HHS Public Access

Author manuscript

Mol Ecol Resour. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Mol Ecol Resour. 2022 July ; 22(5): 1786–1802. doi:10.1111/1755-0998.13588.

High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes

Florian Trigodet[1,‡], **Karen Lolans**[1,‡], **Emily Fogarty**[2], **Alon Shaiber**[3], **Hilary G. Morrison**[4], **Luis Barreiro**[1], **Bana Jabri**[1], **A. Murat Eren**[1,2,3,4,*]

[1]Department of Medicine, The University of Chicago, Chicago, IL 60637, USA

[2]Committee on Microbiology, University of Chicago, Chicago, IL 60637, USA

[3]BioPhysical Sciences Program, The University of Chicago, Chicago, IL 60637, USA

[4]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

Abstract

By offering extremely long contiguous characterization of individual DNA molecules, rapidly emerging long-read sequencing strategies offer comprehensive insights into the organization of genetic information in genomes and metagenomes. However, successful long-read sequencing experiments demand high concentrations of highly purified DNA of high molecular weight (HMW), which limits the utility of established DNA extraction kits designed for short-read sequencing. Challenges associated with input DNA quality intensify further when working with complex environmental samples of low microbial biomass, which requires new protocols that are tailored to study metagenomes with long-read sequencing. Here, we use human tongue scrapings to benchmark six HMW DNA extraction strategies that are based on commercially available kits, phenol-chloroform (PC) extraction, and agarose encasement followed by agarase digestion. A typical end goal of HMW DNA extractions is to obtain the longest possible reads during sequencing, which is often achieved by PC extractions as demonstrated in sequencing of cultured cells. Yet our analyses that consider overall read-size distribution, assembly performance, and the number of circularized elements found in sequencing results suggest that column based kit with enzyme supplementation, rather than PC methods, may be more appropriate for long-read sequencing of metagenomes.

[*]Corresponding Author: A. Murat Eren[1,2,3,4], The University of Chicago, Chicago, IL 60637, USA, meren@uchicago.edu.
[‡]Equal contribution

Author contributions

FT, KL, and AME conceived the study. KL developed and implemented long-read sequencing protocols. FT performed analyses of short- and long-read sequencing data. KL, EF, and AS helped with data analysis. HGM performed short-read sequencing. HGM, LB, BJ, and AME supervised research. FT, KL, and AME wrote the paper with critical input from all authors.

Data Accessibility and Benefit-Sharing

Data Accessibility Statement

Raw sequences for long- and short-read sequencing data for shotgun metagenomes and 16S rRNA gene amplicons are available under the BioProject PRJNA703035. We also made available FASTA files and anvi'o contigs databases for (1) assembled and polished long-read sequences at doi:10.6084/m9.figshare.14141228, (2) unassembled long-read sequences at doi:10.6084/m9.figshare.14141414, and (3) assembled shotgun metagenomes at doi:10.6084/m9.figshare.14141819. Supplementary Tables and Figures are also available via doi:10.6084/m9.figshare.14141918. A reproducible bioinformatics workflow is available at the URL https://merenlab.org/data/hmw-dna-extraction-strategies/, where the URL https://github.com/merenlab/web/tree/master/data/hmw-dna-extraction-strategies and the commit hash 557f6e37381d212a22a17d67c7b0e774e5abee82 gives access to its final version at the time of publication.

Keywords

metagenomics; high-molecular-weight DNA; long-read sequencing; nanopore

Introduction

High-throughput sequencing of metagenomes offers unprecedented insights into the diversity and gene pool of naturally occurring microbes and viruses that occupy soils (Nesme et al., 2016), marine habitats (Gregory, Zayed, et al., 2019; Sunagawa et al., 2015) and host-associated environments (Gregory, Zablocki, et al., 2019; Human Microbiome Project Consortium, 2012). The high accuracy and the high throughput of the modern sequencing platforms are afforded by read lengths that typically remain below 250 bases. These relatively short reads pose significant constraints on the data utility, especially in metagenomics (Wommack et al., 2008).

By stitching together the short reads that partially overlap, metagenomic assembly can reconstruct orders of magnitude longer contiguous segments of input DNA (Nurk et al., 2017) and enable the recovery of microbial genomes from metagenomes (Tyson et al., 2004). In recent years, this strategy has become a primary tool in microbiology to study the ecology and evolution of naturally occurring microbial populations (Al-Shayeb et al., 2020; Delmont et al., 2018; Hug et al., 2016; Pasolli et al., 2019; Spang et al., 2015). But metagenomic assembly is inherently a challenging task (Ayling et al., 2020) and the assembly of complex environments often leads to highly fragmented assemblies (Olson et al., 2019). These fragmented assemblies increase the likelihood of generating composite genomes that include contigs from multiple distinct populations (Chen et al., 2020), which risk erroneous insights into microbial ecology and evolution (Chen et al., 2020; Shaiber & Eren, 2019).

By circumventing the problems associated with short read assembly, long-read sequencing offers a compelling solution to the ideal of reconstructing complete genomes from metagenomes (Driscoll et al., 2017; White et al., 2016). Nanopore sequencing, which resolves the identity and order of nucleotides based on changes in ionic current as a single-stranded RNA or DNA molecule pass through a tiny pore (Kasianowicz et al., 1996), has rapidly gained popularity among researchers (Wang et al., 2014) thanks to its availability through affordable and easy to operate sequencing devices, such as MinION by Oxford Nanopore Technologies (“The Long View on Sequencing,” 2018). Despite the high error rates and relatively lower sequencing depth, long reads from nanopore sequencing of metagenomes led to key insights from challenging systems (Pessi et al., 2020; Reveillaud et al., 2019) and enabled the recovery of circular, complete genomes from metagenomes (Cusco et al., 2020; Moss et al., 2020; Nicholls et al., 2019; Sanderson et al., 2018; Singleton et al., 2021; Somerville et al., 2019).

The efficacy of long-read sequencing heavily depends on the structural integrity of the input DNA (Schalamun et al., 2019), which poses a new and significant challenge. Commercial DNA extraction kits that emerged during the era of short-read sequencing typically include steps that physically disrupt cells through mechanical lysis and generate highly fragmented

DNA molecules. While these commercial kits improve short-read sequencing outcomes as they ensure maximum yield and coverage of DNA in a sample, they set a critical limit to the outcomes of long-read sequencing. Hence, establishing DNA extraction strategies that afford (1) preservation of high molecular weight (HMW) molecules, (2) high degree of sample purity, and (3) increased overall DNA yields have become critical considerations for the successful application of long-read sequencing.

Phenol-chloroform DNA extractions, first popularized back in 1989 (Sambrook et al., 1989), have been making a resurgence as a 'go-to' method for extracting HMW DNA (Maghini et al., 2020; Quick & Loman, 2018). While recent studies have used this approach to recover ultra-long DNA fragments (*e.g.*, >100 kbp) from cultured organisms (Cicha et al., 2020; Hosoe et al., 2020; Kinoshita et al., 2020; Takeshita et al., 2020; Tippelt et al., 2020), the utility of phenol-chloroform extractions for metagenomics is not yet clear. In parallel, long-read sequencing surveys of metagenomes have largely focused on high microbial biomass samples including human stool (Moss et al., 2020) or activated sludge (Singleton et al., 2021), and best practices to study metagenomes of lower biomass samples are yet to emerge. Increasing the breadth of long-read sequencing requires DNA extraction protocols that can both produce long reads and can scale to a range of systems, including those that are associated with low microbial biomass.

Here we designed six DNA extraction protocols and examined their relative effectiveness to extract HMW DNA in terms of total yields, concentration, purity, integrity and applicability for subsequent long-read sequencing. Using an Oxford Nanopore MinION sequencer, we benchmarked our protocols based on the extent of host contamination, fragment size distribution, microbial taxonomy, and metagenomic assembly outcomes across multiple sequencing runs, and compared taxonomic insights that emerged from long-read sequencing to those from Illumina short-read sequencing. We chose the human oral cavity as a model system, since it offers a challenging environment to benchmark DNA extraction strategies as it is home to complex microbial communities (Dewhirst et al., 2010) with relatively low biomass (Duran-Pinedo & Frias-Lopez, 2015) and is typically mixed with eukaryotic host DNA (Marotz et al., 2018).

Materials and Methods

Tongue Dorsum Sample Collection.

A single healthy individual self-collected scraping of their tongue dorsum on 13 separate occasions (1 per day) over the course of 3 weeks. We used BreathRx Gentle Tongue Scraper (Philips Sonicare) for sample collection which was performed prior to eating, drinking or performing oral hygiene. Starting as far back as possible on the tongue, the scraper was passed forward over the entire surface three sequential times. We transferred the collected material to 520 μl of PBS (phosphate-buffered saline) and immediately stored at −20°C until processing. The pooled 13 tongue dorsum samples were used for high molecular weight DNA extraction comparison, while another self-collected tongue dorsum sample (named TD) from this same individual was obtained 2 weeks later for short read sequencing (amplicon and metagenomics).

DNA extraction methods.

We compared six DNA extraction methods tailored for high molecular weight DNA recovery. We included both commercially available kits and non-kit methods used in the published literature which incorporated different combinations of cell lysis mechanisms and DNA purification methodologies. To facilitate direct comparison between all extraction methods, we thawed and pooled the 13 samples immediately prior to DNA extraction. After homogenizing the pooled sample by vortexing for 15-sec, we used a 500 µl aliquot as the starting material for each extraction method, which was performed in duplicate. We resuspended the isolated genomic DNA from each method in a final 100 µl volume in 1.5 ml LoBind microfuge tube (Eppendorf, Hauppauge, NY). We sought to maximize read lengths by implementing best practices for handling HMW DNA throughout all the methods. We eliminated vortexing and mixing by pipetting, when possible, in favor of end-over-end tube rotation to minimize velocity gradients. We used wide-bore pipet tips with gentle pipetting to reduce DNA breakage and avoided unnecessary freeze-thaw cycles by storing DNA at 4ºC until sequence analysis.

DNeasy PowerSoil Isolation Kit with modified bead beating (PB).

DNeasy PowerSoil® DNA Isolation kits (Qiagen) are commonly used in metagenomics (Human Microbiome Project Consortium, 2012; Shaiber et al., 2020) to extract high quality DNA from environmental matrices. We sought to determine its compatibility with nanopore sequencing protocols by amending the PowerSoil® DNA isolation protocol in two ways. In our first modification to the DNeasy PowerSoil® DNA Isolation kit, we incorporated a modified bead beating step (Edwards et al 2019), as a way to minimize velocity gradients and reduce DNA shearing, therefore improving fragment length. We transferred 500 µl of the pooled sample to the kit provided PowerBead tube, which we inserted flat into IKA Works Inc MS2S8 Minishaker for Bioanalyzer DNA chips. Samples were agitated for 10 minutes (in 1 min pulse increments) at 2400 rpm (Edwards et al., 2019). We then followed the remainder of the manufacturer's instructions for DNA isolation and purification.

DNeasy PowerSoil Isolation Kit supplemented with an enzymatic treatment (PE).

In the second modification to the DNeasy PowerSoil® DNA Isolation kit, we fully replaced the use of mechanical cell lysis (or bead beating) with a heated, enzymatic treatment step (Yuan et al., 2012). We added 500 µl of the pooled sample to the kits' PowerBeads tubes including the tube's solution (but lacking the beads). We added a lytic cocktail to facilitate cell lysis: 125 µl lysozyme (10 mg/mL, Sigma-Aldrich, St. Louis, MO), 37.5 µl mutanolysin (10 KU/mL Sigma-Aldrich), 7.5 µl lysostaphin (4000 U/mL, Sigma-Aldrich) and 5 µl RNase A (10mg/mL, Sigma-Aldrich). We incubated for 1 hour at 37ºC. Then we added 50 µl of proteinase K (20 mg/mL, Sigma-Aldrich) alongside kit solution C1, and followed by incubation for 30 min at 56ºC. After centrifugation of the tubes at 10,000 × g for 30 seconds at room temperature (step 6), we continued the rest of the isolation protocol as described by the manufacturer.

DNeasy UltraClean Microbial Kit (UC).

We extracted genomic DNA using the DNeasy UltraClean Microbial Kit (Qiagen) and replaced its bead beating step with the manufacturer's alternative lysis procedure to reduce shearing of the DNA. This commercial kit was evaluated as a result of a direct recommendation by Oxford Nanopore. We added 500 µl of pooled sample to 300 µl of PowerBead solution. After the addition of 50 µl of Solution SL, we incubated the sample for 10 min at 65°C. After centrifuging the tubes at 10,000 × g for 30 seconds at room temperature (step 5), we continued the rest of the isolation protocol as described by the manufacturer.

Qiagen Genomic Tip 20/G supplemented with an enzymatic treatment (GT).

We extracted DNA using the Qiagen Genomic Tip 20/G (Qiagen) and followed the manufacturer's protocol "Preparation Gram-negative and some Gram-positive Bacterial Samples". We centrifuged 500 µl of pooled sample for 10 min at 10,000 × g and resuspended the pellet in Buffer B1 supplemented with 20 µl DNAse-free RNAse (10 mg/mL, Sigma-Aldrich), 45 µl Proteinase K (20 mg/mL, Sigma-Aldrich) and 20 µl lysozyme (100 mg/mL, Sigma-Aldrich), as outlined. We modified the lytic cocktail to include 9 µl lysostaphin (4000 U/mL, Sigma-Aldrich) and 45 µl mutanolysin (10 KU/mL Sigma-Aldrich) in order to improve the lysis potential in Gram-positive bacteria. After incubation for 2 hours at 37°C, we added Buffer B2 and extended the incubation at 50°C to 90 min. As the lysate had not cleared after this initial period, we extended the incubation for an additional two hours. We removed any remaining particulate matter by centrifugation at 5,000 × g for 10 min as recommended by the manufacturer. Finally, we used the columns in the kit according to the described isolation protocol.

Phenol/Chloroform Extraction (PC).

We extracted DNA using a phenol/chloroform extraction protocol modified from Chapter 6, protocol 1 of Sambrook and Russell (Sambrook, Jain). For SDS cell lysis, we added 500 µl of pooled sample to 10 mL of TLB (10 mM Tris-HCL pH 8.0, 25 mM EDTA pH 8.0, 100 mM NaCl, 0.5% (w/v) SDS, 20 µg/mL RNase A) and vortexed at full speed for 5 sec followed by an incubation at 37°C for one hour. Proteinase K (Qiagen) was added to a final concentration of 200 µg/mL and we mixed the sample by slow inversion three times, followed by two hours at 50°C with gentle mixing every 30 min. We purified the lysate with 10 mL buffer saturated phenol using phase-lock gel falcon tubes, followed by phenol:chloroform-isoamyl alcohol (1:1). We precipitated the DNA by adding 4 mL of 5 M ammonium acetate and 30 mL ice-cold ethanol. We recovered DNA by one of the following methods: a glass hook followed by washing twice in 70% ethanol (if a DNA mass was visible) or centrifugation at 4500 × g for 10 min followed by washing twice in 70% ethanol (if no DNA mass was visible). After spinning down at 10,000 × g, we removed ethanol by drying at ambient temperature for 10 min. We added 100 µl EB (elution buffer, 10 mM Tris-HCL, pH 8.5) to the DNA and left at 4°C overnight to resuspend the pellet.

Agarose Plug Encasement & Extraction (AE).

From the 500-ul pooled sample aliquot, we pelleted the cells by centrifugation at 10,000 × g for 10 min and removed the supernatant. We performed DNA extraction on the pellet using a pulsed-field gel electrophoresis (PFGE)-based agarose encasement extraction protocol (Matushek et al., 1996). The pelleted cells were resuspended in a 300 µl 2X lysis solution (12.5 mM Tris-HCl pH 7.6, 2 M NaCl, 20 mM EDTA pH 8.0, 1% (w/v) Brij 58, 1% (w/v) deoxycholate, 1% (w/v) sodium lauroyl sarcosine) to which we added lysozyme,1 mg/mL and DNAse-free RNAse, 30 µg/mL on the day of the experiment. We combined the entire suspension with 300 µL of molten 1.6% low-melting point (LMP) agarose (TopVision, Thermo Fisher Scientific, Waltham, MA) which we pipetted into a plug mold and allowed to solidify. We placed each plug into 3 mL of 1X lysis solution containing lysozyme, 0.5 mg/mL; DNAse-free RNAse,100 µg/mL; lysostaphin, 50 U/mL; and mutanolysin, 0.3 KU/mL. We added the enzymes on the day of the experiment and incubated the plugs overnight at 37°C with gentle shaking. After incubation, we replaced the lysis solution sequentially with ESP (10 mM Tris HCl pH 7.6, 1 mM EDTA) to which we added proteinase K (Sigma) at a final concentration of 100 µg/ml and 1% sodium dodecyl sulfate at 50°C for one hour; and then two rounds of washing in sterile, dilute TE (10 mM Tris HCl pH 7.6, 0.1 mM EDTA) first at 50°C for one hour and then at 35°C for 30 min with gentle shaking. We transferred the plug to 5 ml of fresh dilute TE in a clean tube for storage until we performed a ß-agarase I digestion of the agarose and isopropanol DNA precipitation according to the manufacturer's instructions (New England Biolabs, Ipswich, MA). We resuspended the precipitated DNA in 100 µl EB (10mM Tris-HCL, pH 8.5).

Determination of DNA yield, purity metrics and fragment size distribution.

For each extraction, we quantified the DNA yield on a Qubit® 1.0 Fluorometer (Thermo Scientific), using the dsDNA HS (high sensitivity, 0.2 to 100 ng) Assay kit according to the manufacturer's protocols; a sample volume of 1 µl was added to 199 µl of a Qubit working solution. We assessed the purity of the extracted nucleic acids with the A260/280 and A260/230 absorbance ratios obtained using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). We assessed the DNA fragment size distribution by electrophoresis (90V for 1.5 h) of genomic DNA (22 ng or 44 ng, as available) on a 0.8% (wt/vol) agarose gel followed by staining with ethidium bromide and UV light visualization. We used λ-HindIII DNA size standards to estimate fragment sizes.

MinION Library Preparation and Sequencing.

Upon receipt, and again immediately prior to sequencing, we measured the flowcell pore counts using the Platform QC script (MinKNOW). We stored the flowcells in their original packaging, which we resealed with parafilm and tape, at 4°C until use. Unless stated otherwise, we omitted size selection and the optional shearing step to allow us to evaluate the full distribution of fragment sizes produced by each method. We performed the library preparation using the Ligation Sequencing Kit (SQK-LSK108) and Native Barcoding Kit (EXP-NBD103) for genomic DNA, according to the standard 1D Native barcoding protocol provided by the manufacturer (Oxford Nanopore) (version: NBE_9006_v103_rev2_21DEC16) unless indicated. We carried out all reactions at room

temperature. We followed the input DNA mass recommendation of 1.0 μg gDNA and performed DNA repair (NEBNext FFPE DNA Repair Mix, NEB M6630) to maximize read length. For the 1.0× AMPure clean-up step, we used gentle rotation and an extended elution time (minimum 20 min) to assist in the release of long DNA fragments from the beads. We quantified 1 μl aliquots by fluorometry (Qubit) after each phase of the library preparation (*i.e.* damage repair, end prep, barcoding, pooling and adaptor ligation) to quantify DNA recovery, and identified substantial DNA loss after each step (Table. Quality metrics).

We used two sequencing runs (3 samples multiplexed on each flowcell) for the sequencing. We loaded the completed libraries onto R9.4 flowcells as per instructions from ONT and scheduled sequencing runs for 48 hours. Sequencing continued until time expired or until pore exhaustion (defined by <10 functional pores).

BluePippin Size Selection.

To quantify the magnitude of its impact, if incorporated into a long-read sequencing workflow, we performed size selection using DNA from GT_1 and GT_2 with the BluePippin (Sage Science) system with 0.75% dye-free agarose cassettes and marker S1. We selected fragments >6 kB in high-pass collection mode (an approach enabling the collection of DNA fragments above a user-defined size). Our 3.3 μg sample DNA input was less than the recommended 5 μg; however, this input was maximized given sample DNA concentrations and loading volume constraints.

Long-read Sequence analysis.

We uploaded the raw MinION FAST5 files produced with the MinKNOW software (versions 1.15.4 through 3.1.19) to our cluster to perform the base-calling and demultiplexing with guppy v2.3.1. FastQ files were generated only for reads meeting a minimum quality threshold (quality score of 7). We used minimap2 v2.14 (Li, 2018) and samtools v1.9 (Li et al., 2009) to remove sequences that mapped to the human genome build 38 (GRCh38) from NCBI and estimate the amount of host contamination. We used anvi'o v6.2 and the contig snakemake workflow to compute the sequence metrics (Köster & Rahmann, 2012). Briefly, the workflow created a contigs database with 'anvi-gen-contigs-database', which used Prodigal v2.6.3 (Hyatt et al., 2010) with the metagenome mode to identify open reading frames. It used 'anvi-run-hmm' to detect the single-copy core genes from bacteria (n=71, modified from (Lee, 2019)), archaea (n=76, (Lee, 2019)), eukarya (n=83, http://merenlab.org/delmont-euk-scgs), and ribosomal RNAs (n=12, modified from https://github.com/tseemann/barrnap). We used 'anvi-display-contigs-stats' to get the number of sequences, total length, N50, longest sequence, number of genes, number of single-copy core genes and ribosomal genes. We used BLAST against the NCBI's nr/nt database to get the best taxonomy, percent identity and query alignment for each longest read per extraction method. We used 'anvi-get-sequences-for-hmm-hits' to recover 16S rRNA genes for each condition and used the Human Oral Microbiome Database (HOMD) online blast tool for taxonomic assignment. We assembled the long reads with Flye and the metagenomic option (Kolmogorov et al., 2019), which takes into account the uneven coverage nature of metagenomes. We used short-reads generated using HMW DNA extraction sample PB_2 (see method below) to polish the assemblies using Pilon (Walker et

al., 2014). We created anvi'o contigs databases, as described above, with the polished Flye's contigs to summarize the assembly metrics.

16S rRNA gene amplicon DNA extraction, library preparation, sequencing and analysis.

For the 16S rRNA gene amplicon sequencing, we used a sample (TD), which was collected from the same individual two weeks after the pooled samples that were used for HMW DNA extraction. We performed sample DNA extraction using the DNeasy Powersoil kit (Qiagen) following the manufacturer's protocol. We amplified the V4-V5 hypervariable regions of the bacterial SSU rRNA gene using degenerate primers. 518F (CCAGCAGCYGCGGTAAN) and 926R (CCGTCAATTCNTTTRAGT, CCGTCAATTTCTTTGAGT, and CCGTCTATTCCTTTGANT). Amplification was done with fusion primers containing the 16S-only sequences fused to Illumina adapters. The forward primers included a 5 nt multiplexing barcode and the reverse a 6 nt index. We generated PCR amplicons in triplicate 33 uL reaction volumes with an amplification cocktail containing 0.67 U SuperFi Taq Polymerase (Invitrogen, Carlsbad, CA), 1X enzyme buffer (includes $MgCl_2$), 200 uM dNTP mix (ThermoFisher), and 0.3 uM of each primer. We added approximately 10–25 ng template DNA to each PCR and ran a no-template control for each primer pair. Amplification conditions were: initial 94C, 3 minute denaturation step; 30 cycles of 94C for 30s, 57C for 45s, and 72C for 60s; final 2 minute extension at 72C. The triplicate PCR reactions were pooled after amplification, visualized with the negative controls on a Caliper LabChipGX, and purified using Ampure followed by PicoGreen quantitation and Ampure size selection. Libraries were sequenced on an Illumina Miseq 250-cycle paired-end run. We used illumina-utils v2.7 (Eren et al. 2013) for the quality filtering, following (Minoche et al., 2011) recommendations, and the merging of the paired-end reads. We used Vsearch to remove chimeric sequences (Rognes et al., 2016) and Minimum Entropy Decomposition (MED) (Eren et al., 2015) to cluster the merged reads into oligotypes. We assigned taxonomy using DADA2 (Callahan et al., 2016) and the Silva v132 non-redundant database (Quast et al., 2013).

Short-read metagenomic library preparation, sequencing and analysis.

For the short-read metagenomic sequencing, we used both the sample TD collected two week after the initial sampling, and the HMW DNA extraction sample PB_2 (the closest methodology to gold standard short-read sequencing extraction methodology). Sample DNA concentrations, determined by PicoGreen assay, were 67 ng/ul (TD) and 0.75 ng/ul (PB_2). We used 100 ng and 28 ng, respectively, for library construction. DNA was sheared to ~400 bp using the Covaris S2 acoustic platform and libraries were constructed using the Nugen Ovation Ultralow kit. Each required an amplification step: 7 cycles (TD) or 11 cycles (PB_2). The products were visualized on an Agilent Tapestation 4200 and size-selected to an average of 482 bp using BluePippin (Sage Biosciences). The final library pool was quantified with the Kapa Biosystems qPCR protocol and sequenced on the Illumina NextSeq500 in a 2 × 150 paired-end sequencing run using dedicated read indexing. We used anvi'o v6.2 and the metagenomics snakemake workflow for the assembly and analysis of the short-reads. Briefly, the workflow uses illumina-utils for quality filtering followed by a metagenomics assembly with IDBA-UD v1.1.3 (Peng et al., 2012) and generates contigs databases as described in the long-read sequence analysis section. The workflow also uses

Bowtie2 v2.3.5.1 (Langmead & Salzberg, 2012) to map short-read on the assembled contigs and samtools (Li et al., 2009) to sort, index and convert sam files into bam files used by anvi'o to generate profiles databases. To compute taxonomic profiles of metagenomes we used the anvi'o program 'anvi-estimate-scg-taxonomy', which aligns ribosomal proteins found in a metagenomic assembly to those that are found in reference genomes from the Genome Taxonomy Database (GTDB) (Parks et al., 2020).

Visualization.

We generated the figures in R with ggplot2 (Wickham, 2009) and modified them with inkscape.

Results

The DNA extraction protocols we benchmarked here include four protocols based on three commercially available Qiagen DNA extraction kits (each incorporated modifications to their cell lysis procedures): the Qiagen DNeasy PowerSoil with a modified bead beating step (PB) or with bead beating replaced by enzymatic cell lysis (PE); the Qiagen DNeasy UltraClean Microbial Kit (UC) using the manufacturer's alternative lysis procedure to reduce DNA shearing; and the Qiagen Genomic Tip 20/G extraction kit (GT) augmented with additional enzymatic cell lysis. The remaining two protocols included in our study are a phenol-chloroform protocol (PC), and a pulsed-field gel electrophoresis (PFGE)-based agarose encasement extraction protocol (AE) followed by agarase digestion. Throughout the text we refer to these protocols as PB, PE, UC, GT, PC and AE, and we denote technical replicates as "XX_1" or "XX_2". To benchmark the protocols that are detailed in the Methods section, we used a tongue dorsum sample pooled together from 13 samples collected from the same individual. An additional sample (TD) collected from the same individual has been used to compare taxonomic composition between short and long-read sequencing.

Yield and quality metrics of isolated DNA vary between protocols.

To ascertain their suitability for long-read sequence analysis, we first analyzed the quantity and quality of DNA isolated from each extraction protocol using both fluorometric (Qubit) and spectrophotometric (Nanodrop) methods (Table 1). The fluorescent dye used by Qubit binds specifically to its target molecule, dsDNA, and provides the most accurate DNA quantitation. DNA concentrations were comparable between technical replicates for all methods, and three extraction protocols (GT, PC and AE) were distinguished from the others with concentrations >75 μg/ml. These concentrations were comparable to previously published HMW DNA extractions from oral samples (Yahara et al., 2021). GT exhibited the highest sample concentrations with a mean of 110 μg/ml. In comparison, the mean concentrations from PB, PE and UC were far less at 2.03, 4.15 and 33.45 μg/ml, respectively. The MinION sequencing protocol for the Ligation Sequencing Kit (SQK-LSK108) advises using 1 – 1.5 μg DNA. As we resuspended DNA into a final volume of 100 μl, only UC, GT, PC and AE had sufficient DNA yield to meet the recommended input.

We then considered the absorption spectra to assess sample purity and identify potential non-nucleic acid contamination. The A260/A280 ratio indicates DNA purity with expected values around 1.8 for pure DNA. All extraction methods fell within the desired range (1.74 – 1.97), except for both PB replicates and one replicate of PE; however, samples with concentrations approaching the lower limit of 2 μg/mL may result in unacceptable 260/280 ratios. The A260/A230 ratio signifies possible residual chemical contamination such as EDTA, phenol, guanidine salts (often used in column-based kits), or carbohydrate carryover. UC, GT and PC had ratios close to the expected value of 2. Overall, GT had the best combined yield and purity metrics; PC and AE were deemed suitable alternatives. The purity metrics and congruence of Qubit and Nanodrop concentrations in UC were also desirable, yet UC's ~2–3-fold lower DNA yield proved less appealing.

We ran an agarose gel electrophoresis to visually assess the crude DNA fragment size distribution (Figure 1). All DNA samples migrated predominantly as a single high-molecular-weight DNA band that aligned with (or was larger) than the reference 23.1 kbp fragment of HindIII-digested lambda DNA. A light smear, visible to 2.0 kb, was observed in GT indicating the presence of smaller fragments. Recovery of larger DNA fragments by PC was denoted by the predominant band running higher than the 23.1 kbp fragment.

Low DNA yield can result in sample loss during library preparation or low pore occupancy on MinION during sequencing. While working with samples of low DNA yield is inevitable, spiking in known DNA (such as lambda DNA) to 'pad' samples with low DNA yields may be used to start sequencing as we demonstrated before (Reveillaud et al., 2019). But to minimize the need for additional DNA to 'pad' samples with low DNA yield due to the extraction protocol, we eliminated protocols PB, PE and UC from any further evaluation as they consistently resulted in mediocre DNA yield.

Next, we sequenced the DNA from GT, PC, and AE using two MinION flow cells to ensure the replicates of the same protocol were sequenced on different runs (Run 1: GT_1, PC_1, and AE_1; Run 2: GT_2, PC_2, and AE_2). The sequencing runs generated 4.84 and 7.79 Gbp, respectively, which were within the expected range of MinION sequencing output (Cusco et al., 2020; Moss et al., 2020). The increase in the output in the second run could be attributed to less sample loss during library preparation and subsequent input of 3X more DNA (142.8 ng versus 466.2 ng) into the flow cell. After performing a quality filtering step (using a minimum Q-score of 7), the percentage of reads passing the quality check (Pass_Reads) was similar between extraction methods within each flow cell (96% and 93–94% respectively, Table S1.a). However, a comparison between runs demonstrated that a higher percentage of sequences were removed (Fail_Reads) in the second run (Table S1.a). The total number of nucleotides was comparable between GT and PC within sequencing runs (1,699,213,259 and 1,641,389,967 on average, respectively), and was much smaller for AE (949,615,901 on average) in both runs.

Eukaryotic contamination is enriched in the pool of shorter fragments.

Samples from the human oral cavity, prepared for metagenomic sequencing, are typically associated with extensive eukaryotic contamination which can account for up to 45% of the short-read sequencing product (Shaiber et al., 2020). Therefore, we assessed the amount of

host contamination in each DNA sample by mapping reads to a reference human genome from the NCBI (GRCh38). Host DNA contamination was high for all protocols. AE had the least amount of human DNA (on average, 63%) compared to GT (on average, 75%) and PC (on average, 81%) (Table 2). This trend persisted when comparing their cumulative sequence lengths, although our analysis of read length distribution showed that the human reads were predominantly composed of shorter fragments (Figure 2, insets). The increased representation of human DNA in PC compared to other extraction methods is likely due to the use of detergents versus enzymes. Lytic detergents exert their effect on both bacterial and eukaryotic cells, while lytic enzymes target bacterial cells only. PC lacked lytic enzymes and included SDS, a stronger lytic detergent than the ones used in GT (Tween 20, Triton X-100) and AE (SLS, Brij, deoxycholate).

Read size distribution is not uniform across extraction methods.

After the removal of sequences that match the human genome, we assumed that the vast majority of the sequences were of microbial origin. We further focused on sequences that were longer than 2,500 nucleotides, and quantified the number of reads and their length distribution across GT, PC, and AE (Table 3). Overall, we observed a size distribution that was coherent with other long-read metagenomics studies, with reads reaching the 100,000 bp (Moss et al., 2020; Somerville et al., 2019).

Our comparison of the five longest reads per replicate revealed that PC produced the longest fragments (130,355 to 180,460 bp) while GT ranged from 68,275 to 92,515 bp and AE had a single 116,730 bp read followed by significantly shorter reads (28,218 to 68,189 bp) (Table 3). The replicates of PC yielded 24 and 38 reads that were over 100,000 bp (Table S1.b), but this extraction method was also associated with the smallest N50 score due to the large fraction of shorter reads (Figure 2, insets). The contribution of reads above 2,500 bp to the sequencing yield (total number of nucleotides) was greater in GT (mean, 18.65%) and AE (mean, 30.68%), while dropping to 7.11% for PC (Table 3, Figure 2). AE had a few short reads contributing to the sequencing yield, but also lacked long reads as there were only 40 and 186 reads above 20 kbp in the two runs (Table S1.b). We find it surprising that AE, an extraction strategy that can yield over a million base pair DNA fragments (Anand, 1986), produced only four reads that were longer than 50 kbp when we used identical DNA mass inputs for all extractions. We speculate that, in the absence of a fragmentation step, the very long fragments AE might have produced may have been lost during library preparation steps or become stuck in the pores during sequencing.

Size selection has limited utility and leads to substantial loss of biomass.

Despite adding the nanopore recommended 1X volume of solid-phase reversible immobilization (SPRI) beads to reaction mixtures to remove small fragments during library preparation steps, there was a large contribution of shorter reads (< 2500 bp) in our initial sequencing effort. Thus, we evaluated additional options to reduce their numbers and increase the representation of longer fragments. We sought to determine the effect of using the BluePippin system on fragment size metrics by utilizing the ‘high-pass collection mode’, which enables the collection of DNA molecules above a certain user-defined size. To maximize sequences that might contain ribosomal RNA genes, which are particularly

useful for taxonomic assignments (Camanocha & Dewhirst, 2014), we chose 6 kbp as our minimum threshold value. Due to its ample material availability, we used replicates of GT to compare the size-selected sequencing metrics with data from the previous untreated sequencing runs. High sample loss is a known drawback of BluePippin high-pass size filtering as the manufacturer warns to expect a loss between 20 – 50% of the sample input. However, we were able to recover only 16% to 18% of the 3,300 ng input DNA for each replicate after the size selection step, in agreement with low recovery rates (25% to 35%) also reported by others (Schalamun et al., 2019). Even though we started the size selection step with more than 3X the amount of DNA than was used for the untreated workflow, our recovery post-size selection was 550–600ng, meaning that our sample input into the start of library preparation was half of the DNA input that went into the untreated sequencing run. Strikingly, the number of reads and sequencing yields were reduced by 80–90% compared to the untreated sample (Table 4, Table S1.c), a likely consequence of the reduced sample input. Other notable shifts included a reduced proportion of human contamination and an increased N50 (Table 4). The number of microbial reads above 2.5 kbp and their cumulative length were comparable for both approaches, which is quite impressive given the low amount of input DNA in the size-selected samples. However, this parallel did not persist when evaluating longer DNA fragments as the cumulative length of microbial reads above 20 kbp was greater for non-size selected samples. Propelled by vastly reduced read numbers (and despite the superior N50), size selection step did not result in the substantial improvements we hypothesized in overall read lengths and the cumulative nucleotide sequences. Considering the additional demands on (1) sample requirements, (2) reagent/personnel costs, and (3) sample handling and processing times, we consider size selection of this type to have limited utility in this context.

Extraction method and read length distribution alter taxonomic profiles.

Extraction methods may have distinctive biases which can impact the determination of microbial community from metagenomic data, like a differential ability to lyse Gram-negative versus Gram-positive organisms. To investigate the microbial community composition of our long-read metagenomes we used bacterial, archaeal and eukaryotic single-copy core genes (SCGs) and ribosomal RNAs (rRNAs). Given the high rate of insertion or deletion errors in nanopore sequencing, predicting genes accurately is a significant challenge in uncorrected reads. While we assumed that these biases would similarly impact all extraction methods and elected to not correct our raw reads, we suggest the use of appropriate bioinformatics tools to correct frame-shift errors in long-read sequencing results (Arumugam et al., 2019; Huang et al., 2020).

In agreement with the sequencing yield of bacterial reads across extraction strategies, GT had the most and PC had the fewest number of genes predicted (Table 5). PC also yielded the lowest number of 16S rRNA genes with an average of 123 genes as compared to GT and AE, which yielded an average of 418 and 398 genes, respectively. Read size distribution played an important role in the recovery of SCGs and rRNAs genes: for example, PC_2 and AE_1 had a comparable sequencing yield of microbial reads (436 and 361 Mbp respectively, Table 3) but the numbers of SCGs and rRNAs genes were two times larger in AE_1, despite a smaller number of predicted genes. These two extractions strategies differed drastically

in their read size distribution (Figure 2) with AE_1 having fewer very long reads (>50,000 bp, Table S1.b) but more medium-sized ones (>2,5 kbp, Table 3). To complement this observation, AE_2 had the most SCGs and rRNA genes out of all extractions and also the most reads above 2.5 kbp (Table 3), despite having very few long reads (only three > 50 kbp, Table S1.b) and being the second smallest sample in terms of number of reads and sequencing yield before the removal of human reads (Table 3).

We used the Human Oral Microbiome Database (HOMD) to assign taxonomy to the 16S rRNA genes found in our long reads. The top genera included *Prevotella*, *Rothia*, *Streptococcus*, *Veillonella*, which are commonly found in oral samples (Mark Welch et al., 2016; Zaura et al., 2009). We observed a comparable taxonomic profile between GT and AE, with *Streptococcus*, *Veillonella*, *Prevotella* and *Actinomyces* identified as the most abundant genera (Figure 3). PC differed with relatively more *Prevotella* and *Haemophilus* (both Gram negative bacteria) and less *Streptococcus* (a Gram positive bacterium).

We then compared these profiles to the short-read sequencing of 16S rRNA gene amplicons found in a sample (TD), collected for the same individual two weeks after the initial sampling. While long-read and amplicon sequencing results were similar to each other qualitatively, the relative abundance estimates between these approaches differed (Figure 3, Table S2). Multiple sources of bias may have contributed to these differences, including the variable detection of rare taxa due to differences in depth of sequencing, differences in sampling time, the use of a single time point for TD versus the pooling of samples from all other sites, and the use of primers. For instance, our 16S rRNA gene primers did not match to TM7, a prevalent taxon in the human oral cavity, which explains their absence in the amplicon sequencing.

Finally, we compared taxonomic profiles across protocols to the short-read sequencing of two metagenomes from the same individual generated from (1) the sample used for PB_2 and (2) the sample used for 16S rRNA gene amplicon sequencing (sample TD). To estimate the relative abundance of taxa in metagenomes we used short read coverage of single-copy core genes matching to ribosomal protein S7, which was the most frequently found ribosomal protein in the assemblies of short-read metagenomes (Table S3). While the genera composition was very comparable with long-read metagenomes, we observed a lower relative abundance of *Streptococcus* and higher relative abundance TM7, which is likely due to copy number of the rRNA operons in these genera (Stoddard et al., 2015) that skew the relative abundance estimates based on short-read amplicons and long-read metagenomes. We also investigated the distribution and taxonomic assignment of ribosomal proteins in long-read metagenomes (Table S3, Figure S1), however, relatively low sequencing depth of these data and challenges associated with gene calling in uncorrected MinION sequences prevented reliable insights.

Overall, these results highlight (1) the differences observed between extraction methods (i.e. Gram negative/positive biases), but also (2) the critical role that the read size distribution has on the recovery of SCGs, ribosomal RNAs and consequently, its impact on taxonomic profiling.

Read size distribution dramatically influences assembly outcomes

We finally compared different extraction methods by assembling microbial reads using a long-read assembler, Flye (Kolmogorov et al., 2020). We also used short-reads generated with PB_2 to polish the assemblies using Pilon (Walker et al., 2014). GT and AE resulted in larger assemblies with an average size of 31,518,599 bp and 28,354,223 bp respectively, while the average assembly size of PC was much smaller at 11,269,952 bp (Table 6). The extraction method GT stood out for its high N50, but also by having the longest contig assembled (1,176,789 bp) and the most contigs above 100 kbp. The number of predicted genes was directly related to the assembly size. For that reason, GT had on average more genes predicted and more SCGs and ribosomal RNAs, which lead to more bacterial genomes expected in the assembly (based on the most frequently occurring count of SCG hits). The assembly step did not result in any circular bacterial chromosomes but there were a few circular plasmids and phage genomes. PC had the fewest circular contigs (Table 6), and while AE had more, they were shorter. GT stood out again with the most circular contigs.

Discussion

Prior to undertaking this evaluation, we established two key conditions that a successful high molecular weight DNA extraction method to study host-associated environments with low microbial biomass would need to fulfill: (1) optimizing a read length distribution profile to improve downstream sequence analyses and (2) minimizing the proportion of host-associated DNA contamination. While the utility of minimizing non-target DNA in sequencing libraries is relatively obvious from a resource conservancy point of view, the impact of the read length distribution on studies of metagenomes may escape attention when a sequencing strategy focuses on obtaining the longest possible reads. Based on the assembly of 2,267 genomes and the average length of rDNA operons, Koren et al. had suggested 7 kbp as the ideal read length that could span through most repeats for dramatic improvements in genome assembly (Koren et al., 2013). The modified column-based extraction method (GT) consistently yielded the greatest number of reads over 10 kbp and led to most successful downstream assemblies as indicated by the total size, the length of longest contigs, and the number of circularized elements.

The representation of host DNA contamination varied between the extraction methods, and reached its maximum in the phenol-chloroform extraction (PC). As a consequence, the phenol-chloroform extraction resulted in fewer microbial reads for an equal amount of DNA input (Table S2). While host DNA was enriched in the pool of shorter sequences and thus could be eliminated through size selection, our analyses revealed a high cost for this step as size selection removed many of the very long reads (> 60 kbp) and required three times more DNA as input. Size selection can also be performed using solid-phase reversible immobilization (SPRI), which holds great potential for improvements on read-size distributions. Indeed, a recent analysis of diverse and customized SPRI formulations have demonstrated significant increases in the sizing threshold of SPRI beads from 150 – 800 bp to 1.5 – 7 kbp (Stortchevoi et al., 2020). The modified PowerSoil kit methods resulted in low DNA yield, which was likely due to our modifications of the manufacturer protocols regarding the bead beating step. Extraction methods that rely on mechanical sample lysis

will likely result in much lower yields when modified to diverge from the manufacturer's guidelines to maximize the recovery of high molecular weight DNA. A unique advantage of nanopore sequencing platforms is the real-time access to the product that is being sequenced, which promotes new methods for reference-based identification of non-target DNA molecules and their real-time rejection from an active pore to minimize their impact on sequencing results (Kovaka et al., 2020). Maximizing the input DNA quality, and the use of novel assembly algorithms (Kolmogorov et al., 2020; Koren et al., 2017; Li, 2016; Ruan & Li, 2020), effective polishing strategies (Loman et al., 2015; Morisse et al., 2021; Vaser et al., 2017; Walker et al., 2014), will lead to new insights in metagenomics especially with complete circular MAGs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Nicole Robichaud for technical help with short-read sequencing. This research was supported by the Gastro-Intestinal Research Foundation, and the Simons Foundation (#687269, AME). In addition, a Helmsley Foundation grant to LB, BJ, and AME supported FT, and an NIH NIDDK grant (RC2 DK122394) supported KL and AME.

References

Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y, He C, Méheust R, Brooks B, Thomas A, Lavy A, Matheus-Carnevali P, Sun C, Goltsman DSA, Borton MA, … Banfield JF (2020). Clades of huge phages from across Earth's ecosystems. Nature, 578(7795), 425–431. 10.1038/s41586-020-2007-4 [PubMed: 32051592]

Anand R. (1986). Pulsed field gel electrophoresis: a technique for fractionating large DNA molecules. Trends in Genetics: TIG, 2, 278–283. 10.1016/0168-9525(86)90269-6

Arumugam K, Bağcı C, Bessarab I, Beier S, Buchfink B, Górska A, Qiu G, Huson DH, & Williams RBH (2019). Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. Microbiome, 7(1), 61. 10.1186/s40168-019-0665-y [PubMed: 30992083]

Ayling M, Clark MD, & Leggett RM (2020). New approaches for metagenome assembly with short reads. Briefings in Bioinformatics, 21(2), 584–594. 10.1093/bib/bbz020 [PubMed: 30815668]

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, & Holmes SP (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581–583. 10.1038/nmeth.3869 [PubMed: 27214047]

Camanocha A, & Dewhirst FE (2014). Host-associated bacterial taxa from Chlorobi, Chloroflexi, GN02, Synergistetes, SR1, TM7, and WPS-2 Phyla/candidate divisions. Journal of Oral Microbiology, 6. 10.3402/jom.v6.25468

Chen LX, Anantharaman K, Shaiber A, Murat Eren A, & Banfield JF (2020). Accurate and complete genomes from metagenomes. In Genome Research (Vol. 30, Issue 3, pp. 315–333). Cold Spring Harbor Laboratory Press. 10.1101/gr.258640.119

Cicha C, Hedges J, Novak I, Snyder D, Jutila M, & Wiedenheft B. (2020). Complete Genome Sequence of Brucella abortus Phage EF4, Determined Using Long-Read Sequencing. Microbiology Resource Announcements, 9(18). 10.1128/MRA.00212-20

Cusco A, Perez D, Viñes J, & Francino O. (2020). Long-read metagenomics to retrieve high-quality metagenome-assembled genomes from canine feces. In Research Square. 10.21203/rs.3.rs-60068/v1

Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, McLellan SL, Lücker S, & Eren AM (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nature Microbiology, 3(7), 804–813. 10.1038/s41564-018-0176-9

Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, Lakshmanan A, & Wade WG (2010). The human oral microbiome. Journal of Bacteriology, 192(19), 5002–5017. 10.1128/JB.00542-10 [PubMed: 20656903]

Driscoll CB, Otten TG, Brown NM, & Dreher TW (2017). Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Standards in Genomic Sciences, 12, 9. 10.1186/s40793-017-0224-8 [PubMed: 28127419]

Duran-Pinedo AE, & Frias-Lopez J. (2015). Beyond microbial community composition: functional activities of the oral microbiome in health and disease. Microbes and Infection / Institut Pasteur, 17(7), 505–516. 10.1016/j.micinf.2015.03.014

Edwards A, Debbonaire AR, Nicholls SM, Rassner SME, Sattler B, Cook JM, Davy T, Soares A, Mur LAJ, & Hodson AJ (2019). In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. In bioRxiv (p. 073965). 10.1101/073965

Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, & Sogin ML (2015). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. The ISME Journal, 9(4), 968–979. 10.1038/ismej.2014.195 [PubMed: 25325381]

Gregory AC, Zablocki O, Howell A, Bolduc B, & Sullivan MB (2019). The human gut virome database (p. 655910). 10.1101/655910

Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, … Sullivan MB (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell, 177(5), 1109–1123.e14. 10.1016/j.cell.2019.03.040 [PubMed: 31031001]

Hosoe A, Suenaga T, Sugi T, Iizumi T, Nagai N, & Terada A. (2020). Complete Genome Sequence of Pseudomonas putida Strain TS312, Harboring an HdtS-Type N-Acyl-Homoserine Lactone Synthase, Isolated from a Paper Mill. Microbiology Resource Announcements, 9(13). 10.1128/MRA.00055-20

Huang Y-T, Liu P-Y, & Shih P-W (2020). High-Quality Genomes of Nanopore Sequencing by Homologous Polishing. In Cold Spring Harbor Laboratory (p. 2020.09.19.304949). 10.1101/2020.09.19.304949

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, & Banfield JF (2016). A new view of the tree of life. Nature Microbiology, 1, 16048. 10.1038/nmicrobiol.2016.48

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. Nature, 486(7402), 207–214. 10.1038/nature11234 [PubMed: 22699609]

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, & Hauser LJ (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11, 119. 10.1186/1471-2105-11-119 [PubMed: 20211023]

Kasianowicz JJ, Brandin E, Branton D, & Deamer DW (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences of the United States of America, 93(24), 13770–13773. 10.1073/pnas.93.24.13770 [PubMed: 8943010]

Kinoshita Y, Niwa H, Uchida-Fujii E, & Nukada T. (2020). Complete Genome Sequence of Mycoplasma felis Strain Myco-2, Isolated from an Equine Tracheal Wash Sample in Japan. Microbiology Resource Announcements, 9(9). 10.1128/MRA.00057-20

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, & Pevzner PA (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature Methods, 17(11), 1103–1110. 10.1038/s41592-020-00971-x [PubMed: 33020656]

Kolmogorov M, Rayko M, Yuan J, Polevikov E, & Pevzner P. (2019). metaFlye: scalable long-read metagenome assembly using repeat graphs. bioRxiv, 637637. 10.1101/637637

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, & Phillippy AM (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology, 14(9), R101. 10.1186/gb-2013-14-9-r101 [PubMed: 24034426]

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, & Phillippy AM (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), 722–736. 10.1101/gr.215087.116 [PubMed: 28298431]

Köster J, & Rahmann S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics, 28(19), 2520–2522. 10.1093/bioinformatics/bts480 [PubMed: 22908215]

Kovaka S, Fan Y, Ni B, Timp W, & Schatz MC (2020). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nature Biotechnology. 10.1038/s41587-020-0731-9

Langmead B, & Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. 10.1038/nmeth.1923 [PubMed: 22388286]

Lee MD (2019). GToTree: a user-friendly workflow for phylogenomics. Bioinformatics, 35(20), 4162–4164. 10.1093/bioinformatics/btz188 [PubMed: 30865266]

Li H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103–2110. 10.1093/bioinformatics/btw152 [PubMed: 27153593]

Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. 10.1093/bioinformatics/bty191 [PubMed: 29750242]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. 10.1093/bioinformatics/btp352 [PubMed: 19505943]

Loman NJ, Quick J, & Simpson JT (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods, 12(8), 733–735. 10.1038/nmeth.3444 [PubMed: 26076426]

Maghini DG, Moss EL, Vance SE, & Bhatt AS (2020). Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. Nature Protocols. 10.1038/s41596-020-00424-x

Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, & Borisy GG (2016). Biogeography of a human oral microbiome at the micron scale. Proceedings of the National Academy of Sciences of the United States of America, 113(6), E791–E800. 10.1073/pnas.1522149113 [PubMed: 26811460]

Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, & Zengler K. (2018). Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome, 6(1), 42. 10.1186/s40168-018-0426-3 [PubMed: 29482639]

Matushek MG, Bonten MJ, & Hayden MK (1996). Rapid preparation of bacterial DNA for pulsed-field gel electrophoresis. Journal of Clinical Microbiology, 34(10), 2598–2600. https://www.ncbi.nlm.nih.gov/pubmed/8880529 [PubMed: 8880529]

Minoche AE, Dohm JC, & Himmelbauer H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology, 12(11), R112. 10.1186/gb-2011-12-11-r112 [PubMed: 22067484]

Morisse P, Marchet C, Limasset A, Lecroq T, & Lefebvre A. (2021). Scalable long read self-correction and assembly polishing with multiple sequence alignment. In Scientific Reports (Vol. 11, Issue 1). 10.1038/s41598-020-80757-5

Moss EL, Maghini DG, & Bhatt AS (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. In Nature Biotechnology (pp. 1–7). Nature Research. 10.1038/s41587-020-0422-6

Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, Brunel D, Frostegård Å, Heulin T, Jansson JK, Jurkevitch E, Kruus KL, Kowalchuk GA, Lagares A, Lappin-Scott HM, Lemanceau P, Le Paslier D, Mandic-Mulec I, Murrell JC, Myrold DD, … Simonet P. (2016). Back to the Future of Soil Metagenomics. Frontiers in Microbiology, 7, 73. 10.3389/fmicb.2016.00073 [PubMed: 26903960]

Nicholls SM, Quick JC, Tang S, & Loman NJ (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. GigaScience, 8(5). 10.1093/gigascience/giz043

Nurk S, Meleshko D, Korobeynikov A, & Pevzner PA (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Research, 27(5), 824–834. 10.1101/gr.213959.116 [PubMed: 28298430]

Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, & Pop M. (2019). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Briefings in Bioinformatics, 20(4), 1140–1150. 10.1093/bib/bbx098 [PubMed: 28968737]

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, & Hugenholtz P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology, 38(9), 1079–1086. 10.1038/s41587-020-0501-8

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, & Segata N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell, 176(3), 649–662.e20. 10.1016/j.cell.2019.01.001 [PubMed: 30661755]

Peng Y, Leung HCM, Yiu SM, & Chin FYL (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics, 28(11), 1420–1428. 10.1093/bioinformatics/bts174 [PubMed: 22495754]

Pessi IS, Viitamäki S, Eronen-Rasimus E, Delmont TO, Luoto M, & Hultman J. (2020). Truncated denitrifiers dominate the denitrification pathway in tundra soil metagenomes. In Cold Spring Harbor Laboratory (p. 2020.12.21.419267). 10.1101/2020.12.21.419267

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, & Glöckner FO (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(Database issue), D590–D596. 10.1093/nar/gks1219 [PubMed: 23193283]

Quick J, & Loman NJ (2018). DNA Extraction Strategies for Nanopore Sequencing. In Nanopore Sequencing (pp. 91–105). WORLD SCIENTIFIC. 10.1142/9789813270619_0007

Reveillaud J, Bordenstein SR, Cruaud C, Shaiber A, Esen ÖC, Weill M, Makoundou P, Lolans K, Watson AR, Rakotoarivony I, Bordenstein SR, & Eren AM (2019). The Wolbachia mobilome in Culex pipiens includes a putative plasmid. Nature Communications, 10(1), 1051. 10.1038/s41467-019-08973-w

Rognes T, Flouri T, Nichols B, Quince C, & Mahé F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. 10.7717/peerj.2584 [PubMed: 27781170]

Ruan J, & Li H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature Methods, 17(2), 155–158. 10.1038/s41592-019-0669-3 [PubMed: 31819265]

Sambrook J, Fritsch EF, & Maniatis T. (1989). Molecular cloning: a laboratory manual. In Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press. https://www.cabdirect.org/cabdirect/abstract/19901616061

Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ, McNally MA, Oakley S, Taylor A, Peto TEA, Crook DW, & Eyre DW (2018). Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. BMC Genomics, 19(1), 714. 10.1186/s12864-018-5094-y [PubMed: 30261842]

Schalamun M, Nagar R, Kainer D, Beavan E, Eccles D, Rathjen JP, Lanfear R, & Schwessinger B. (2019). Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. Molecular Ecology Resources, 19(1), 77–89. 10.1111/1755-0998.12938 [PubMed: 30118581]

Shaiber A, & Eren AM (2019). Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories [Review of Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories]. mBio, 10(3). 10.1128/mBio.00725-19

Shaiber A, Willis AD, Delmont TO, Roux S, Chen L-X, Schmid AC, Yousef M, Watson AR, Lolans K, Esen ÖC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, & Eren AM (2020).

Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biology, 21(1), 292. 10.1186/s13059-020-02195-w [PubMed: 33323122]

Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, & Albertsen M. (2021). Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nature Communications, 12(1), 2009. 10.1038/s41467-021-22203-2

Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmler S, Frey JE, & Ahrens CH (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiology, 19(1), 143. 10.1186/s12866-019-1500-0 [PubMed: 31238873]

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, & Ettema TJG (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature, 521(7551), 173–179. 10.1038/nature14447 [PubMed: 25945739]

Stoddard SF, Smith BJ, Hein R, Roller BRK, & Schmidt TM (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research, 43(Database issue), D593–D598. 10.1093/nar/gku1201 [PubMed: 25414355]

Stortchevoi A, Kamelamela N, & Levine SS (2020). SPRI Beads-based Size Selection in the Range of 2–10kb. Journal of Biomolecular Techniques: JBT. 10.7171/jbt.20-3101-002

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, … Bork P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science, 348(6237), 1261359. 10.1126/science.1261359

Takeshita K, Jang S, & Kikuchi Y. (2020). Complete Genome Sequence of Burkholderia sp. Strain THE68, a Bacterial Symbiont Isolated from Midgut Crypts of the Seed Bug Togo hemipterus. Microbiology Resource Announcements, 9(10). 10.1128/MRA.00041-20

The long view on sequencing. (2018). Nature Biotechnology, 36(4), 287. 10.1038/nbt.4125

Tippelt A, Busche T, Rückert C, & Nett M. (2020). Complete Genome Sequence of the Cryptophycin-Producing Cyanobacterium Nostoc sp. Strain ATCC 53789. Microbiology Resource Announcements, 9(14). 10.1128/MRA.00040-20

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, & Banfield JF (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 428(6978), 37–43. 10.1038/nature02340 [PubMed: 14961025]

Vaser R, Sović I, Nagarajan N, & Šikić M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research, 27(5), 737–746. 10.1101/gr.214270.116 [PubMed: 28100585]

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, & Earl AM (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One, 9(11), e112963. 10.1371/journal.pone.0112963

Wang Y, Yang Q, & Wang Z. (2014). The evolution of nanopore sequencing. Frontiers in Genetics, 5, 449. 10.3389/fgene.2014.00449 [PubMed: 25610451]

White RA 3rd, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, Fansler SJ, Glaesemann KR, Glass K, & Jansson JK (2016). Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. mSystems, 1(3). 10.1128/mSystems.00045-16

Wickham H. (2009). ggplot2. Springer New York. 10.1007/978-0-387-98141-3

Wommack KE, Bhavsar J, & Ravel J. (2008). Metagenomics: read length matters. Applied and Environmental Microbiology, 74(5), 1453–1463. 10.1128/AEM.02181-07 [PubMed: 18192407]

Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y, & Okazaki Y. (2021). Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. Nature Communications, 12(1), 27. 10.1038/s41467-020-20199-9

Yuan S, Cohen DB, Ravel J, Abdo Z, & Forney LJ (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. PloS One, 7(3), e33865. 10.1371/journal.pone.0033865

Zaura E, Keijser BJF, Huse SM, & Crielaard W. (2009). Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiology, 9(1), 259. 10.1186/1471-2180-9-259 [PubMed: 20003481]

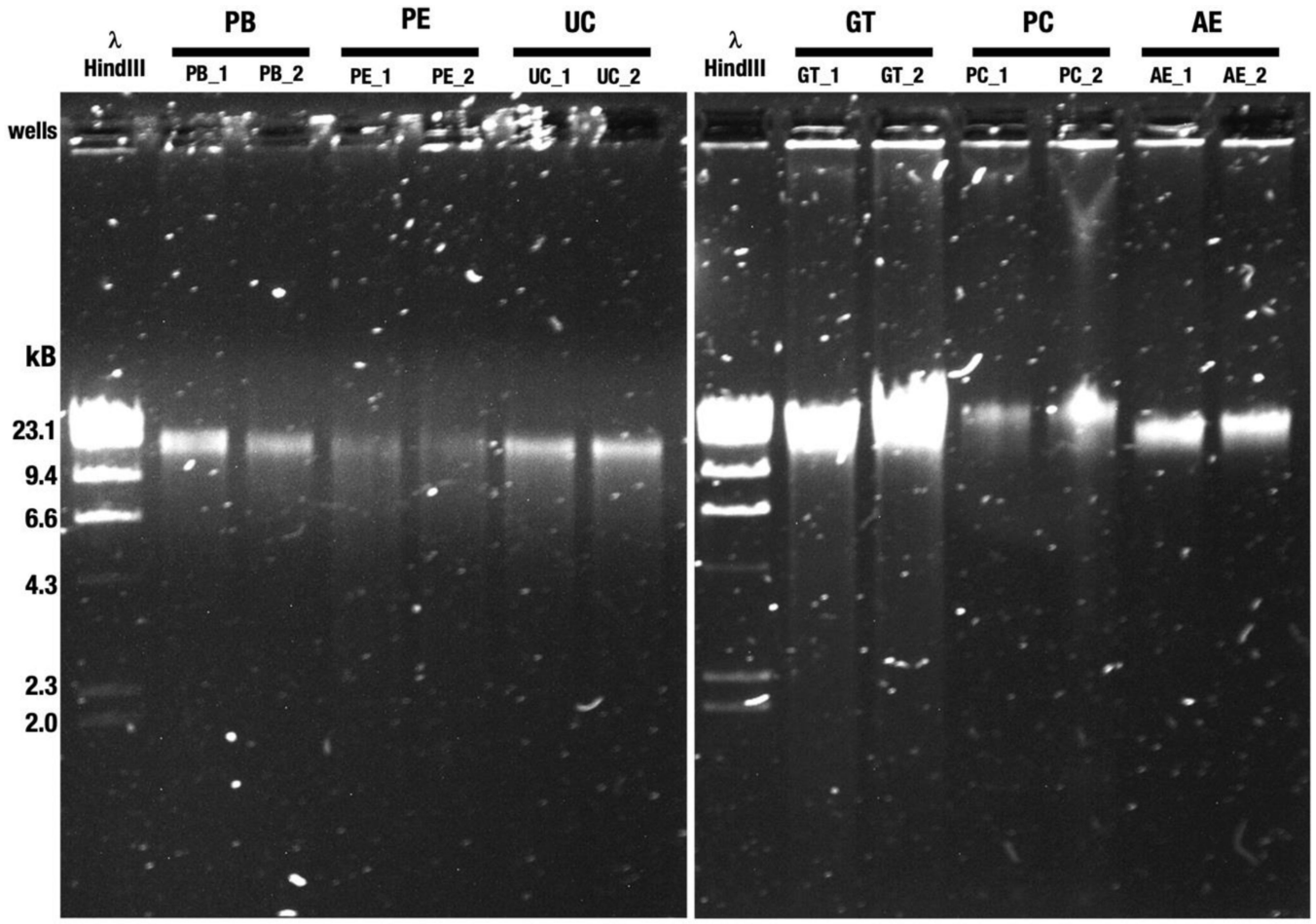

Figure 1.

Agarose gel electrophoresis of genomic DNA isolated from a pool of tongue dorsum samples. Genomic DNA was electrophoresed on a 0.8% (w/v) agarose gel. PB, PE and UC with replicates (22 ng input, left panel) and GT, PC and AE replicates (44 ng input, right panel) are shown. Different DNA inputs were used based on overall sample availability. λ-HindIII, Lambda DNA, digested with the restriction endonuclease HindIII, was used to assess fragment size distribution. PB, DNeasy PowerSoil with modified bead beating; PE, DNeasy PowerSoil with enzymatic treatment; UC, DNeasy UltraClean Microbial Kit; GT, Qiagen Genomic Tip 20/G with enzymatic treatment; PC, Phenol-Chloroform; AE, agarose encasement. The designations "_1" and "_2" indicates replicate 1 and replicate 2, respectively.

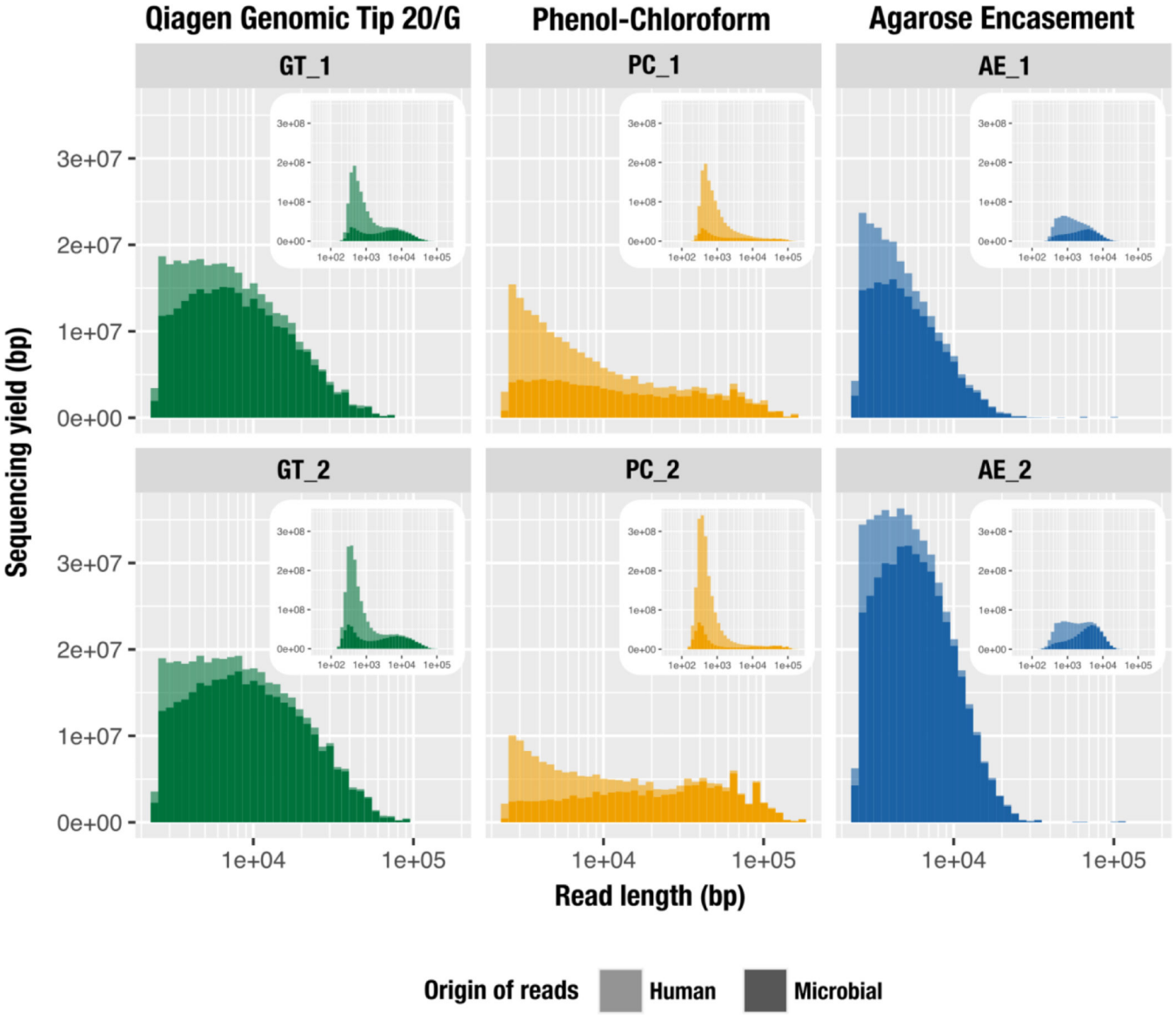

Figure 2.

The impact of DNA extraction protocol on the distribution of human (lighter color) and microbial (darker color) read lengths from MinION sequencing. These histograms visualize the total accumulative length (total number of nucleotides sequenced) per range of individual read lengths. The x-axis represents the read length in log scale and the y-axis represents the cumulative length for a given size bin (bar width). The main panel shows the size distribution of reads >2,500 bp for GT (green), PC (yellow) and AE (blue) while the inset panel shows the size distribution of all reads, using the same data. Results are outlined vertically by extraction method (replicate 1, top panel; replicate 2, lower panel).

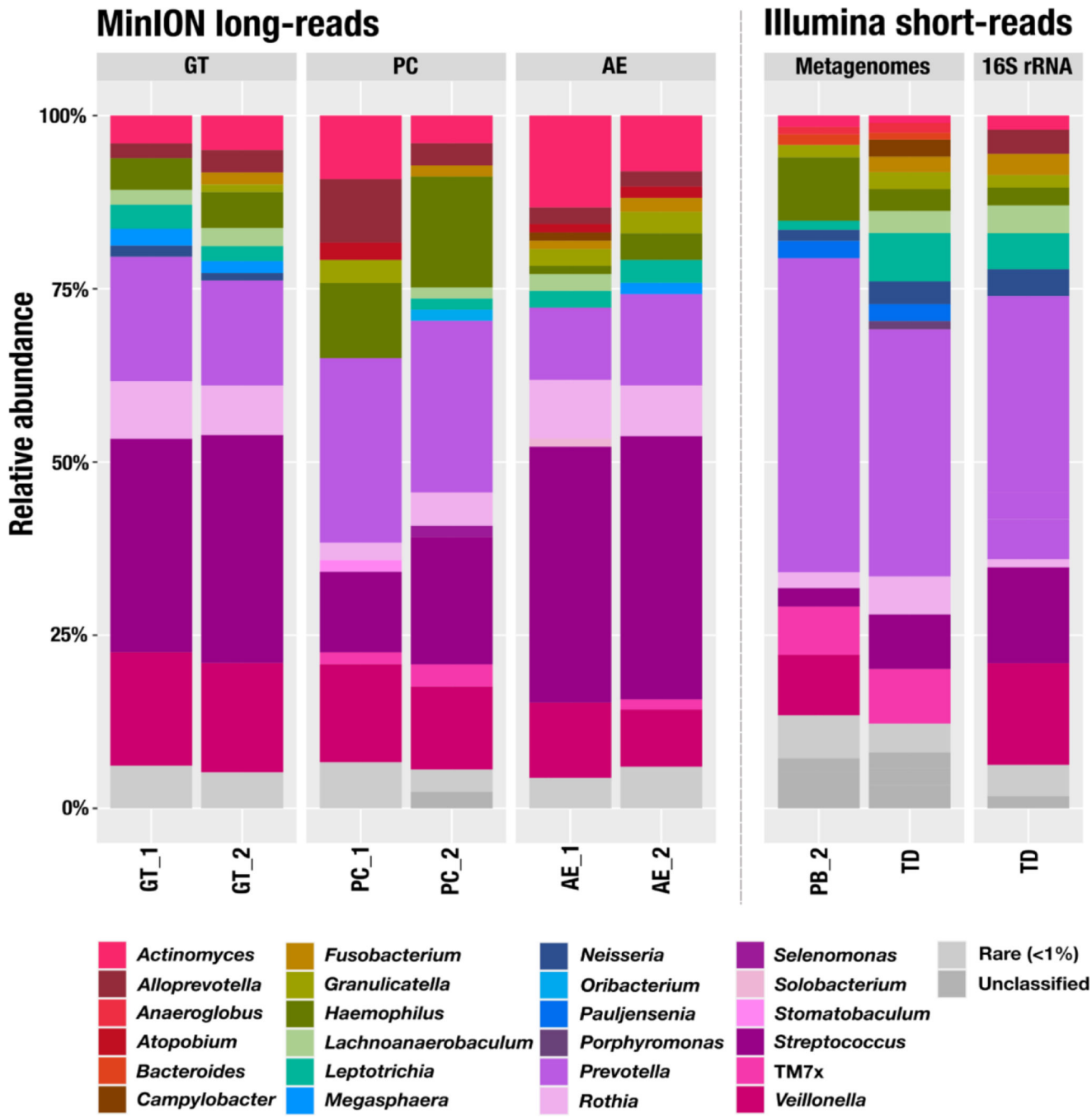

Figure 3.

Relative abundance of 16S rRNA at the genus level. We used the Human Oral Microbiome Database (HOMD) to assign taxonomy to the 16S rRNA from the MinION reads. For the short-read metagenomes, we used the taxonomy of the ribosomal gene S7 with the Genome Taxonomy Database (GTDB). We processed the 16S rRNA amplicons with the Minimum Entropy Decomposition (MED) algorithm and used Silva v132 to assign taxonomy. Genera representing less than 1% of a sample were pooled as rare (in light grey). Samples noted as TD correspond to an additional sampling performed two weeks after the initial pool

of samples used for the long-read extractions. PB, DNeasy PowerSoil with modified bead beating; GT, Qiagen Genomic Tip 20/G with enzymatic treatment; PC, Phenol-Chloroform; AE, agarose encasement. The designations "_1" and "_2" indicates replicate 1 and replicate 2, respectively.

Table 1.

Summary of DNA concentration and quality metrics. Technical replicates are denoted as “XX_1” and “XX_2”.

		Extraction Methodology											
		PB		PE		UC		GT		PC		AE	
Method	**Metric**	**PB_1**	**PB_2**	**PE_1**	**PE_2**	**UC_1**	**UC_2**	**GT_1**	**GT_2**	**PC_1**	**PC_2**	**AE_1**	**AE_2**
Qubit	**DNA concentration (μg/ml)**	2.34	1.71	4.08	4.21	30.7	36.2	110	110	75.8	77.9	90.6	95.5
Nanodrop	**A260/A280** [1]	6.01	5.53	1.75	2.27	1.95	1.74	1.74	1.82	1.97	1.95	1.8	1.76
Nanodrop	**A260/A230** [2]	0.26	0.16	0.33	0.08	1.98	2.02	1.86	2.27	2.07	2.07	1.56	1.36

[1] Primary measure of nucleic acid purity – Expected value for “pure” DNA: ~1.8 while >2 indicates RNA contamination.

[2] Secondary measure of nucleic acid purity - evaluates residual chemical contamination (phenol, guanidine HCL, carbohydrate carryover). Expected values are in the range of 1.8 – 2.2. Values significantly lower indicate chemical contamination and are undesired.

Table 2.

The impact of HMW DNA extraction protocol on proportional read numbers and sequence lengths according to read type (microbial versus human).

	Qiagen Genomic Tip (GT)		Phenol-Chloroform (PC)		Agarose Encasement (AE)	
	GT_1	**GT_2**	**PC_1**	**PC_2**	**AE_1**	**AE_2**
All Reads						
Number of reads	2,052,300	3,681,083	2,008,795	4,338,575	739,186	962,065
Sequencing yield (Gbp)	1.440	1.959	1.311	1.972	0.767	1.133
Human Reads						
Number of reads	1,584,229	2,674,013	1,664,447	3,433,184	490,377	585,733
%	77.19	72.64	82.86	79.13	66.34	60.88
Sequencing yield (Gbp)	0.908	1.199	1.026	1.536	0.405	0.463
%	63.07	61.23	78.26	77.88	52.79	40.91
Microbial Reads						
Number of reads	468,071	1,007,070	344,348	905,391	248,809	376,332
%	22.81	27.36	17.14	20.87	33.66	39.12
Sequencing yield (Gbp)	0.532	0.759	0.285	0.436	0.362	0.669
%	36.93	38.77	21.74	22.12	47.21	59.09

Table 3.

Microbial read size distribution. All percentages are relative to the total reads (or sequencing yield) of the quality filtered reads, prior to removal of human reads.

	Qiagen Genomic Tip (GT)		Phenol-Chloroform (PC)		Agarose Encasement (AE)	
	GT_1	**GT_2**	**PC_1**	**PC_2**	**AE_1**	**AE_2**
All reads						
Number of reads	2,052,300	3,681,083	2,008,795	4,338,575	739,186	962,065
Sequencing yield (Gbp)	1.440	1.959	1.311	1.972	0.767	1.133
All microbial reads						
Number of reads	468,071	1,007,070	344,348	905,391	248,809	376,332
%	22.81	27.36	17.14	20.87	33.66	39.12
Sequencing yield (Gbp)	0.532	0.759	0.285	0.436	0.362	0.669
%	36.93	38.77	21.74	22.12	47.21	59.09
N50	2,810	1,929	1,116	449	2,524	3,649
L50	38,513	62,126	34,874	155,142	38,768	52,182
Median length (bp)	468	326	427	307	782	837
Longest microbial reads (bp)	73,029	90,424 *	163,320	180,460	68,189	116,730
The top hit for the longest microbial read on NCBI's nr database (identity/ alignment)	*Streptococcus salivarius* (93.6%/ 99%)	*Streptococcus salivarius* (92%/ 98%) *	*Streptococcus sp.* (89.9%/81%)	*Veillonella dispar* (88.7%/ 69%)	*Veillonella nakazawae* (90.2%/89%)	*Prevotella histicola* (92.5%/96%)
Microbial reads > 2.5 kb						
Number of Reads	43,173	49,572	13,752	10,182	39,270	82,221
%	2.10	1.35	0.68	0.23	5.31	8.55
Sequencing yield (Gbp)	0.278	0.352	0.109	0.117	0.182	0.426
%	19.32	17.98	8.29	5.92	23.77	37.58

* showing the second longest read as the first longest read (92,515 bp) had no hits on NCBI.

Table 4.

Comparison of sequencing run read metrics between untreated and BluePippin size-selected samples. All percentages are relative to the total reads (or sequencing yield) of the quality filtered reads. GT, Qiagen Genomic Tip 20/G with enzymatic treatment; SS size-selection

	Untreated		BluePippin High-Pass Size Selection	
	GT_1	**GT_2**	**GT_1 SS**	**GT_2 SS**
All reads				
Total reads	2,052,300	3,681,083	221,344	430,986
Sequencing yield (Gbp)	1.440	1.959	0.410	0.450
Human reads				
Number of reads	1,584,229	2,674,013	117,071	282,113
%	77.19	72.64	52.89	65.46
Sequencing yield (Gbp)	0.908	1.199	0.104	0.162
%	63.07	61.23	25.31	35.92
Microbial reads				
Number of reads	468,071	1,007,070	104,273	148,873
%	22.81	27.36	47.11	34.54
Sequencing yield (Gbp)	0.532	0.759	0.306	0.289
%	36.93	38.77	74.69	64.08
N50	2,810	1,929	6,106	5,594
Microbial reads > 2.5 kb				
Number of reads	43,173	49,572	40,248	34,949
%	2.10	1.35	18.18	8.11
Sequencing yield (Gbp)	0.278	0.352	0.252	0.215
%	19.32	17.98	61.26	47.74
Microbial reads > 20 kb				
Number of reads	1,269	2,294	300	407
%	0.06	0.06	0.14	0.09
Sequencing yield (Gbp)	0.035	0.067	0.008	0.011
%	2.42	3.41	1.86	2.49

Table 5.

Results of prodigal gene calling and HMM hits for single-copy core genes (SCGs) and ribosomal RNAs (rRNAs). Analysis was performed after removal of human-reads.

	Qiagen Genomic Tip (GT)		Phenol-Chloroform (PC)		Agarose Encasement (AE)	
	GT_1	**GT_2**	**PC_1**	**PC_2**	**AE_1**	**AE_2**
Number of genes	676,577	994,050	391,675	665,986	440,845	771,653
Bacterial SCGs	2,893	3,131	1,309	1,210	2,136	3,559
Ribosomal RNAs (per 1,000 genes)	901 (1.33)	1,107 (1.11)	295 (0.75)	307 (0.46)	655 (1.49)	1,355 (1.76)
Bacterial 16S rRNA	373	462	120	125	249	547

Table 6.

Flye assembly statistics. Assemblies were polished using Pilon and short-reads from the extraction PB_2.

	Qiagen Genomic Tip (GT)		Phenol-Chloroform (PC)		Agarose Encasement (AE)	
	GT_1	**GT_2**	**PC_1**	**PC_2**	**AE_1**	**AE_2**
Total length (bp)	28,002,213	35,034,984	10,630,822	11,909,082	17,557,138	39,151,308
Number of contigs	401	466	137	106	483	855
Num contigs > 5 kb	369	412	130	101	467	775
Num contigs > 10 kb	347	383	124	100	440	714
Num contigs > 20 kb	299	336	107	97	354	564
Num contigs > 50 kb	159	194	73	67	84	203
Num contigs > 100 kb	63	84	33	35	21	69
Longest contig (bp)	1,025,627	1,176,789	504,505	676,998	304,763	875,613
Shortest contig (bp)	512	528	2103	1025	913	511
N50 (bp)	129,677	155,366	122,828	166,496	44,461	71,755
Number of genes	32,540	41,818	11,628	13,025	21,392	48,784
Single-copy core genes						
Bacteria_71	845	1072	298	339	551	1163
Archaea_76	428	554	139	167	283	614
Protista_83	34	48	11	14	26	54
Ribosomal RNAs	75	116	30	43	44	100
Num of expected bacterial genome	12	12	5	5	7	15
Circular contigs						
Number of circular contigs	20	32	8	5	9	24
Max length	86,329	155,366	155,422	155,411	24,494	88,098