Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing

Søren M. Karst[1,3], Ryan M. Ziels[2,3], Rasmus H. Kirkegaard[1] and Mads Albertsen[1]

Affiliations:
[1]Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Denmark.
[2]Department of Civil Engineering, The University of British Columbia, Vancouver, Canada
[3]These authors contributed equally to this work.

Correspondence:
Mads Albertsen (ma@bio.aau.dk)

Abstract

High-throughput amplicon sequencing of large genomic regions represents a challenge for existing short-read technologies. Long-read technologies can in theory sequence large genomic regions, but they currently suffer from high error rates. Here, we report a high-throughput amplicon sequencing approach that combines unique molecular identifiers (UMIs) with Oxford Nanopore sequencing to generate single-molecule consensus sequences of large genomic regions. We demonstrate the approach by generating nearly 10,000 full-length ribosomal RNA (rRNA) operons of roughly 4,400 bp in length from a mock microbial community consisting of eight bacterial species using a single Oxford Nanopore MinION flowcell. The mean error rate of the consensus sequences was 0.03%, with no detectable chimeras due to a rigorous UMI-barcode filtering strategy. The simplicity and accessibility of this method paves way for widespread use of high-accuracy amplicon sequencing in a variety of genomic applications.

Introduction

High throughput amplicon sequencing is a powerful method for analysing variation in defined genetic regions when sample amounts are limited, insights into low abundant

subpopulations are important, or samples need to be analysed in an economical manner. The method is therefore ideal for studying genetic populations with low abundant variants or high heterogeneity such as cancer driver genes[1–3], virus populations[4–6] and microbial communities[7].

For years, short-read Illumina sequencing has dominated amplicon related research due to its unprecedented throughput and low native error-rate of 0.1%, but with a limitation in maximum amplicon size of ~500 bp (merging of 2x300 bp PE reads)[8]. To enable a lower error-rate and sequencing of longer amplicons, unique molecular identifiers (UMI's) have been applied extensively. Each template nucleotide sequence molecule in a sample is tagged with a UMI sequence consisting of 10-20 random bases. All derived products throughout processing and sequencing will contain the UMI tag, which can subsequently be used to sort and analyse reads based on their original template molecule. This concept has many applications in high-throughput sequencing, such as absolute quantification[9], generating molecule-level consensus sequences with a low error rate[10] , and assembly of synthetic long reads[11]. These applications have enabled key advances across diverse fields of research, such as absolute counting of transcripts in single cells[12], detecting low-frequency cancer mutations in plasma cell-free DNA[13], and generating full-length microbial SSU ribosomal RNA (rRNA) sequences in a high throughput manner[14], to mention a few. The lowest possible error rate of Illumina based consensus sequencing is impressive ($< 10^{-7}$ %), but the upper limit of target length for UMI synthetic long-reads remains approximately 2000 bp due to inefficient cluster generation of longer DNA fragments on the flowcells[15]. UMI-based protocols exist that can generate longer consensus sequences from short reads[16], but they are not widely adopted due to complicated laboratory

protocols. Partitioning based methods such as 10x Genomics and TruSeq Synthetic Long-Reads struggle resolve complex amplicon populations, as there is a high risk of >1 amplicon ending up in the same partition which will result in a chimeric assembly[8]. Lastly, as synthetic long reads depend on *de novo* assembly of the short-reads, this approach will never be able to resolve internal molecule repeats larger than the read length.

In order to analyse amplicons larger than 2000 bp in high throughput, the only feasible approach would be to use long-read sequencing technologies such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). However, these methods are limited by a relatively high raw error-rate of 9 and 13% respectively[17]. For PacBio, the circular consensus sequencing (CCS) approach, where a template molecule is circularized and read multiple times by the polymerase, can produce mean error rates to as low as 0.04%[18]. Strategies also exist for reducing the error-rate of amplicon sequencing on the ONT platform with template circularization and rolling circle amplification before sequencing to generate single-molecule consensus sequences, but these methods suffer from insufficient molecule coverage to effectively reduce mean error rates below 2%[19]. In principle, lower error rates can be achieved with different clustering strategies[19], but at the cost of missing variants, which are critical in many applications[20].

In principle, UMIs can be used in long-read amplicon sequencing to reduce sequencing error-rate, and eliminate PCR artefacts (e.g. chimeras), which are present irrespective of polymerase[21] and can make up > 20% of the amplicons[22]. This is also true for PacBio CSS, where errors are introduced before sequencing during amplicon PCR amplification. Despite the benefits, the combination of UMIs with long-read sequencing is relatively unexplored, and only recently has it been applied with PacBio sequencing, but without

profiling the error of the generated consensus sequences[23,24]. For ONT sequencing the raw error rate of 5-25%[25] has, until now, made it difficult to efficiently extract UMI sequences and confidently determine the true UMI sequences necessary for read binning.

Here, we created a UMI design containing recognizable internal patterns, which together with UMI length filtering now makes it possible to robustly determine true UMI sequences in raw nanopore data. We incorporated this patterned-UMI design into a simple, generally applicable laboratory and bioinformatics protocol that combines UMIs and ONT sequencing of long amplicons (>4500 bp) from low template amounts with high accuracy. As a proof of concept, we apply the method to sequence full-length ribosomal RNA (rRNA) operons in a mock microbial community of eight bacterial species (ZymoBIOMICS Microbial Community DNA Standard) and generate consensus sequences with a mean error rate of 0.03% and no detectable chimeras.

Results and discussion

The method is simple and comprised of two PCR amplifications, Nanopore library preparation, Nanopore sequencing and custom data processing (Figure 1). First, the DNA template is diluted according to the desired number of output sequences. The final yield is impacted by the initial dilution, as well as the amplicon length and PCR efficiency; thus, the dilution should be calibrated empirically for an amplicon target of interest. For rRNA operon sequencing, we found that 5 ng of template produced ~10,000 consensus sequences, and is a good general starting point for further optimization. The genetic region of interest is targeted using 2 cycles of PCR with a custom set of tailed primers, which include a target-specific primer, a UMI sequence and a synthetic priming site used for

downstream amplification (Figure 1A, step 1). Here we used the 27F (16S)[26] and the 2490R (23S)[27] primers to target the bacterial rRNA operon. The result from the initial PCR is a dsDNA amplicon copy of the genetic target with UMIs and synthetic primers in both ends. This template is subsequently amplified by PCR (Figure 1A, step 2) and prepared for long read sequencing, in this case using the using the ONT 1D ligation kit and ONT MinION (Figure 1A, step 3) followed by base-calling. After sequencing, the data is trimmed, filtered and reads are binned according to both terminal UMIs (Figure 1B, steps 1 and 2). To overcome the obstacle of binning UMIs in raw nanapore data with a mean error rate ~9.5%, we designed `patterned` UMIs, with the structure "NNNYRNNNYRNNNYRNNN". The YR [C/T][A/G] patterns limit the length of homopolymer in the UMIs to 4 bases, which mitigates the higher homopolymer error rate present in ONT sequencing[8]. UMI sequences that have a high probability of being correct are detected based on the presence of the above pattern, as well as an expected UMI length of 18 bp. The two terminal UMIs in the amplicons make up a combined UMI pair of 36 bases with a theoretical complexity of 1.2×10^{18} combinations, which means it is extremely unlikely that two molecules contain the same UMI pairs if aiming for 10,000 – 1,000,000 molecules. Chimeric amplicons will form during the later cycles of PCR amplification step, especially if proof-reading polymerases are used[21]. UMI pairs from these chimeric sequences are *de novo* filtered by removing reads with UMI pairs in which either UMI has been observed before in a more abundant UMI pair (Figure 1B, step 2)[28]. The filtered, high-quality UMI pair sequences are used as a reference for binning of the raw dataset according to UMIs (Figure 1B, step 3).

Sequencing of the mock community rRNA operon library resulted in 7.4 Gbp of base-called raw data, of which 3.3 Gbp was binned based on UMIs. The mean read coverage per UMI bin was 67x. The consensus sequence for each UMI bin was generated by initially finding the centroid sequence in the bin, and polishing this centroid with all the data in the UMI bin using 5 rounds of racon[29] followed by 2 rounds of Medaka (Figure 1B, step 4).

Initially, we observed error-rates that were highly correlated with the individual rRNA operons in the Zymo mock (Supplementary Figure 4), which indicated errors in the available reference genomes, as was also reported by others[18]. The reference genomes were generated using the Unicycler assembler with both Illumina and Nanopore reads and polished with pilon (personal communication with Zymo Research). As Unicycler uses a short-read assembly as starting point[30] and short-read polishing has been used for final curation, repeat regions are bound to contain errors resulting from ambiguous assembly and mapping[31]. To generate improved rRNA operon references, we first used a long-read assembly approach, in which publicly available ONT sequence data of the Zymo mock community[32] was assembled into individual reference genomes with miniasm[33] followed by racon and Medaka polishing. rRNA operons were then extracted from the high-quality long-read assemblies, and SNPs with no Illumina short-read support were manually curated, which were mainly indel errors in homopolymers. In total, we found 49 bacterial rRNA operons with 4-10 copies/species, where 44 operons were unique and had 1-379 intra-species difference (Supplementary Figure 2). The mean difference between the original references and our curated sequences was 0.063% (~2.8 SNP/operon), with a range of 0 – 0.47% (0 – 21 SNP/operon) (Supplementary Figure 3).

A total of 9759 amplicon UMI consensus sequences with an average length of 4372 bp were generated with a read coverage of ≥ 30x, a mean error rate of 0.03% and no detected chimeras (Figure 2C). Of these sequences, 2570 were perfect with no errors. The error rate is markedly different in non-homopolymer regions compared to homopolymer regions (Figure 2B). The non-homopolymer error rate stabilizes above a coverage of 10x for all error types (deletions, inserts and mismatches), with mismatches contributing to a majority of the remaining error (Figure 2D). Within homopolymer regions, the error rate is higher and continues to drop beyond 100x coverage, which is primarily due to the indel errors (Figure 2B). The mismatch error rate is similar between non-homopolymer and homopolymer regions over all coverage values. This demonstrates that the major obstacles for achieving a lower error rate are generally mismatch errors, as well as indel errors specifically in the homopolymer regions. The mismatch error rate of 0.012-0.016% is most likely derived from the 2 cycles of initial PCR performed to target the rRNA operon. For this PCR, Platinum Taq DNA high-fidelity polymerase (Thermo Fisher) was used, which should have an error rate in the range of 0.003 - 0.005% (6x lower than Taq) [34,35] per duplication which theoretically would result in a cumulative error rate over 2 PCR cycles of up to 0.01%. Other high-fidelity polymerases with lower error rates were tested, but we were unable to consistently produce amplicons, which we might be due to unwanted intra- or inter-molecule annealing. The homopolymer indel error rate is a consequence of the nanopore read-head structure in the CsgG pore used in the current R9.4 chemistry[8]. Generally, the homopolymer indel rate depends on homopolymer length and specific nucleotide (Supplementary Table 2), i.e. A-homopolymers have markedly lower errors than G-homopolymers. Yet, a closer inspection of the homopolymer error rates reveals a more complicated picture. For example, some positions of 3xC

homopolymers contained more frequent insertions than longer C-homopolymers (Supplementary Figure 1). This problem is likely rooted within the calibration of the neural networks of the base-caller and consensus algorithms[36], and is bound to change significantly in the future, and will probably be reduced with the introduction of the R10 pores. Despite residual systematic errors, the error-rate presented here is the lowest documented for long read amplicons yet (Supplementary Table 4). We did not identify any chimera's in the generated long-read amplicon data.

An important application of high-accuracy amplicon sequencing is the ability to confidently call variants, even if they are present in low relative abundance. To test our method, we performed naive variant calling based on the consensus sequences. Consensus sequences initially were grouped via clustering, and SNPs within each cluster were phased and called as a variant if present ≥2x coverage. Subsequently, the consensus reads were binned according to variants, and variant consensus sequences were generated. To reduce impact of systematic homopolymer errors, the homopolymers were masked before phasing and variant calling, and reintroduced before final consensus calling. Of 44 unique rRNA operons, 40 variant consensus sequences were found with no errors, and 4 with 1 error in homopolymer regions (Figure 3B, Supplementary Table 3). An additional 26 spurious variants were detected with a mean error count of 1.4 (0.03% error rate) and a maximum of 3 (0.07% error rate). These spurious variants are supported by 1.6% of the total data, and seem to occur due to systematic errors at specific positions outside homopolymer regions.

The relative rRNA operon abundance within each species were very similar, as was expected (Figure 3C). For some species the internal coverage variance was small (E. coli

percent sd=4.9) and for others it is higher (L. fermentum sd=12.8) (Supplementary Figure 6 and Supplementary Table 5). By investigating the read coverage of the mock community genomes within the publicly-available metagenomic nanopore data[32], we found evidence of heavy coverage skew across the genome in some species, likely due to different growth rates of the cultures at the time of sampling (Figure 3D, Supplementary Table 7). This skew can impact the relative template abundances of the operons up to +/- 50% (Supplemental Table 5), depending on their distance to the origin of replication, and could to some degree explain the variance we see among inter-species operon abundances. The observed relative abundance between species did not match the theoretical abundance for all species reported by the vendor (Supplementary Table 6). Possible explanations are erroneous mixing of the mock community, species-dependent DNA fragment size, PCR primer mismatch, operon/genome GC content, and different amplification efficiencies. To our surprise, none of these potential causes could alone explain the observed discrepancy in relative abundance (Supplementary Tables 6-7 and Figures 7-9). However, it is evident that multiple factors have to be considered when interpreting this kind of data, especially template DNA size distribution impact on template availability (Supplementary Figure 7), growth dependent coverage bias (Supplementary Figure 6) and template amplification efficiency (Supplementary Figure 9).

The data presented here was generated in 48 hrs (6 hrs lab work, 24 hrs sequencing, 6 hrs data processing) at a reagent cost of 1100 USD, which is ~0.1 USD/consensus sequence. Using this method on the PacBio Sequel system with the SMRT Cell 1m chips, we anticipate the output would be around 100,000 UMI consensus sequences at a cost of ~0.02 USD/consensus sequence with a marginally better error rate, as the PacBio errors

seem more random and therefore better suited for consensus calling[37]. The throughput will likely change by a factor of 10x with the introduction of Sequel II and the SMRT Cell 8m chips. The turnaround for PacBio sequencing is theoretically < 24 hrs, but as most users would need sequencing out-of-house, this is more likely > 7 days. We predict that the ease of use, fast turn-around time and accessibility will favour sequencing of high-accuracy amplicons on the ONT platform.

Over the past several decades, the amplification and sequencing of ribosomal RNA (rRNA) genes, primarily 16S and 18S, has become an integral method used to study the diversity and taxonomic composition of microbial communities in a variety of environments[38]. With our method, it is now possible to effortlessly improve upon high-throughput sequencing of environmental samples with databases based on full rRNA operon (SSU-ITS-LSU), which has not been previously feasible due to the length of the operon (≈ 5 kbp) and the method limitations aforementioned. A database of full operon rRNA sequences will help improve upon rRNA phylogeny, allow higher phylogenetic resolution[39–42], especially critical if the method is applicable to eukaryotes[43,44], and will present a wider range of target regions for designing short-read amplicon sequencing assays and fluorescent in situ hybridization probes[45,46].

High-accuracy amplicon sequencing of long targets has many applications, and the ease and accessibility of this method now makes it possible for the wider scientific community to develop new solutions – all one needs is a modified version of their favourite primers, a few generic molecular laboratory instruments, and a MinION starter kit from Oxford Nanopore Technologies. While the residual error rate in the Nanopore consensus data is

negligible, the remaining systematic indel errors could still be an issue in some contexts, such as sensitive assays where low abundant variants are important, or if shifts in reading frames cannot be tolerated. These systematic indel errors will hopefully be solved soon, and until then, this method can be applied with the PacBio platform for the specific purposes above. By exchanging the initial PCR for a ligation step, high-accuracy amplicon sequencing could also be applied to fragmented DNA with tight size distributions (5-15 kbp) to produce long reads with low error rate, which holds great promise for human genome sequencing[47] and resolving strain-diversity in metagenomes[48].

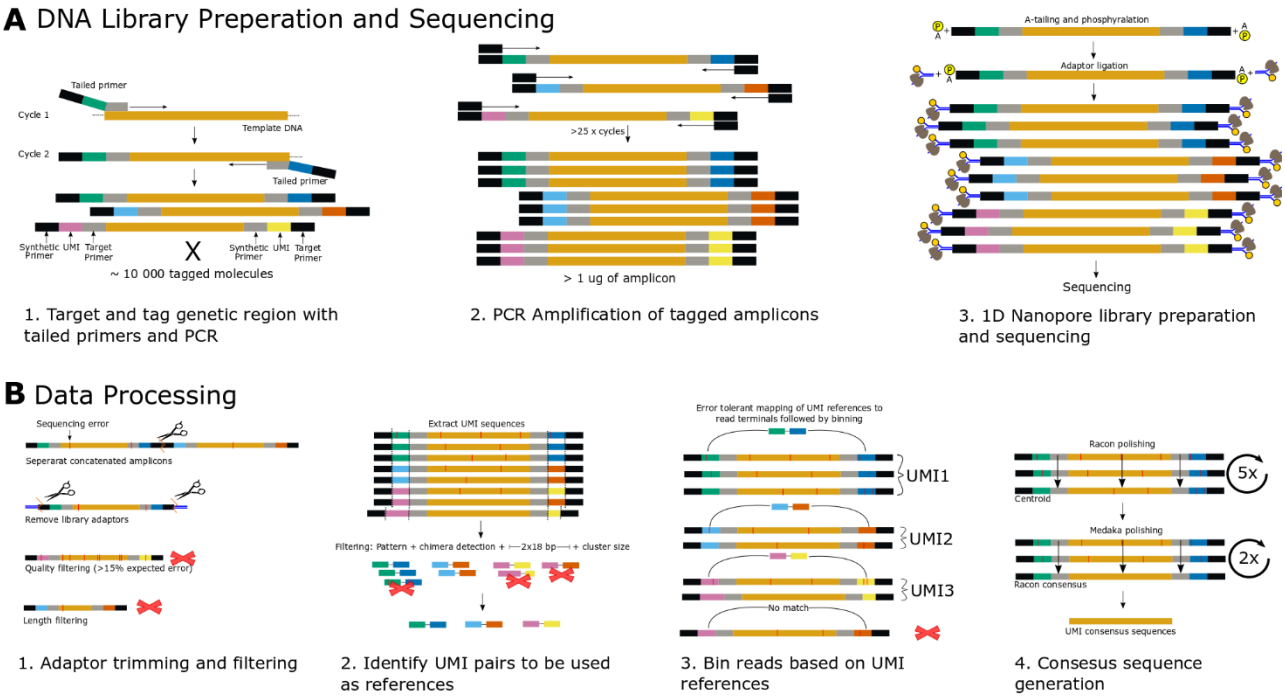


Figure 1: Overview of laboratory (A) and bioinformatics workflow (B).

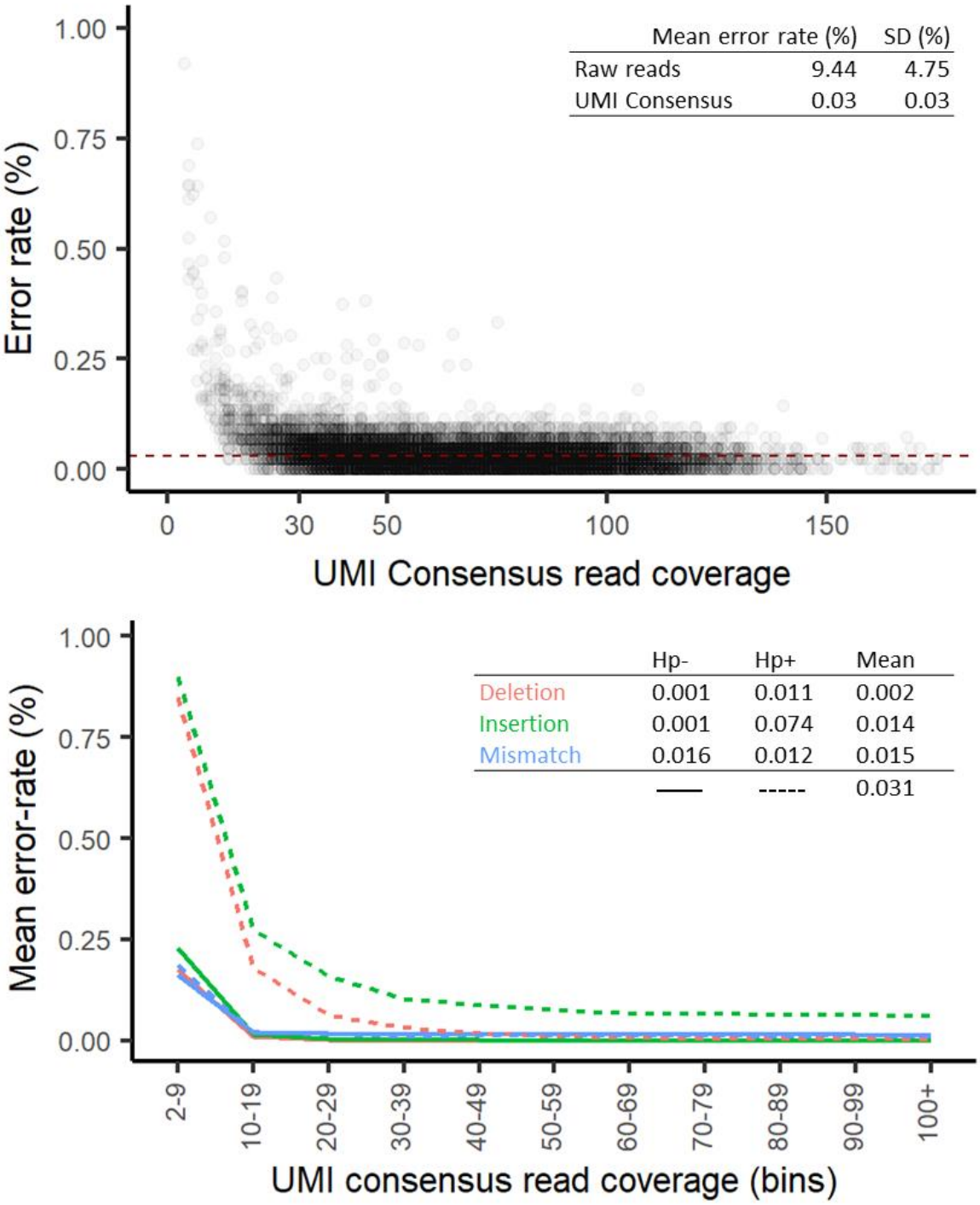


Figure 2: (A) Error rate as function of the number of reads in each UMI bin. The inlaid table shows the mean error rate of raw reads and UMI consensus sequences with a coverage >= 30. (B) Error rate as function of the number of reads in each UMI bin split by error type and whether the error fell inside (Hp+) or outside (Hp-) a homopolymer region. The inlaid table gives the mean error rates summaries for all error and homopolymer types for bins with >= 30 coverage.

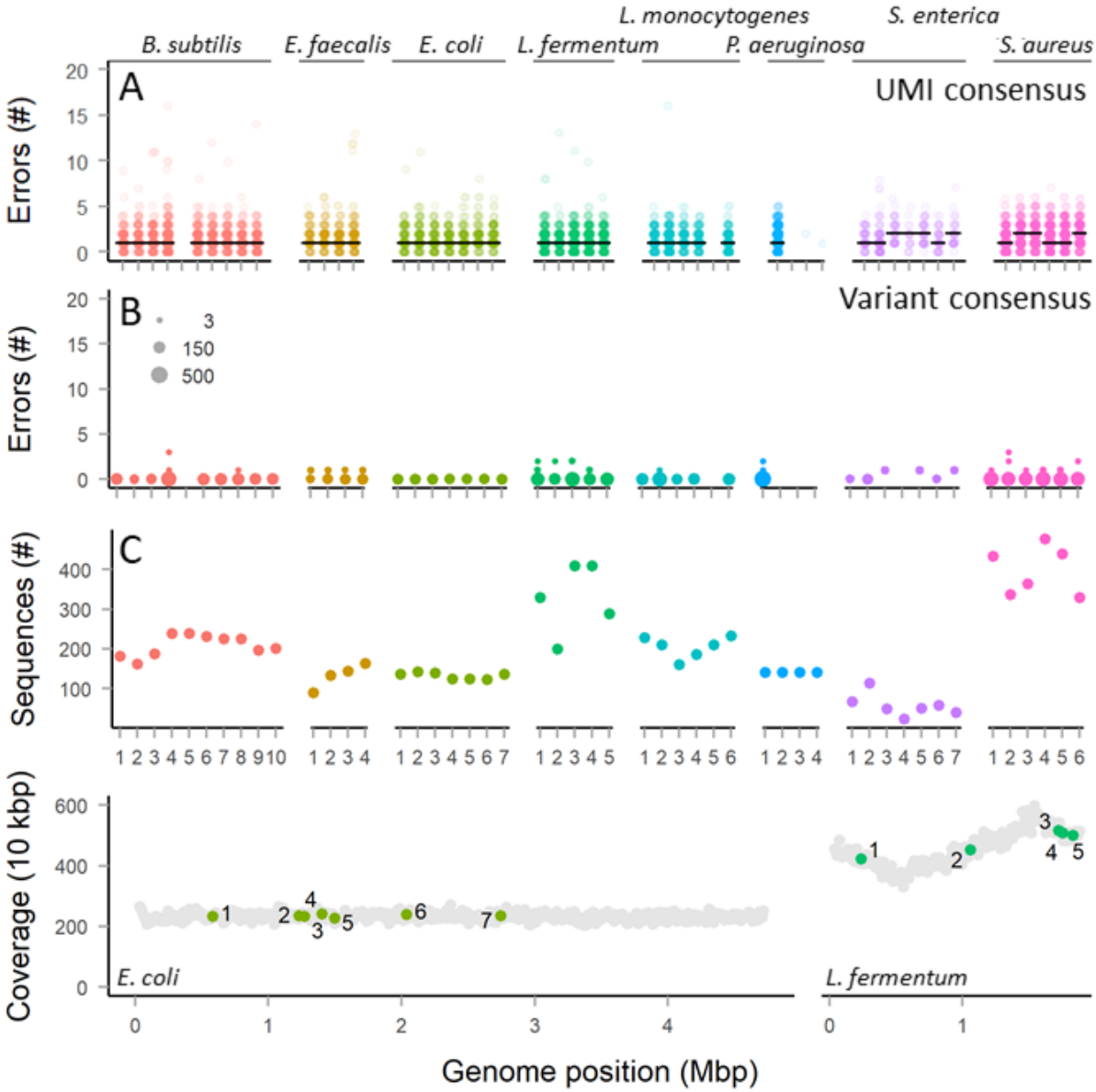


Figure 3: UMI sequencing results for the mock community. Note that the following operons are identical: *B. subtilis* 4 & 5; *L. monocytogenes* 2 & 5; *P. aeruginosa* 1-4. The *S. enterica* operon 2 & 4 differ by 1 bp within a 7 bp homopolymer. A) The number of errors in the UMI consensus sequences. Black bars represent the median number of errors. B) The number of errors in the variant consensus sequences. The size of the circles are scaled by the number of UMI consensus sequences in each variant. C) Number of UMI consensus sequences pr. operon. The count for identical operons have been divided evenly on each operon. D) Coverage profile of the mock community based on shotgun Nanopore sequencing data. Each grey point is a 10 kbp average. Colored points represents the position of the individual rRNA operons.

References

1. Meldrum, C., Doyle, M. A. & Tothill, R. W. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* **32**, 177–95 (2011).

2. Guibert, N. *et al.* Amplicon-based next-generation sequencing of plasma cell-free DNA for detection of driver and resistance mutations in advanced non-small cell lung cancer. *Ann. Oncol.* **29**, 1049–1055 (2018).

3. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci.* **105**, 13081–13086 (2008).

4. Goldsmith, D. B., Parsons, R. J., Beyene, D., Salamon, P. & Breitbart, M. Deep sequencing of the viral phoH gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral phoH genes in the Sargasso Sea . *PeerJ* **3**, e997 (2015).

5. Adriaenssens, E. M. & Cowan, D. A. Using Signature Genes as Tools To Assess Environmental Viral Ecology and Diversity. *Appl. Environ. Microbiol.* **80**, 4470–4480 (2014).

6. Uyaguari-Diaz, M. I. *et al.* A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* **4**, 20 (2016).

7. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**, 4516–22 (2011).

8. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

9. Fu, G. K., Hu, J., Wang, P. & Fodor, S. P. A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 9026–31 (2011).

10. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 9530–5 (2011).

11. Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7**, 119–122 (2010).

12. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–6 (2014).

13. Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).

14. Karst, S. M. *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **36**, 190–195 (2018).

15. Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. *Europe PMC Funders*

Group Improved Protocols for Illumina Sequencing. *Current Protocols in Human Genetics* **18**, (2009).

16. Stapleton, J. A. *et al.* Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing. 1–20 (2016). doi:10.5061/dryad.kr8kk

17. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Res.* **46**, 2159–2168 (2018).

18. Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *bioRxiv* 392332 (2018). doi:10.1101/392332

19. Calus, S. T., Ijaz, U. Z. & Pinto, A. J. NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience* **7**, 1–16 (2018).

20. Hathaway, N. J., Parobek, C. M., Juliano, J. J. & Bailey, J. A. SeekDeep: Single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Res.* **46**, 1–13 (2018).

21. Sze, M. A. & Schloss, P. D. The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. *bioRxiv* 565598 (2019). doi:10.1101/565598

22. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. 494–504 (2011). doi:10.1101/gr.112730.110.Freely

23. Karlsson, K. & Linnarsson, S. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* **18**, 1–11 (2017).

24. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).

25. Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner. 1–14 (2018).

26. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, 1–11 (2013).

27. Hunt, D. E. *et al.* Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Appl. Environ. Microbiol.* **72**, 2221–5 (2006).

28. Burke, C. M. & Darling, A. E. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ* **4**, e2492 (2016).

29. Nagarajan, N., Mile, Š., Vaser, R. & Sovic, I. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 1–10 (2017). doi:10.1101/gr.214270.116.5

30. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**,

1–22 (2017).

31. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).

32. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**, 1–7 (2019).

33. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).

34. Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by Single- Molecule Sequencing. 1–19 (2017). doi:10.1371/journal.pone.0169774

35. Tindall, K. R. & Kunkel, T. A. Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry* **27**, 6008–6013 (1988).

36. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *bioRxiv* 1–14 (2019). doi:10.1101/543439

37. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).

38. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Front. Microbiol.* **8**, 1–22 (2017).

39. Martijn, J. *et al.* Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. *Environ. Microbiol.* **00**, 1462-2920.14636 (2019).

40. Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J. & Martin Embley, T. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. B Biol. Sci.* **279**, 4870–4879 (2012).

41. Ferla, M. P., Thrash, J. C., Giovannoni, S. J. & Patrick, W. M. New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. *PLoS One* **8**, 1–14 (2013).

42. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

43. Mallatt, J. M., Garey, J. R. & Shultz, J. W. Ecdysozoan phylogeny and Bayesian inference: First use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.* **31**, 178–191 (2004).

44. Marvaldi, A. E., Duckett, C. N., Kjer, K. M. & Gillespie, J. J. Structural alignment of 18S and 28S rDNA sequences provides insights into phylogeny of Phytophaga (Coleoptera: Curculionoidea and Chrysomeloidea). *Zool. Scr.* **38**, 63–77 (2009).

45. Stoffels, M., Amann, R., Ludwig, W., Hekmat, D. & Schleifer, K. H. Bacterial community dynamics during start-up of a trickle-bed bioreactor degrading aromatic

compounds. *Appl. Environ. Microbiol.* **64**, 930–939 (1998).

46. Fang, Q., Brockmann, S., Botzenhart, K. & Wiedenmann, A. Improved detection of Salmonella spp. in foods by fluorescent in situ hybridization with 23S rRNA probes: a comparison with conventional culture methods. *J. Food Prot.* **66**, 723–31 (2003).

47. Wenger, A. M. *et al.* Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *bioRxiv* 519025 (2019). doi:10.1101/519025

48. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).

Methods

Sources of DNA

The ZymoBIOMICS Microbial Community DNA Standard (D6305, lot no. ZRC190633) was obtained from Zymo Research (Irvine, California). The mock community DNA contained genomic material from 10 species (8 bacteria and 2 yeasts): *Bacillus subtilis*, *Cryptococcus neoformans*, *Enterococcus faecalis*, *Escherichia coli*, *Lactobacillus fermentum*, *Listeria monocytogenes*, *Pseudomonas aeruginosa*, *Saccharomyces cerevisiae*, *Salmonella enterica*, *Staphylococcus aureus*. Note, 2 of the yeast species were not targeted by PCR amplification of rRNA operons. The concentration of mock DNA was measured on a Qubit 3.0 fluorometer and Qubit dsDNA HS assay kit (Thermo Fisher Scientific) and the quality of the mock DNA was measured by gel electrophoresis on an Agilent 2200 Tapestation using Genomic screentapes (Agilent Technologies).

DNA Sequence Library Preparation

Target gene and add UMIs

PCR was used to target the bacterial 16S-23S rRNA operon and simultaneously tag each template molecule with terminal unique molecular identifiers (UMIs).

The following primers were used for the PCR. Forward primer (ncec_16S_8F_v7): 5'-CAAGCAGAAGACGGCATACGAGAT NNNYRNNNYRNNNYRNNN AGRGTTYGATYMTGGCTCAG. Reverse primers (ncec_23S_2490R_v7): 5'-AATGATACGGCGACCACCGAGATC NNNYRNNNYRNNNYRNNN CGACATCGAGGTGCCAAAC. The first section of both primers is a synthetic priming site used for downstream amplification. The second section is the `patterned` UMI consisting of a total of 12 random nucleotides (N) and 6 degenerate nucleotides (Y or R) which results in a total

of $1.2x10^{18}$ possible UMI combinations if the UMIs in both ends of a molecule are concatenated (4^{12*2} x 2^{6*2} = $1.2x10^{18}$). The last section of the primers consists of the rRNA operon specific primer site for 27f[1] and 2490r[2], respectively.

The PCR reaction contained 10 ng of ZymoBIOMICS Microbial Community DNA Standard, 0.5 U Platinum Taq DNA Polymerase High Fidelity (Thermo Fisher Scientific, USA), and a final concentration of 1x High Fidelity PCR buffer, 100 mM of each dNTP, 1.5 mM $MgSO^4$, 500 nM of each ncec_16S_8F_v7/ ncec_23S_2490R_v7 primers in 50 µL. The PCR program consisted of initial denaturation (3 minutes at 95˚C) and 2 cycles of denaturation (30 seconds at 95˚C), annealing (30 seconds at 55˚C) and extension (6 minutes at 72˚C). The PCR product was purified using CleanPCR (CleanNA, Netherlands) following the manufacturer's instructions (CleanPCR, manual revision v1.02) with the exception of an EtOH concentration of 80%, post wash dry time of < 3 minutes and 0.6x bead solution/sample ratio.

Amplification of UMI tagged amplicons

A second PCR was used to amplify the UMI-tagged template molecules. All of the UMI-tagged template molecules were added to the reaction along with a final concentration of 1x High Fidelity PCR buffer, 100 mM of each dNTP, 1.5 mM $MgSO^4$, 500 nM of each ncec_pcr_fw_v7 (5- CAAGCAGAAGACGGCATACGAGAT)[1] and ncec_pcr_rv_v7 (5- AATGATACGGCGACCACCGAGATC)[1] primers and 0.5 U Platinum Taq DNA Polymerase High Fidelity (Thermo Fisher Scientific, USA) in 100 µL. The PCR program consisted of initial denaturation (3 minutes at 95˚C) and then 25 cycles of denaturation (15 seconds at 95˚C),

[1] Oligonucleotide sequences © 2007-2018 Illumina, Inc. All rights reserved. Illumina adaptor sequences were used as synthetic priming sites, as they are proven to work robustly for PCR amplification and to allow the option to validate libraries with Illumina sequencing, which is useful during troubleshooting.

annealing (30 seconds at 60˚C) and extension (6 minutes at 72˚C) followed by final extension (5 minutes at 72˚C). The PCR product was purified using a custom bead purification protocol "SPRI size selection protocol for >1.5-2 kb DNA fragments" (Oxford Nanopore, England) based on: dx.doi.org/10.17504/protocols.io.idmca46. CleanPCR (CleanNA, Netherlands) bead solution was used for preparing the custom buffer. The purification was performed according to the custom protocol with the exception of an EtOH concentration of 80% and 0.9x bead solution/sample ratio. The concentration and quality of the PCR amplicons was measured as described before.

To obtain sufficient PCR product for Oxford Nanopore sequencing, a third PCR was performed using amplicons from the second PCR and the same procedure as before, but with 4 x 100 µl reactions and 10 cycles of amplification. The final amount of amplicon generated was 10 µg in 55 µL.

DNA Sequencing

2000 ng of the purified amplicon from the third PCR was used as template for library preparation using the protocol "1D amplicon/cDNA by ligation (SQK-LSK109)" (Oxford Nanopore, England) with omission of the AMPure purification after the end-prep step. A R9.4.1 FLO-MIN106 flowcell was used for sequencing on a MinION and MinKNOW v18.12.9 (Oxford Nanopore, England). Basecalling was performed with Guppy v3.0.3 in GPU mode and the dna_r9.4.1_450bps_hac.cfg model (Oxford Nanopore, England).

Data generation workflow

Trimming and filtering of raw data

Raw fastq sequence data was adaptor trimmed using porechop with the commands: *--min_split_read_size 3500 --adaptor_threshold 80 --extra_end_trim 0 --extra_middle_trim_good_side 0 --extra_middle_trim_bad_side 0 --middle_threshold 80 --check_reads 1000* (v0.2.4 https://github.com/rrwick/Porechop). Additionally, the adaptors.py file in porechop was modified to include possible end-to-end ligation combinations of the custom primers (ncec_pcr_fw_v7/ ncec_pcr_rv_v7 5-GTCTTCTGCTTGAATGATACGGCG; ncec_pcr_fw_v7/ ncec_pcr_fw_v7 5-GTCTTCTGCTTGCAAGCAGAAGAC; ncec_pcr_rv_v7/ ncec_pcr_rv_v7: 5-CGCCGTATCATTAATGATACGGCG). The custom settings and modifications to adaptors.py were necessary to correctly split amplicons concatenated in the ligation step of the library preparation, which made up a substantial amount of the data.
The adaptor trimmed data was filtered using filtlong --min_length 3500 --min_mean_q 70 (v0.2.0 https://github.com/rrwick/Filtlong) and cutadapt[3] (v2.1) *-m 3500 –M 6000*. The final result from these pre-processing steps was trimmed and filtered raw read data.

Extraction of UMI reference sequences

To efficiently bin reads according to UMIs, it was critical to extract and validate true UMI sequences that could be used as references. UMI sequences of the correct length (18 bp) were extracted from the reads by locating the flanking sequences within the custom primers. The first 200 bp from each terminal end of all reads were extracted using *awk*, and saved into individual files. UMI sequences were extracted from each file with cutadapt[3] (v2.1) in paired-end input mode, using the commands: *-e 0.2 -O 11 -m 18 -M 18 --discard-untrimmed -g CAAGCAGAAGACGGCATACGAGAT…AGRGTTYGATYMTGGCTCAG -g AATGATACGGCGACCACCGAGATC…CGACATCGAGGTGCCAAAC -G*

GTTTGGCACCTCGATGTCG…GATCTCGGTGGTCGCCGTATCATT -G CTGAGCCAKRATCRAACYCT…ATCTCGTATGCCGTCTTCTGCTTG. This step insured that only reads with UMIs of the correct length in both ends were extracted. UMI pairs were then concatenated and filtered to remove UMI pairs that did not follow the expected pattern (NNNYRNNNYRNNNYRNNNNNNYRNNNYRNNNYRNNN). Filtered UMI pairs were clustered using usearch[4] (v11.0.667) with the commands: *-fastx_uniques -minuniquesize 2 -strand both and usearch -cluster_fast -id 0.85 -centroids -sizein -sizeout -strand both*. Potential chimeras were removed by filtering all UMI pairs containing a single UMI that was observed in another UMI pair with a higher abundance. The final result from these steps was a list of trusted UMI pairs that could be used as references for binning reads.

Binning reads according to UMI

The first 55-65 bp of each terminal of the trimmed and filtered reads were extracted with awk and saved into individual files. The UMI pair reference sequences were split into their corresponding single UMIs and mapped to the read terminals using bwa[5] (v0.7.17-r1198-dirty) with the commands: *index, aln -n 3 –N*, and *samse –n 10000000*. The mapping results were then filtered using samtools[6] (v1.9) with the command *view -F 20*. Mapping results from each end of the reads were merged, and a read was assigned to a specific UMI pair reference if two conditions were met: A) the UMI was the best hit; B) the mapping difference between the query read and each sub UMI was ≤ 3 bp. Based on these designations, the trimmed and filtered reads were divided into UMI bins.

Generation of UMI consensus sequences

For each individual UMI bin, a consensus sequences was initially generated using usearch (v11.0.667) with the commands *-cluster_fast -id 0.75 -strand both -centroids*, and picking the most abundant centroid. The centroid sequence was used as template for five rounds of polishing using all the UMI bin reads with minimap2[7] (v2.16-r922) with the command *-x ava-ont* and racon[8] (v1.3.1) with the command *-m 8 -x -6 -g -8 -w 800*. The racon-polished consensus sequence was further polished using all of the reads in that UMI bin using two rounds of Medaka (v0.7.0) with the commands *-m r941_min_high_model.hdf5* (https://github.com/nanoporetech/medaka). The polished consensus sequences from all UMI bins were then pooled and trimmed and filtered using cutadapt with the commands *-m 3000 -M 6000 –g AGRGTTYGATYMTGGCTCAG…GTTTGGCACCTCGATGTCG*. Consensus sequences not containing both primers were discarded.

Phasing of consensus sequences

Consensus sequences were phased and used to call variants using a custom workflow. The homopolymers were masked in the consensus sequences by converting homopolymers of length ≥3 into length 2 to prevent them from impacting the phasing. The masked consensus sequences were clustered using two rounds of usearch with the commands *-cluster_fast -id 0.995 -strand both -consout -clusters -sort length -sizeout*, and removing clusters of size < 3. The reads belonging to each cluster were mapped back to the consensus sequence of the cluster using minimap2 with the command *–ax asm5*. Genotype likelihoods were estimated from the mappings with bcftools[9] (v1.9) with the command *mpileup –Ov –d 1000000 –L 1000000 –a "FORMAT/AD,FORMATDP"*, and the results were filtered to show positions of SNPs present in

≥2x coverage using bcftools view -i 'AD[0:1-]>2' for each cluster. The list of SNP positions were used to phase the reads within a cluster, and a variant was called if ≥3 reads supported a combination of SNPs. Consensus reads were then grouped according to called variants, and consensus sequences were generated for each variant group. First, the homopolymers were unmasked in the consensus reads and a crude variant consensus was generated using usearch with commands *-cluster_fast -id 0.99 -strand both -consout –sizeout*. The crude variant consensus was polished with workflow using minimap2 with commands *–ax map-ont, bcftools mpileup –Ov –d 1000000 –L 1000000 –a "FORMAT/AD,FORMAT/DP"*, bcftools *norm –Ov, bcftools view -i 'AD[0:1]/FORMAT/DP>0.5' –Oz* and bcftools *consensus*.

Pipeline parallelization

Many steps in the pipeline has been parallelized using GNU parallel[10].

Generation of Reference Sequences for Mock Community

We obtained raw fast5 files from a previously-reported sequencing effort of the ZymoBIOMICS Microbial Community DNA Standard using Oxford Nanopore Technologies GridION flowcells (available from: https://github.com/LomanLab/mockcommunity). The fast5 data was basecalled using the GPU-basecaller guppy v. 2.2.3 with "flipflop" mode. The basecalled reads mapped to the existing reference sequences using minimap2 (v.2.12) using default settings. The mapped reads were assembled separately for each reference using minimap2 (v.2.12) to create overlaps and miniasm (v.0.3) to perform the assembly using default settings. The reads were mapped to the assembled genomes using minimap2 (v.2.12) using default settings and racon (v.1.3.1) was used to retrieve corrected consensus sequences using default settings. The corrected sequences were

subsequently polished with medaka (v.0.6.0, https://github.com/nanoporetech/medaka) with the "r941_flip_model" model. Ribosomal RNA operons were extracted from the draft reference genome assemblies using *in silico* PCR with our forward and reverse primers using the *ipcress* command from the package exonerate (v.2.2), and were verified with genome coordinates for rRNA operons predicted by barrnap (v.0.9) (available from: https://github.com/tseemann/barrnap).

To further remove any residual errors from the rRNA operon reference sequences after assembly and polishing, high-quality short reads generated from Illumina sequencing were downloaded from NCBI for each bacterial strain in the mock community (accessions: ERR2935851, ERR2935850, ERR2935852, ERR2935857, ERR2935854, ERR2935853, ERR2935848, ERR2935849) and used for final polishing. The Illumina reads were randomly subsampled to an expected average coverage of 100 for each bacterial strain using the *sample* command in seqtk (v.1.0) (available from: https://github.com/lh3/seqtk). The subsampled Illumina reads were mapped to the draft rRNA operon sequences using minimap2 with the settings: *-ax sr*. The BAM files were sorted and indexed by samtools. We performed variant calling using bcftools (v1.9) with the commands *mpileup* and *call* using the settings: *ploidy =1*. Variant calls were filtered using bcftools *filter* with the settings: *quality > 200*. Variant calls were manually inspected and corrected, if needed, by visualizing mapping profiles in CLC Workbench. Polished consensus sequences were generated with bcftools *consensus* to generate high-quality references for use in benchmarking error rates in this study.

Data analysis

Chimera detection

Chimeras in the consensus sequences were detected by usearch[12] with the commands *-uchime2_ref -strand plus -mode sensitive,* using our curated rRNA operon reference sequences from the ZymoBIOMICS Microbial Community DNA Standard (see above).

Error profiling

Detection of error was based on a mapping of the sequence data (raw reads, consensus sequences, variant consensus sequences) to our curated rRNA operon reference sequences from the ZymoBIOMICS Microbial Community DNA Standard (see above). Mapping was performed with minimap2 *-ax map-ont --cs* and filtered using samtools *view -F 2308*. The references and mappings were imported into R software environment[13] (v3.5.1), where errors in the sequences were profiled using mainly the tidyverse (v1.2.1 https://www.tidyverse.org/) and Biostrings[14] (v2.48.0) R-packages and custom scripts (see Code availability). In brief, errors and their type (mismatch, deletion, insert) were detected from the SAM --cs tags. The relative positions of the errors was determine in respect to the reference and this was used to categorize the errors as being homopolymers errors (hp+) or no (hp-). The error information was combined with metadata (UMI bin sizes, most similar reference etc.) and used to explore and visualize error as function of different parameters.

Exploration of relative abundance inconsistencies

We observed a difference between the relative abundance estimated with our UMI consensus data and the theoretical abundance for the rRNA operons of the mock community. We investigated several different potential causes of this discrepancy by importing relevant data and

metadata into the R software environment[13] (v3.5.1), using mainly the tidyverse (v1.2.1 https://www.tidyverse.org/) and Biostrings[14] (v2.48.0) R-packages and custom scripts (see Code availability).

Validate content of ZymoBIOMICS mock. Oxford Nanopore data from the ZymoBIOMICS Microbial Community DNA Standard described above was used for the analysis. The data was divided per species and imported into R. Based on read lengths, the total bp count was estimated for each species, and used together with the theoretical genome sizes and rRNA operon copy numbers to estimate the theoretical relative abundance of 16S (equal to rRNA operons). The read length data was used to estimate the amount of DNA theoretically available for rRNA operon PCR. A DNA fragment has to be equal to or larger than the rRNA operon to be a valid PCR template. Furthermore, DNA fragments are generated randomly and break points introduced within the operon will also render the DNA fragment useless as a template for PCR. Hence, all fragments below 4500 bp were discarded and 4500 bp were subtracted from all longer fragment lengths > 4500 bp to take broken operons into account. Based on the adjusted read lengths we estimated an adjusted theoretical relative abundance of 16S rRNA.

Investigate impact of GC and operon length. Possible impact of GC content (genome/rRNA operon) and operon read lengths was investigated by plotting relative difference between observed abundance and theoretical abundances.

Investigate PCR primer match. A bias in relative abundance can be introduced in the first PCR where the rRNA operon is targeted with region specific primers. If there are mismatches between

primers and template, we would expect a lower annealing/amplification efficiency. Primer/template mismatches were estimated using *ipcress* as described above.

Investigate PCR amplification bias. A bias in relative abundance can also be introduced in the second PCR where the UMI tagged amplicons are amplified with > 25 cycles of PCR. If a specific template has a relatively poor amplification efficiency we would expect this to impact the general bin size of this template. To investigate this, we imported UMI bin size statistics and UMI classifications into R and plotted bin sizes as function of species, operon and operon size.

Analysis of genomic coverage skew due to growth. A bias in relative abundance could also occur due to the mock species being in different growth phases at the time of sampling. To investigate the potential contribution of growth to coverage bias, we used the previously generated genomes of the mock community species. Nanopore data was mapped to each species genome using minimap2 *-ax map-ont* and calculated genome position depth using samtools. Ribosomal RNA operon genome coordinates were predicted by barrnap as described before. The data was imported into R, and used to create read coverage plots.

Code Availability

Source code and analysis scriptes are freely available at https://github.com/SorenKarst/longread-UMI-pipeline

Data Availability

Raw and assembled sequencing data is available at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under the project number PRJEB32674 and a complete data overview can be found in supplementary table 8.

References

1. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, 1–11 (2013).
2. Hunt, D. E. *et al.* Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Appl. Environ. Microbiol.* **72**, 2221–5 (2006).
3. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
4. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
5. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
6. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
7. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
8. Nagarajan, N., Mile, Š., Vaser, R. & Sovic, I. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 1–10 (2017). doi:10.1101/gr.214270.116.5
9. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
10. Tange, O. GNU Parallel: the command-line power tool. ;*login USENIX Mag.* **36**, 42–47 (2011).
11. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**, 1–7 (2019).
12. Edgar, R. C. UCHIME2 : improved chimera prediction for amplicon sequencing. (2016).
13. Team, R. C. R: A Language and Environment for Statistical Computing. (2018).
14. DebRoy, H. P. and P. A. and R. G. and S. Biostrings: Efficient manipulation of biological strings. (2018).

Supplementary information

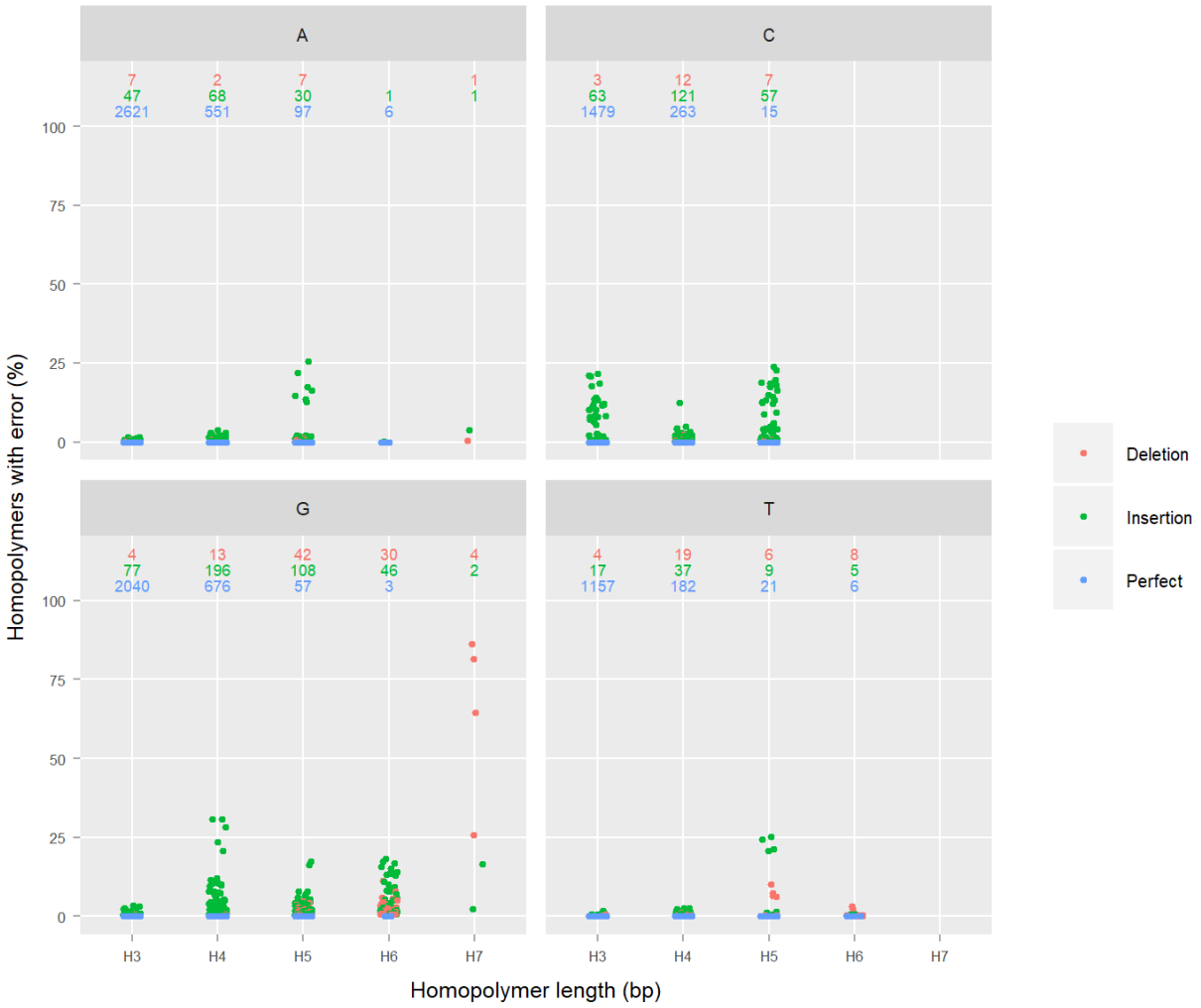


Figure S1: Error as function of homopolymers in reference sequences. All homopolymers in the reference sequences (44 unique curated rRNA operons from the ZymoBIOMICS Microbial Community DNA Standard) have been categorized according to nucleotide (four boxes) and length of homopolymer (x-axis). Each dot represents a specific homopolymer in a specific reference operon. The y-axis denotes the fraction of sequences with a specific operon that either has a deletion, insertion or is perfect. The numbers at the top of the boxes show the total number of homopolymers in each category.

		1	2	3	4	5	6	7	8	9	10
Bacillus_1	1		177	177	0	0	0	0	0	0	0
Bacillus_2	2	183		0	177	177	177	177	177	177	177
Bacillus_3	3	183	2		177	177	177	177	177	177	177
Bacillus_4	4	3	182	184		0	0	0	0	0	0
Bacillus_5	5	3	182	184	0		0	0	0	0	0
Bacillus_6	6	2	183	185	3	3		0	0	0	0
Bacillus_7	7	1	184	184	4	4	1		0	0	0
Bacillus_8	8	5	184	186	4	4	3	4		0	0
Bacillus_9	9	4	183	185	3	3	4	5	5		0
Bacillus_10	10	4	183	185	1	1	4	5	5	4	

		1	2	3	4	5	6	7
Escherichia_1	1		117	117	1	117	12	40
Escherichia_2	2	238		0	116	0	123	95
Escherichia_3	3	238	2		116	0	123	95
Escherichia_4	4	18	223	223		116	11	39
Escherichia_5	5	239	1	1	224		123	95
Escherichia_6	6	47	232	234	31	233		46
Escherichia_7	7	223	147	149	207	148	199	

		1	2	3	4
Pseudomonas_1	1		0	0	0
Pseudomonas_2	2	0		0	0
Pseudomonas_3	3	0	0		0
Pseudomonas_4	4	0	0	0	

		1	2	3	4	5	6	7
Salmonella_1	1		0	162	1	162	0	263
Salmonella_2	2	1		162	1	162	0	263
Salmonella_3	3	254	253		161	0	162	101
Salmonella_4	4	2	1	252		161	1	262
Salmonella_5	5	253	254	3	253		162	101
Salmonella_6	6	2	1	254	2	253		263
Salmonella_7	7	378	379	128	378	127	380	

		1	2	3	4
Enterococcus_1	1		0	102	102
Enterococcus_2	2	5		102	102
Enterococcus_3	3	108	106		0
Enterococcus_4	4	109	105	1	

		1	2	3	4	5
Lactobacillus_1	1		205	10	10	205
Lactobacillus_2	2	216		195	195	0
Lactobacillus_3	3	18	206		0	195
Lactobacillus_4	4	19	207	9		195
Lactobacillus_5	5	219	12	207	208	

		1	2	3	4	5	6
Staphylococcus_1	1		46	268	167	174	231
Staphylococcus_2	2	54		222	121	128	185
Staphylococcus_3	3	314	266		145	98	197
Staphylococcus_4	4	173	127	173		51	108
Staphylococcus_5	5	217	171	104	74		101
Staphylococcus_6	6	243	204	248	123	147	

		1	2	3	4	5	6
Listeria_1	1		0	251	251	0	0
Listeria_2	2	1		251	251	0	0
Listeria_3	3	263	262		0	251	251
Listeria_4	4	262	261	1		251	251
Listeria_5	5	1	0	262	261		0
Listeria_6	6	2	1	261	262	1	

Figure S2: Difference between intra species rRNA operons. Each table show intra species difference between rRNA operons. Below the diagonal is total differences and above is total indels. The analysis was performed on the curated rRNA operons from the ZymoBIOMICS Microbial Community DNA Standard using CLC genomics workbench v9.5.5 (Qiagen) using the ´Create Alignment´ tool (Gap open cost = 10.0, Gap extension cost = 1.0, End gap cost = Free, Alignment mode = Very accurate (slow), Redo alignments = No, Use fixpoints = No) and the `Create pairwise comparison` tool (default settings).

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
S. enterica 1	S. enterica 1	0	0	1	0.02
S. enterica 2	S. enterica 2	0	0	3	0.07
S. enterica 3	S. enterica 3	1	0	1	0.04
S. enterica 4	S. enterica 2	1	2	1	0.09
S. enterica 5	S. enterica 5	1	0	1	0.04
S. enterica 6	S. enterica 6	0	0	0	0.00
S. enterica 6	S. enterica 4	1	0	1	0.05
S. enterica 7	S. enterica 7	0	0	0	0.00

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
L. monocytogenes 1	L. monocytogenes 1	0	0	0	0.00
L. monocytogenes 2	L. monocytogenes 2 & 5	0	0	0	0.00
L. monocytogenes 2	L. monocytogenes 6	0	0	1	0.02
L. monocytogenes 3	L. monocytogenes 3	0	0	1	0.02
L. monocytogenes 4	L. monocytogenes 4	0	0	0	0.00
L. monocytogenes 5	L. monocytogenes 2 & 5	0	0	1	0.02
L. monocytogenes 6	L. monocytogenes 2 & 5	0	2	1	0.07

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
P. aeruginosa 1	P. aeruginosa 1 & 2 & 3 & 4	0	0	0	0.00
P. aeruginosa 2	P. aeruginosa 1 & 2 & 3 & 4	0	0	0	0.00
P. aeruginosa 3	P. aeruginosa 1 & 2 & 3 & 4	0	0	0	0.00
P. aeruginosa 4	P. aeruginosa 1 & 2 & 3 & 4	0	0	0	0.00

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
S. aureus 1	S. aureus 1	0	0	0	0.00
S. aureus 2	S. aureus 2	0	0	0	0.00
S. aureus 3	S. aureus 3	0	0	1	0.02
S. aureus 4	S. aureus 4	8	10	1	0.43
S. aureus 5	S. aureus 5	0	0	0	0.00
S. aureus 6	S. aureus 6	0	0	0	0.00

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
B. subtilis 1	B. subtilis 1	0	0	0	0.00
B. subtilis 2	B. subtilis 2	0	0	0	0.00
B. subtilis 3	B. subtilis 3	0	0	0	0.00
B. subtilis 4	B. subtilis 4	0	0	0	0.00
B. subtilis 5	B. subtilis 5	0	0	0	0.00
B. subtilis 6	B. subtilis 6	0	0	0	0.00
B. subtilis 7	B. subtilis 7	0	0	0	0.00
B. subtilis 8	B. subtilis 8	0	0	0	0.00
B. subtilis 9	B. subtilis 9	0	0	0	0.00
B. subtilis 10	B. subtilis 10	0	0	0	0.00

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
E. coli 1 & 4	E. coli 1	3	2	15	0.45
E. coli 1 & 4	E. coli 4	2	0	4	0.13
E. coli 2 & 5	E. coli 2	0	0	9	0.20
E. coli 2 & 5	E. coli 5	0	0	10	0.22
E. coli 3	E. coli 3	0	0	10	0.23
E. coli 6	E. coli 6	0	0	6	0.13
E. coli 7	E. coli 7	1	1	19	0.47

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
E. faecalis 1	E. faecalis 4	0	0	2	0.05
E. faecalis 2	E. faecalis 3	0	0	1	0.02
E. faecalis 3	E. faecalis 2	0	0	0	0.00
E. faecalis 4	E. faecalis 1	0	0	0	0.00

Mock References (Zymo)	Curated References	Del.	Ins.	Mism.	Error (%)
L. fermentum 1	L. fermentum 3	0	0	4	0.09
L. fermentum 2 & 4	L. fermentum 2	0	0	5	0.11
L. fermentum 2 & 4	L. fermentum 5	0	0	7	0.16
L. fermentum 3	L. fermentum 1	0	0	4	0.09
L. fermentum 5	L. fermentum 4	0	0	4	0.09

Figure S3: Comparison between non-curated and curated rRNA reference sequences. Comparison between curated and non-curated reference sequences at species level. Each overview table corresponds to a species in the mock community, in which each ZymoBIOMICS reference operon is listed next to the closest corresponding curated reference, differences between the two divided by type (deletion, insert and mismatch) and total error rate. The determined differences between the ZymoBIOMICS references and our curated references were determined through minimap2.

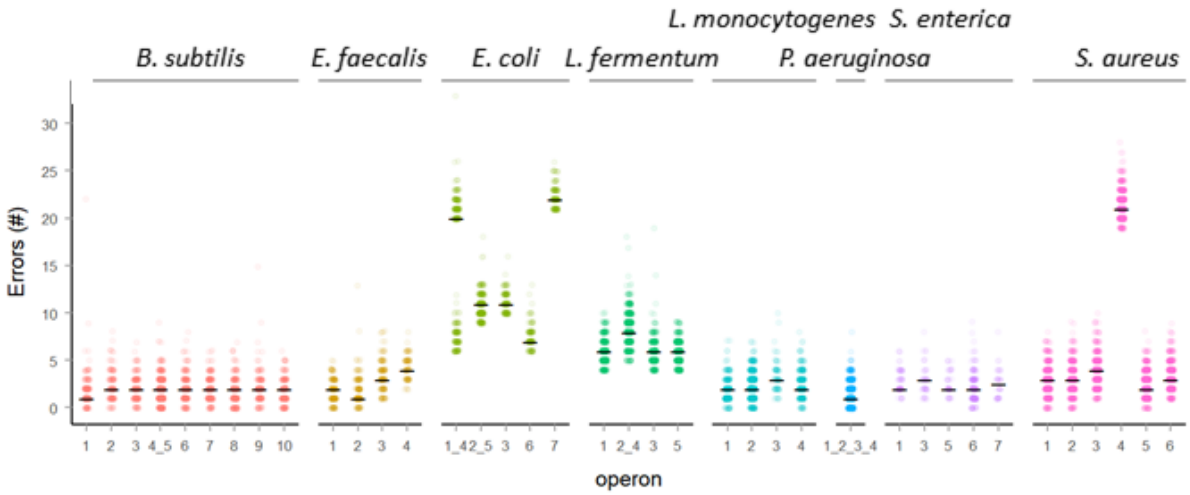
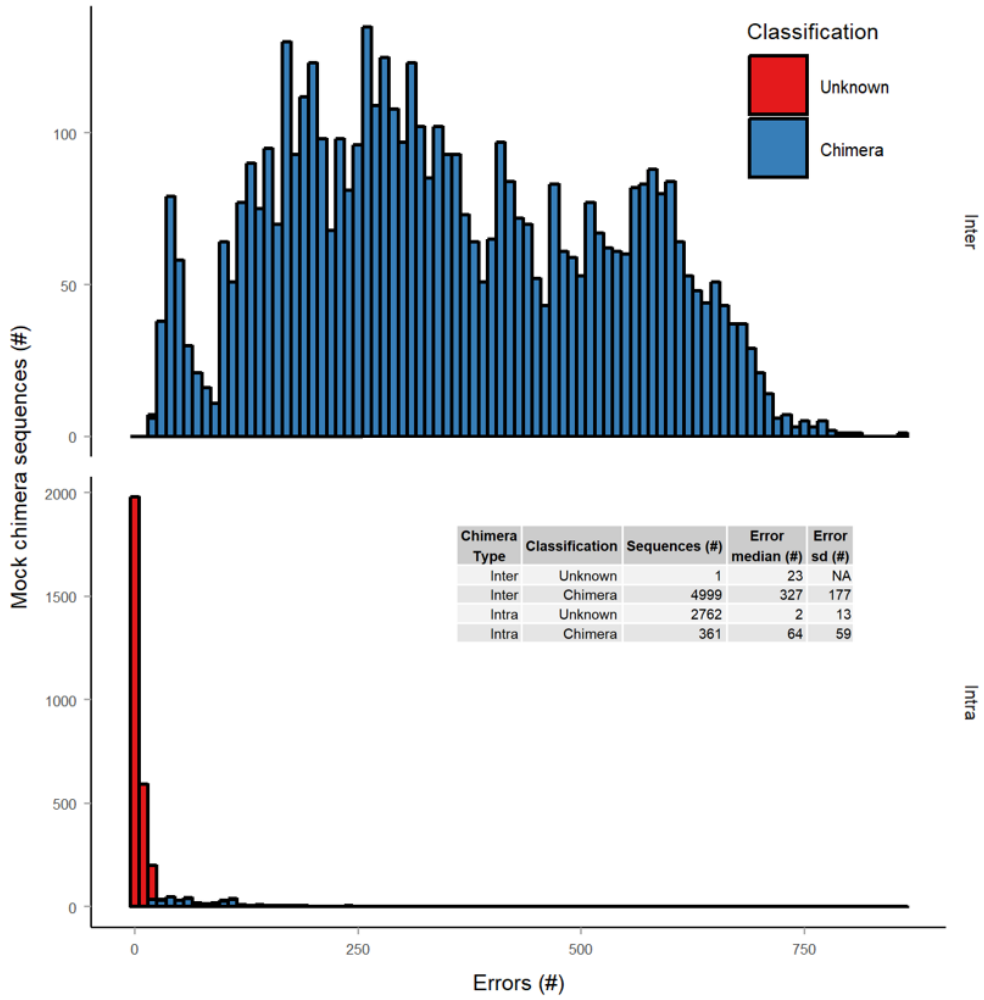


Figure S4: Number of errors in UMI consensus using non-curated references. The number of errors in the UMI consensus sequences estimated based on non-curated rRNA reference sequences. Each point represents a single polished consensus sequence that aligns to a specific reference operon. Black bars represent the median number of errors.



Chimera Type	Classification	Sequences (#)	Error median (#)	Error sd (#)
Inter	Unknown	1	23	NA
Inter	Chimera	4999	327	177
Intra	Unknown	2762	2	13
Intra	Chimera	361	64	59

Figure S5: Validation of chimera detection. Chimera detection is notoriously difficult when sequencing errors are present and uchime2_ref, which we used, will only call a chimera if a sequence is an error-free combination of the references[1]. We estimate that approximately ~10% of our consensus sequences are error-free, and hence the chimera detection only works as intended on that fraction of the data. To validate that closely related chimeras would be identified with uchime2_ref, we generated a mock chimera dataset from the references sequences, which had from 1 to 842 bp differences to the closest matching references. 99.98% of the inter species chimeras (n = 5000) were detected and 11.6% of the intra species chimeras (n = 3123). The plot shows the test results; the data is divided by inter and intra species chimeras and the x-axis shows number of differences between chimera and closest matching reference and the y-axis shows number of chimeras. It is mainly chimeras with few SNPs that an not classified.

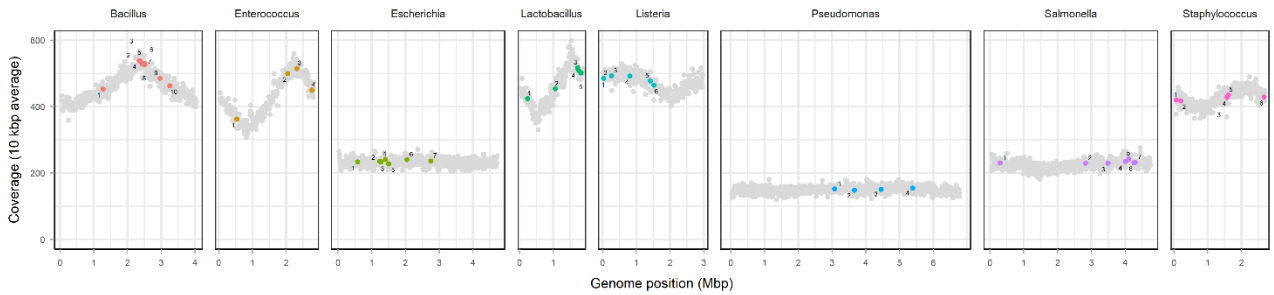


Figure S6: Mock community metagenome read coverage across genomes. Coverage profile of the mock community based on shotgun Nanopore sequencing data. Each grey point is a 10 kbp average coverage value. Colored points represents the position of the individual rRNA operons.

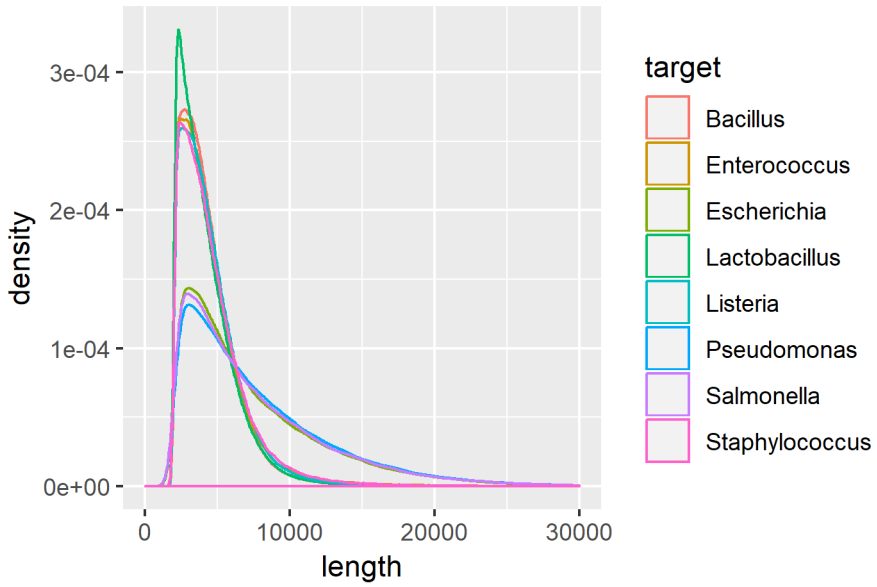


Figure S4: Read size distribution of mock community metagenome data. Each line plot represents the read size distribution from each mock community species estimated from the Nanopore metagenome data. Some species have significantly more high molecular DNA over 5000 bp compared to some of the other species, which is important for effective template availability in PCR. The distributions seem to be gram+/- dependent. Were different DNA extraction kits used?

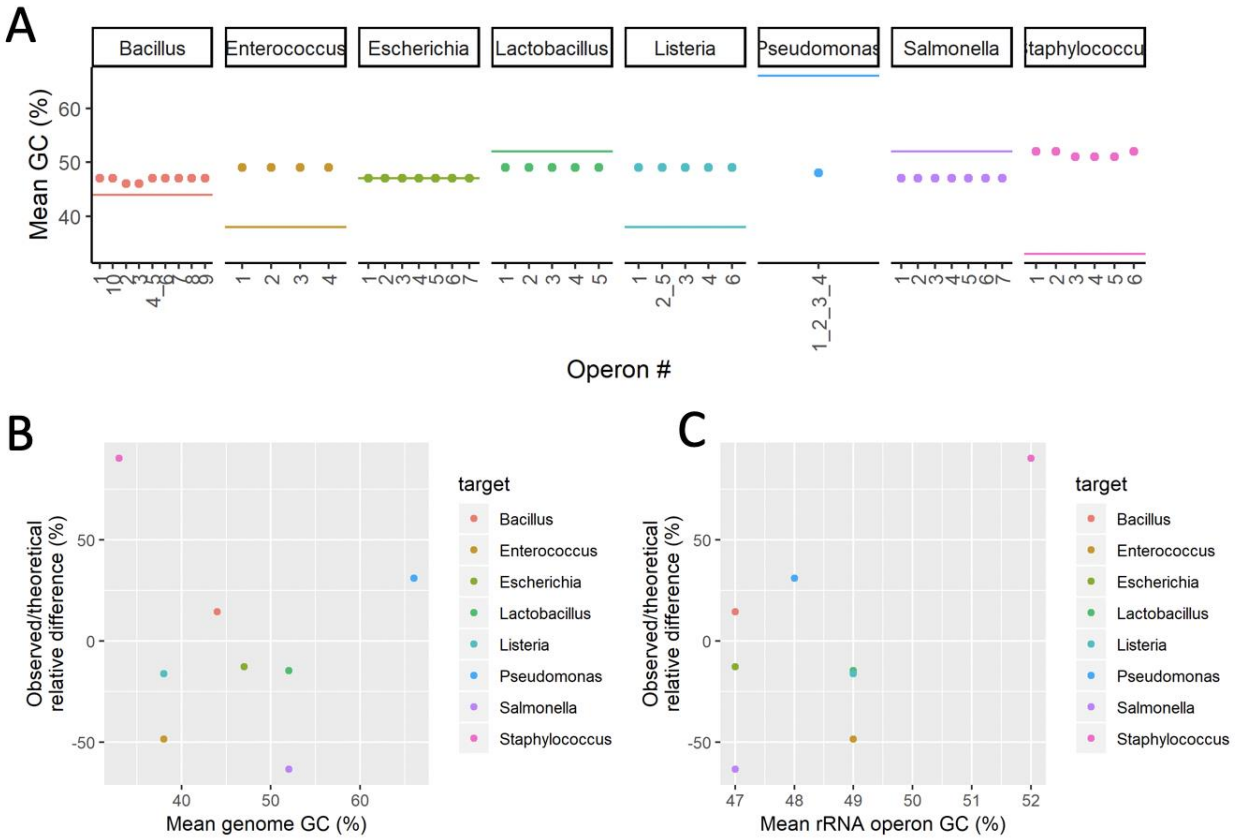


Figure S8: Correlation between GC content and difference in observed and theoretical relative abundance. (A) Mean GC content of the genome (solid lines) along with the GC content of each rRNA operon (points), arranged by species. (B) Fractional difference in percent read abundances (relative to ZymoBIOMICS values) of rRNA operons versus mean genome GC content. Mean genome GC content was determined from the assembled genomes. (C). Fractional difference in relative read abundances of rRNA operons (relative to ZymoBIOMICS values) versus mean rRNA operon GC content. The mean rRNA operon GC content was averaged across all operons of each respective species.

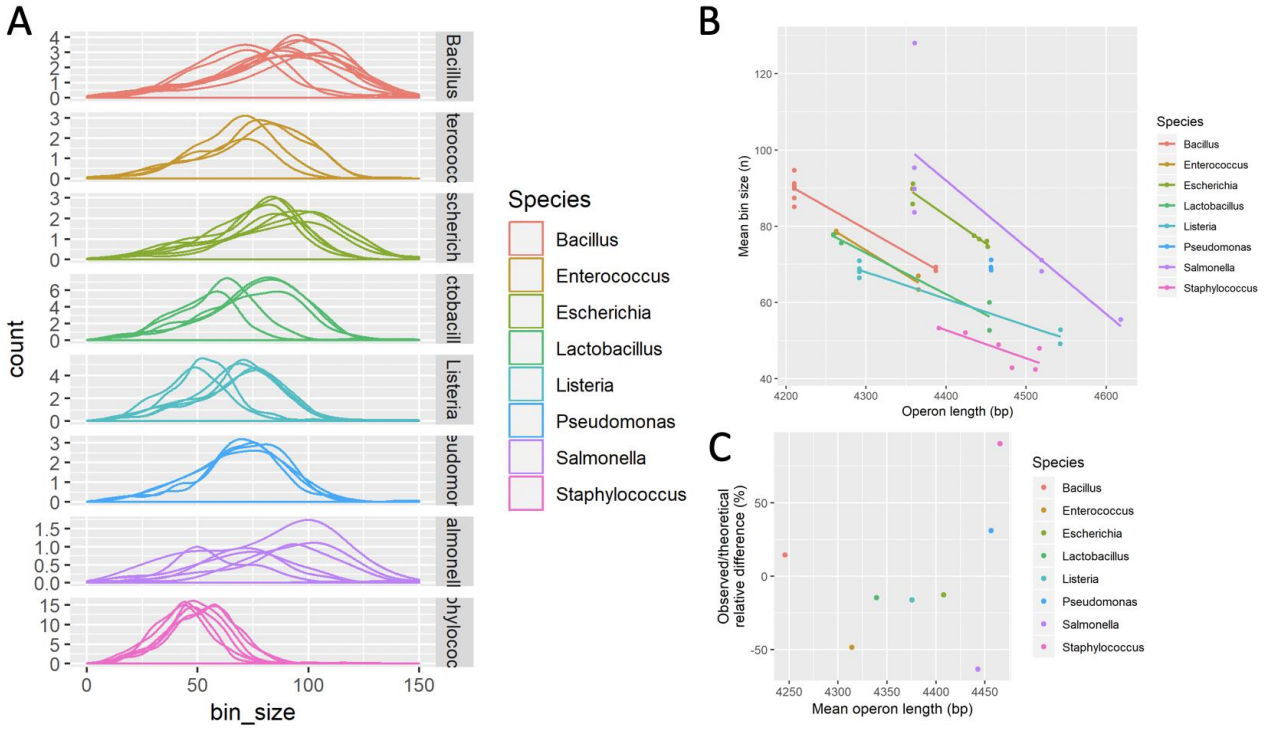


Figure S9: Correlation between operon length and difference in observed and theoretical relative abundance. (A) Density plots showing the number of sequences versus UMI bin size for all rRNA operons in each species. (B) Linear correlation between mean bin size and rRNA operon length, coloured by species. (C) Scatter plot of fractional difference in percent read abundances (relative to ZymoBIOMICS values) versus mean rRNA operon length.

Table S1: Error rate and types depending on UMI read coverage. 'Consensus coverage' refers to the UMI bin size, the error rate is averaged for all consensus sequences in that bin size range, and 'mm', 'ins', and 'del' refer to mean absolute counts of mismatches, insertions, and deletions, respectively.

Consensus coverage	Error rate (%)	mm	ins	del	Sequences (n)
0-10	0.72	6.33	13.57	11.65	51
10-20	0.12	0.97	2.55	1.58	238
20-30	0.06	0.72	1.37	0.54	612
30-40	0.04	0.71	0.86	0.26	1246
40-50	0.04	0.66	0.73	0.16	1392
50-60	0.03	0.69	0.63	0.09	1232
60-70	0.03	0.63	0.54	0.07	1133
70-80	0.03	0.66	0.52	0.06	1201
80-90	0.03	0.64	0.50	0.05	1170
90-100	0.03	0.61	0.50	0.04	918
100+	0.02	0.58	0.47	0.02	1384

Table S2: Error rate (%) divided by homopolymer type and length. 'hp-' indicates non-homopolymer regions, while homopolymer regions are separated by length (3-7 bp) and nucleotide type (A,C,G,T).

hp size (bp)	A	C	G	T
hp-	0.001	0.001	0.001	0.001
3	0.004	0.086	0.012	0.003
4	0.026	0.105	0.152	0.057
5	0.175	1.086	0.438	1.19
6	0.017	-	1.409	0.136
7	0.825	-	8.961	-

Table S3: List of called variants from UMI consensus data.

The error rates shown are relative to the curated reference database. The closest operon relative is shown in the far left column, and the assigned variant name in the far right column.

rname	del	ins	mm	error	cluster_size	cluster_name
Bacillus_1	0	0	0	0.0000	180	Cluster26_var1
Bacillus_10	0	0	0	0.0000	196	Cluster26_var10
Bacillus_2	0	0	0	0.0000	163	Cluster16_var2
Bacillus_3	0	0	0	0.0000	185	Cluster16_var1
Bacillus_4_5	0	0	0	0.0000	450	Cluster26_var5
Bacillus_4_5	0	0	1	0.0238	3	Cluster27_var36
Bacillus_4_5	0	3	0	0.0713	3	Cluster27_var14
Bacillus_6	0	0	0	0.0000	227	Cluster26_var3
Bacillus_7	0	0	0	0.0000	220	Cluster26_var2
Bacillus_8	0	0	0	0.0000	211	Cluster27_var2
Bacillus_8	0	1	0	0.0237	4	Cluster26_var11
Bacillus_9	0	0	0	0.0000	187	Cluster26_var6
Enterococcus_1	0	0	0	0.0000	75	Cluster19_var2
Enterococcus_1	1	0	0	0.0229	13	Cluster19_var4
Enterococcus_2	0	0	0	0.0000	111	Cluster19_var1
Enterococcus_2	1	0	0	0.0229	22	Cluster19_var3
Enterococcus_3	0	0	0	0.0000	132	Cluster25_var3
Enterococcus_3	1	0	0	0.0235	11	Cluster25_var6
Enterococcus_4	0	0	0	0.0000	147	Cluster25_var2
Enterococcus_4	1	0	0	0.0235	16	Cluster25_var5
Escherichia_1	0	0	0	0.0000	135	Cluster12_var1
Escherichia_2	0	0	0	0.0000	140	Cluster20_var2
Escherichia_3	0	0	0	0.0000	139	Cluster20_var3
Escherichia_4	0	0	0	0.0000	125	Cluster11_var1
Escherichia_5	0	0	0	0.0000	123	Cluster20_var1
Escherichia_6	0	0	0	0.0000	122	Cluster15_var1
Escherichia_7	0	0	0	0.0000	137	Cluster13_var1
Lactobacillus_1	0	0	0	0.0000	320	Cluster21_var1
Lactobacillus_1	0	0	1	0.0234	7	Cluster21_var3
Lactobacillus_1	0	1	1	0.0469	3	Cluster21_var2
Lactobacillus_2	0	0	0	0.0000	196	Cluster5_var1
Lactobacillus_2	0	1	1	0.0449	3	Cluster5_var2
Lactobacillus_3	0	0	0	0.0000	395	Cluster23_var1
Lactobacillus_3	0	1	1	0.0470	9	Cluster23_var5
Lactobacillus_4	0	0	0	0.0000	399	Cluster23_var2
Lactobacillus_4	0	0	1	0.0235	8	Cluster24_var3
Lactobacillus_5	0	0	0	0.0000	289	Cluster6_var1
Listeria_1	0	0	0	0.0000	225	Cluster22_var2
Listeria_2_5	0	0	0	0.0000	409	Cluster22_var1
Listeria_2_5	0	0	1	0.0233	6	Cluster22_var16
Listeria_3	0	0	0	0.0000	160	Cluster2_var1
Listeria_4	0	0	0	0.0000	186	Cluster2_var2
Listeria_6	0	0	0	0.0000	228	Cluster22_var3
Pseudomonas_1_2_3_4	0	0	0	0.0000	554	Cluster4_var1
Pseudomonas_1_2_3_4	0	0	1	0.0224	7	Cluster4_var2
Pseudomonas_1_2_3_4	1	1	0	0.0457	4	Cluster9_var1
Salmonella_1	0	0	0	0.0000	70	Cluster18_var3
Salmonella_2	0	0	0	0.0000	130	Cluster18_var1
Salmonella_3	1	0	0	0.0221	48	Cluster7_var1
Salmonella_5	1	0	0	0.0221	49	Cluster7_var2
Salmonella_6	0	0	0	0.0000	62	Cluster18_var2
Salmonella_7	1	0	0	0.0216	40	Cluster0_var1
Staphylococcus_1	0	0	0	0.0000	428	Cluster10_var1
Staphylococcus_1	0	0	1	0.0224	3	Cluster10_var2
Staphylococcus_1	0	1	0	0.0224	3	Cluster10_var3
Staphylococcus_2	0	0	0	0.0000	328	Cluster3_var1
Staphylococcus_2	0	1	2	0.0665	4	Cluster3_var2
Staphylococcus_2	0	2	0	0.0443	4	Cluster3_var3
Staphylococcus_3	0	0	0	0.0000	354	Cluster1_var1
Staphylococcus_3	0	0	1	0.0221	6	Cluster1_var3
Staphylococcus_3	0	1	0	0.0221	4	Cluster1_var2
Staphylococcus_4	0	0	0	0.0000	465	Cluster17_var1
Staphylococcus_4	0	0	1	0.0228	6	Cluster17_var3
Staphylococcus_4	0	1	0	0.0228	6	Cluster17_var2
Staphylococcus_5	0	0	0	0.0000	425	Cluster14_var1
Staphylococcus_5	0	0	1	0.0226	10	Cluster14_var3
Staphylococcus_5	0	1	0	0.0226	5	Cluster14_var2
Staphylococcus_6	0	0	0	0.0000	325	Cluster8_var1
Staphylococcus_6	0	2	0	0.0446	4	Cluster8_var2

Table S4: Overview of high-accuracy long amplicon sequencing from the literature.

Sequencing Platform	Average length of sequences	Yield (Mbp)	Error Rate of Consensus Sequences (%)	Error Rate of Clustered Consensus Sequences (%)	Reference
PacBio	1,460 [a]	90 [b]	0.21 [c]	0.027 [d]	(Schloss et al., 2016)[2]
	1,400 [e]	16 [e]	-	0.50	(Singer et al., 2016)[3]
	1,500 [a]	16	-	0.0073 [d]	(Wagner et al., 2016)[4]
	5,000	1,170	1.1	-	(Volden et al., 2018)[5]
	13,500	89,000	0.20	-	(Wenger et al., 2019)[6]
	1500	117	0.04	-	(Callahan et al., 2019)[7]
Nanopore	1,386 [f]	7.8 [g]	2.0 [h]	0.50 [i]	(Calus et al., 2018)[8]
	5,000	2,180	6.0	-	(Volden et al., 2018)[5]
Illumina	1,530	16	0.17	-	(Karst et al., 2018)[9]
	6,000	17.3	0.04	-	(Stapleton et al., 2016)[10]

[a] Actual length was not provided for V1-V9 amplicon after filtering. This value is based on expected amplicon length.
[b] Based on 61,721 sequences, and 51.33% of sequences remaining after filtering.
[c] Sequence accuracy following *de-novo* clustering
[d] Consensus sequence accuracy after pre-clustering sequences at 99% similarity.
[e] Based on mock community dataset
[f] With $1D^2$ sequencing
[g] Based on 5,622 reads passing filters
[h] Sequence accuracy following *de-novo* correction and size selection
[i] Consensus sequence accuracy after clustering into OTUs at 97% similarity with nanoclust algorithm

Table S5: Overview of metagenome read coverage stats for all species. The mean coverage, standard deviation (sd), maximum coverage, minimum coverage, and the relative standard deviation (sd_pct) were all based on whole-genome Nanopore sequence data (available from: https://github.com/LomanLab/mockcommunity).

mean	sd	max	min	sd_pct	Species
459.21	41.13	568.09	359.38	8.96	Bacillus
424.87	61.7	559.43	306.71	14.52	Enterococcus
233.29	11.53	271.17	204.58	4.94	Escherichia
454.63	58.39	599.36	330.46	12.84	Lactobacillus
471.7	27.74	530.09	401.15	5.88	Listeria
149.35	10.07	181.35	119.92	6.74	Pseudomonas
225.14	13.72	275.94	180.99	6.09	Salmonella
425.34	25.31	489.68	364.85	5.95	Staphylococcus

Table S6: Overview of different estimates of relative abundance in mock community. ‘Theoretical 16S relative abundance’ is the abundance provided by the vendor. ‘Theoretical 16S relative abundance based on metagenome’ is estimated based on the Nanopore metagenome data and the number of rRNA operons per genome. ‘Observed 16S relative abundance UMI consensus data’ is based on the UMI consensus data. GC content is estimated from the genome and rRNA operon reference sequences.

Species	Theoretical 16S relative abundance (%)	Theoretical 16S relative abundance based on metagenome (%)	Observed 16S relative abundance UMI consensus data (%)	Genome mean GC (%)	rRNA operon mean GC(%)
Bacillus	17.4	25.6	19.9	44	47
Enterococcus	9.9	9.5	5.1	38	49
Escherichia	10.1	9.1	8.8	47	47
Lactobacillus	18.4	13.4	15.7	52	49
Listeria	14.1	15.8	11.8	38	49
Pseudomonas	4.2	3.3	5.5	66	48
Salmonella	10.4	8.8	3.8	52	47
Staphylococcus	15.5	14.5	29.5	33	52

Table S7: Estimation of mismatches between primers and rRNA operon sequences.

Species	5' primer hit	5' primer error (n)	3' primer hit	3' primer error (n)
Bacillus_1	27f	0	2490r	0
Bacillus_10	27f	0	2490r	0
Bacillus_2	27f	0	2490r	0
Bacillus_3	27f	0	2490r	0
Bacillus_4_5	27f	0	2490r	0
Bacillus_6	27f	0	2490r	0
Bacillus_7	27f	0	2490r	0
Bacillus_8	27f	0	2490r	0
Bacillus_9	27f	0	2490r	0
Enterococcus_1	27f	0	2490r	0
Enterococcus_2	27f	0	2490r	0
Enterococcus_3	27f	0	2490r	0
Enterococcus_4	27f	0	2490r	0
Escherichia_1	27f	0	2490r	0
Escherichia_2	27f	0	2490r	0
Escherichia_3	27f	0	2490r	0
Escherichia_4	27f	0	2490r	0
Escherichia_5	27f	0	2490r	0
Escherichia_6	27f	0	2490r	0
Escherichia_7	27f	0	2490r	0
Lactobacillus_1	27f	0	2490r	0
Lactobacillus_2	27f	0	2490r	0
Lactobacillus_3	27f	0	2490r	0
Lactobacillus_4	27f	0	2490r	0
Lactobacillus_5	27f	0	2490r	0
Listeria_1	27f	0	2490r	0
Listeria_2_5	27f	0	2490r	0
Listeria_3	27f	0	2490r	0
Listeria_4	27f	0	2490r	0
Listeria_6	27f	0	2490r	0
Pseudomonas_1_2_3_4	27f	0	2490r	0
Salmonella_1	27f	0	2490r	0
Salmonella_2	27f	0	2490r	0
Salmonella_3	27f	0	2490r	0
Salmonella_4	27f	0	2490r	0
Salmonella_5	27f	0	2490r	0
Salmonella_6	27f	0	2490r	0
Salmonella_7	27f	0	2490r	0
Staphylococcus_1	27f	0	2490r	0
Staphylococcus_2	27f	0	2490r	0
Staphylococcus_3	27f	0	2490r	0
Staphylococcus_4	27f	0	2490r	0
Staphylococcus_5	27f	0	2490r	0
Staphylococcus_6	27f	0	2490r	0

Table S8: Data overview

Data type	ENA Accession
ENA study	PRJEB32674
Raw Nanopore reads (fastq)	ERR3336963
Raw Nanopore reads (fast5)	ERR3336964
UMI consensus sequences (fasta)	ERZ940787
Variants consensus sequences (fasta)	ERZ940796

Supplementary references

1. Edgar, R. C. UCHIME2 : improved chimera prediction for amplicon sequencing. (2016).

2. Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L. & Highlander, S. K. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. 1–16 (2016). doi:10.7717/peerj.1869

3. Singer, E. *et al.* High-resolution phylogenetic microbial community profiling. *ISME J.* **10**, 2020–2032 (2016).

4. Wagner, J. *et al.* Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene

classification. *BMC Microbiol.* **16**, 1–17 (2016).

5. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci.* **115**, 9726–9731 (2018).

6. Wenger, A. M. *et al.* Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *bioRxiv* 519025 (2019). doi:10.1101/519025

7. Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *bioRxiv* 392332 (2018). doi:10.1101/392332

8. Calus, S. T., Ijaz, U. Z. & Pinto, A. J. NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience* **7**, 1–16 (2018).

9. Karst, S. M. *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **36**, 190–195 (2018).

10. Stapleton, J. A. *et al.* Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing. 1–20 (2016). doi:10.5061/dryad.kr8kk