an EE-rate threshold of 1%. Sequences were trimmed to the same length as the 16S rRNA amplicons generated in this study by truncating the reads at the 8F and 1391R primer sequences using `cutadapt v2.7`, as described above.

Synthetic long reads of amplicons targeting the V1–V9 region of 16S rRNA genes from the ZymoBIOMICS Microbial Community DNA Standard (cat. D6306, Lot ZRC190811) generated with the LoopSeq 16S rRNA Kit and Illumina 2 × 150 bp sequencing (36) were downloaded from NCBI under BioProject PRJNA644197. LoopSeq amplicons were quality filtered using VSEARCH v11 with the `fastq_filter` command and an EE-rate threshold of 1%. Sequences were trimmed at the primers used for their generation (forward: “AGAGTTTGATCMTGGC”; reverse: “TACCTTGTACGACTT”) (36) using `cutadapt v2.7`, as described above.

Paired-end 2 × 250 bp Illumina reads of amplicons targeting the V4 region of 16S rRNA genes from the ZymoBIOMICS Microbial Community DNA standard (cat. D6305) were downloaded from NCBI under SRA accession SRP155048. Illumina V4 amplicons were processed using USEARCH v11. The paired-end reads were first merged using the “`-fastq_mergepairs`” command with “`-fastq_maxdiffs 5 -fastq_pctid 90 -fastq_minqual 5`” arguments, and length filtered to 230 to 300 bp. Adapters were trimmed using the “`-fastx_truncate`” command with “`-stripleft 19 -stripripleft 20`” arguments, and finally the merged and trimmed reads were quality filtered using the `-fastq_filter` command with the “`-fastq_maxee 1.0`” argument.

Illumina V4–V5 amplicons (2 × 300 bp) were processed using DADA2 pipeline (68). Only forward reads were included in the analysis. Quality filtering of forward reads were conducted with the “`filterAndTrim`” function using “`trimLeft = 15, truncLen = 230, maxN = 0, maxEE = 1`” arguments.

## Characterizing sequence accuracy and abundances with microbial community standards

For the ZymoBIOMICS Microbial Community DNA Standard, curated reference rRNA operon sequences (16S–23S rRNA) were obtained from Karst et al. (22). Reference 16S rRNA gene sequences were retrieved using `barmap` (<https://github.com/tseemann/barmap>), trimmed to the 8F/1391R primer sequences using `cutadapt v2.7`, and were de-replicated using USEARCH v11.0 with the “`-fastx_uniques`” command. For the ZymoBIOMICS Gut Microbiome Standard, the genome assembly and polishing approaches were not adequately described by the vendor, and therefore, we downloaded their provided reference genome assemblies for prokaryotic members (RefSeq Accessions: GCA\_028743295.1, GCA\_028743435.1, GCA\_028743335.1, GCA\_028743755.1, GCA\_028743555.1, GCA\_028743355.1, GCA\_028743375.1, GCA\_028743635.1, GCA\_028743535.1, GCA\_028743315.1, GCA\_028743095.1, GCA\_028743735.1, GCA\_028743275.1, GCA\_028743415.1, GCA\_028743255.1, GCA\_028743395.1, GCA\_028743455.1, GCA\_028743775.1, GCA\_028743475.1), concatenated the scaffolds, and polished the assembly using PacBio HiFi reads from three metagenomes of the same mock community (NCBI accession: PRJNA680590) using one round of Racon. Reference 16S rRNA gene sequences were retrieved and de-replicated as described above. Based on this workflow, we detected 66 unique bacterial 16S rRNA gene sequences in the ZymoBIOMICS Gut Microbiome

Standard. However, `ssUMI_std` sequences generated 65 error-free ASVs without the detection of two low-abundant bacterial species (Fig. 4; Table S2); therefore, the true number of bacterial 16S rRNA ASVs in this community is uncertain (Fig. 4).

To assess sequence accuracy of different datasets, reads were mapped to their corresponding 16S rRNA gene reference databases using `minimap2 v2.17` with the “`-cs`” flag, and the mapping statistics were filtered using `samtools v1.9` (69) with “`view -F 2,308`.” The `minimap2` flag “`-ax sr`” was used for mapping Illumina short reads, “`-ax map-ont`” for mapping raw and UMI-corrected Nanopore reads, and “`-ax map-pb`” was used for mapping PacBio HiFi reads. Mapping files were parsed in R v4.2.1, as previously described (22), and error rates were calculated as the sum of mismatches, insertions, and deletions divided by the alignment length. Reads mapping to contaminants (see below) were filtered before summarizing sequence accuracies. Read mapping files were also parsed to investigate the relative read abundances based on total sum scaling of species within the ZymoBIOMICS Gut Microbiome Standard using UMI-based consensus sequences (Fig. 5). All data types (e.g. `ssUMI`, Illumina, PacBio, and LoopSeq) were normalized by random subsampling to the same number of filtered sequences ( $n = 18,000$ ), which was the number of obtained LoopSeq reads, when comparing sequence accuracy in R.

Following previous guidelines used for characterizing sequence accuracy of mock communities with ambiguous reference genomes and closely related strains (39), we manually confirmed the sequence accuracy of ASVs and 97% OTUs generated with the ZymoBIOMICS Gut Microbiome Standard that did not match our reference sequences by querying them with BLASTn against the NCBI `nr` database. If an ASV or OTU sequence matched a 16S rRNA gene with 100% identity and 100% query cover from a species that is present in the Microbiome Standard, it was considered a true positive sequence (Table S15). Otherwise, the sequence feature was assigned the error rate observed via read mapping above. If an ASV or OTU sequence matched a species from a different genus than was present in the Microbiome Standard at >97% identity and 100% query cover, this was considered a contaminant (Table S16). Reads that mapped to contaminant ASVs or OTUs were filtered when characterizing the error rate profiles of Nanopore raw reads, UMI-based consensus sequences, PacBio HiFi reads, and Illumina short reads.

## Wastewater sample analysis

Reads from wastewater samples were processed with the `ssUMI` workflow in standard mode. Microbial community analysis of wastewater near full-length 16S rRNA ASVs was conducted using the `vegan` package in R v4.2.1 (70). The alpha diversity was assessed with ASV richness. Beta diversity was estimated using PCoA with Bray–Curtis dissimilarity.

## Human gut sample analysis

Raw data from Illumina short-read sequencing (2 × 300 bp) of the V4–V5 region of 16S rRNA genes from different human donors was downloaded from the Borealis repository ([doi:10.5683/SP3/TVH0NY](https://doi.org/10.5683/SP3/TVH0NY), 20220831\_Figure\_3\_HumanFecalSamples.zip), and processed using USEARCH v11. The paired-end reads were first merged using the `-fastq_mergepairs` command with “`-fastq_maxdiffs 10 -fastq_pctid 80 -fastq_minqual 5`” arguments, and length filtered to 370–450 bp. Adapters were trimmed using “`-fastx_truncate`” command with “`-stripleft 19 -stripripleft 20`” arguments, and finally the merged and trimmed reads were quality filtered using `-fastq_filter` command with “`-fastq_maxee 1.0`”

argument. Nanopore reads were processed with the ssUMI workflow in standard mode. For comparison of the two sequencing methods, ssUMI consensus sequences were first rarefied using seqtk to the same number of merged and filtered reads obtained with Illumina sequencing ( $6.61 \times 10^5$  reads). Unique sequences from Illumina sequencing (V4–V5) and the ssUMI pipeline (V1–V9) were generated with USEARCH v11 using “-fastx\_uniques” command with a minimum number of sequence observations of two, then separately denoised into ASVs using USEARCH v11 with a minimum size of 10, as described above. Quality-filtered reads and ssUMI reads from each sample were mapped back to their corresponding ASV sequences using USEARCH v11 with the “-usearch\_global” command and “-strand both -id 1.0 -maxaccepts 8 -maxrejects 64 -top\_hit\_only” arguments, to infer abundances. Taxonomy was assigned using SINTAX (71) and SILVA v138 SSU Ref NR99 database (72), using an 80% bootstrap confidence threshold and the “-strand plus” argument. Beta diversity was estimated using PCoA with Bray–Curtis dissimilarity of ASV abundances using the vegan package in R.

The absolute concentration of near full-length 16S rRNA gene copies for each ASV generated with ssUMI pipeline was calculated based on ddPCR measurements, using the following equations:

$$N_{16S_i} = \frac{N_{ddPCR_i}(\text{copies}/\mu\text{L}) \times DF_i \times V_{DNA}(\mu\text{L})}{V_S(\mu\text{L})}, \quad (1)$$

$$N_{ASV_{ij}} = N_{16S_i} \times f_{ASV_{ij}}, \quad (2)$$

where  $N_{16S_i}$  is the total number of full-length 16S gene copies in each sample  $i$  (copies/ $\mu\text{L}$  – stool);  $N_{ddPCR_i}$  is the ddPCR measured total number of full-length 16S gene copies in sample  $i$  ((copies/ $\mu\text{L}$  – stool); see Text S1);  $DF_i$  is the dilution factor used for ddPCR on sample  $i$  (see Text S1);  $V_{DNA}$  is the DNA elution volume during DNA extraction;  $V_S$  is the volume of stool sample used as inoculum;  $N_{ASV_{ij}}$  is the total number of full-length 16S gene copies for ASV  $j$  from sample  $i$ ; and  $f_{ASV_{ij}}$  is the relative abundance of ASV  $j$  from sample  $i$  estimated with the ssUMI workflow.

To assess the correspondence of ASVs generated with the ssUMI workflow to that of Illumina, we mapped all *Enterococcaceae* ASVs generated with ssUMI pipeline to those generated with Illumina V3–V4 16S rRNA sequencing using minimap2 v2.17 with the “-ax map-ont” argument. We also verified the identities of all *Enterococcaceae* ASVs with BLASTn against the NCBI nr database (Table S11).

## Statistical analysis

A correlation test with Pearson's coefficient was conducted with the wastewater samples to examine potential relationships between the number of ssUMI consensus sequences generated per sample and the corresponding number of raw reads per sample ( $n = 87$ ). To compare the ssUMI UMI-tagging efficiencies for wastewater samples and mock communities, a one-sided Student's  $t$  test was performed on the number of ssUMI consensus sequences generated with wastewater samples and ZymoBIOMICS Microbial Community DNA Standard at the same sequencing depth. Wastewater samples were binned into four group based on the sequencing depths (e.g. 0.1–0.2, 0.2–0.3, 0.3–0.7, and 0.7–1.3 M raw reads;  $n = 6, 13, 65,$  and  $5$  samples, respectively), and the average number of ssUMI consensus sequences in each group was compared with that generated after randomly subsampling the raw reads in the mock community to 0.15, 0.25, 0.5, and 1.0 M raw

reads, respectively. The overall performance of ssUMI workflow with different wastewater sample matrices ( $n = 7$  matrices), and human gut samples originating from different donors ( $n = 4$  donors) were further examined using ANOVA on the number of Nanopore raw reads and ssUMI consensus sequences obtained per sample group type ( $n = 87$  wastewater samples total;  $n = 27$  human stool samples total). All statistical tests were performed using R v.4.1.2.

## Acknowledgments

The authors thank Metro Vancouver staff, including Parisa Chegounian, and David Blair, for assisting with wastewater sample collection. They also thank Mads Albertsen for his helpful insights on this work.

## Supplementary Material

Supplementary material is available at PNAS Nexus online.

## Funding

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance Grant (556792-2020) and Discovery Grant (RGPIN-2018-04585) to R.M.Z.

## Author Contributions

X.L. contributed to formal analysis, methodology development, investigation, software, visualization, and writing of the original draft. R.M.Z. contributed to conceptualization, formal analysis, methodology development, investigation, software, visualization, funding acquisition, supervision, project administration, and writing of the original draft. K.W. helped with methodology development and validation. H.G. and C.T. provided resource, and assisted with formal analysis and visualization. J.T. helped with formal analysis and investigation. All authors contributed to the content and revision of the manuscript.

## Preprints

This manuscript was posted on a preprint server at: <https://doi.org/10.1101/2023.06.19.544637>.

## Data Availability

The sequencing data for this project, including both ZymoBIOMICS mock microbial community standards, are available in the NCBI under BioProject PRJNA974480. Accessions for individual wastewater samples and human gut samples are provided in Tables S8 and S9. All code associated with the ssUMI data processing pipeline is available at: <https://github.com/ZielsLab/ssUMI>.

## References

- 1 Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*. 74: 5088–5090.
- 2 Pace NR. 2009. Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev*. 73:565–576.
- 3 Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. 1990. Genetic diversity in Sargasso sea bacterioplankton. *Nature*. 345:60–63.

- 4 Andersson AF, et al. 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One*. 3:e2836.
- 5 Methé BA, et al. 2012. A framework for human microbiome research. *Nature*. 486:215–221.
- 6 Sunagawa S, et al. 2015. Structure and function of the global ocean microbiome. *Science*. 348:1261359.
- 7 Zinger L, et al. 2011. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One*. 6:e24570.
- 8 Sogin ML, et al. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere.”. *Proc Natl Acad Sci U S A*. 103:12115–12120.
- 9 Amann RI, et al. 1990. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Env Microbiol*. 56:1919–1925.
- 10 Amann RI, Krumholz L, Stahl DA. 1990. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *J Bacteriol*. 172:762–770.
- 11 Quince C, et al. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods*. 6:639–641.
- 12 Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Env Microbiol*. 79:5112–5120.
- 13 Caporaso JG, et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*. 6:1621–1624.
- 14 Tringe SG, Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol*. 11:442–446.
- 15 Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 17:333–351.
- 16 Schloss PD. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol*. 6:e1000844.
- 17 Yarza P, et al. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*. 12:635–645.
- 18 Johnson JS, et al. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun*. 10:5029.
- 19 Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ*. 4:e1869.
- 20 Dueholm MKD, et al. 2022. MiDAS 4: a global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nat Commun*. 13:1908.
- 21 Dueholm MS, et al. 2020. Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax). *mBio*. 11:e01557–e01520.
- 22 Karst SM, et al. 2021. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods*. 18:165–169.
- 23 Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. 2014. Accuracy of next generation sequencing platforms. *Next Gener Seq Appl*. 1.
- 24 Kerkhof LJ. 2021. Is Oxford Nanopore sequencing ready for analyzing complex microbiomes? *FEMS Microbiol Ecol*. 97:fiab001.
- 25 Benítez-Páez A, Portune KJ, Sanz Y. 2016. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*. 5:4.
- 26 Santos A, van Aerle R, Barrientos L, Martínez-Urtaza J. 2020. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Comput Struct Biotechnol J*. 18:296–305.
- 27 Tederloo L, Albertsen M, Anslan S, Callahan B. 2021. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Appl Environ Microbiol*. 87:e0062621–e21.
- 28 Quick J, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 530:228–232.
- 29 Zorz J, et al. 2023. SituSeq: an offline protocol for rapid and remote Nanopore 16S rRNA amplicon sequence analysis. *ISME Commun*. 3:33–11.
- 30 Goordial J, et al. 2017. In situ field sequencing and life detection in remote (79°26'N) Canadian high Arctic permafrost ice wedge microbial communities. *Front Microbiol*. 8:2594.
- 31 Castro-Wallace SL, et al. 2017. Nanopore DNA sequencing and genome assembly on the international space station. *Sci Rep*. 7:18022.
- 32 Li C, et al. 2016. INC-Seq: accurate single molecule reads using Nanopore sequencing. *GigaScience*. 5:34.
- 33 Calus ST, Ijaz UZ, Pinto AJ. 2018. NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the Nanopore sequencing platform. *GigaScience*. 7(12):giy140.
- 34 Volden R, et al. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci U S A*. 115:9726–9731.
- 35 Sereika M, et al. 2022. 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*. 19:823–826.
- 36 Callahan BJ, Grinevich D, Thakur S, Balamotis MA, Yehezkel TB. 2021. Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome*. 9:130.
- 37 Ng KM, et al. 2023. Single-strain behavior predicts responses to environmental pH and osmolality in the gut microbiota. *mBio*. 14(4):e0075323. <https://doi.org/10.1128/mbio.00753-23>
- 38 Wagner J, et al. 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol*. 16:274.
- 39 Callahan BJ, et al. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res*. 47:e103–e103.
- 40 Singer E, et al. 2016. High-resolution phylogenetic microbial community profiling. *ISME J*. 10:2020–2032.
- 41 Earl JP, et al. 2018. Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome*. 6:190.
- 42 Curry KD, et al. 2022. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nat Methods*. 19:845–853.
- 43 Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD. 2018. TaxAss: leveraging a custom freshwater database achieves fine-scale taxonomic resolution. *mSphere*. 3(5):e00327–e00318.
- 44 Giovannoni SJ, DeLong EF, Olsen GJ, Pace NR. 1988. Phylogenetic group-specific oligodeoxynucleotide probes for identification of single microbial cells. *J Bacteriol*. 170:720–726.
- 45 Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. 11:2639–2643.

- 46 Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*. 31:3476–3482.
- 47 Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*. 20:129.
- 48 Jain M, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 36:338–345.
- 49 Edgar RC. 2016. UCHIME2: improved chimera prediction for amplicon sequencing. *BioRxiv*. 074252.
- 50 Kivioja T, et al. 2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*. 9:72–74.
- 51 Dottorini G, et al. 2021. Mass-immigration determines the assembly of activated sludge microbial communities. *Proc Natl Acad Sci U S A*. 118:e2021589118.
- 52 Lee S-H, Kang H-J, Park H-D. 2015. Influence of influent wastewater communities on temporal variation of activated sludge communities. *Water Res*. 73:132–144.
- 53 Pan P, Gu Y, Sun D-L, Wu QL, Zhou N-Y. 2023. Microbial diversity biased estimation caused by intragenomic heterogeneity and interspecific conservation of 16S rRNA genes. *Appl Environ Microbiol*. 89:e0210822.
- 54 Boshier FAT, et al. 2020. Complementing 16S rRNA gene amplicon sequencing with total bacterial load to infer absolute species concentrations in the vaginal microbiome. *mSystems*. 5:e00777–e00719.
- 55 Vandeputte D, et al. 2017. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*. 551:507–511.
- 56 Sambrook J, Russell DW. 2006. Purification of nucleic acids by extraction with phenol:chloroform. *CSH Protoc*. 2006:pdb.prot4455.
- 57 Green MR, Sambrook J. 2016. Precipitation of DNA with ethanol. *Cold Spring Harb Protoc*. 2016:pdb.prot093377.
- 58 Tighe S, et al. 2017. Genomic methods and microbiological technologies for profiling novel and extreme environments for the extreme microbiome project (XMP). *J Biomol Tech*. 28:31–39.
- 59 Oxford Nanopore Technology. 2023. Ligation sequencing amplicons—custom PCR UMI (SQK-LSK109). *Nanopore community*. Oxford Nanopore Technologies.
- 60 Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol*. 18:1403–1414.
- 61 Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 4:e2584.
- 62 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 17:10–12.
- 63 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 26:2460–2461.
- 64 Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257. <https://doi.org/10.1101/081257>, preprint: not peer reviewed.
- 65 Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 27:737–746.
- 66 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094–3100.
- 67 Li H. 2012. Seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub*. 767:69.
- 68 Callahan BJ, et al. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods*. 13:581–583.
- 69 Li H, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*. 25:2078–2079.
- 70 Oksanen J, et al. 2022. *vegan: community ecology package*. R Package.
- 71 Edgar RC. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv* 074161. <https://doi.org/10.1101/074161>, preprint: not peer reviewed.
- 72 Quast C, et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 41:D590–D596.