High molecular weight DNA extraction, nanopore sequencing, and metagenomic assembly from the human gut microbiome

Dylan G. Maghini[1], Eli L. Moss[1], Ami S. Bhatt[1,2*]

[1] Department of Genetics, Stanford University, Stanford, California, USA

[2] Department of Medicine (Hematology, Blood and Marrow Transplantation), Stanford University, Stanford, California, USA

[*] To whom correspondence should be addressed: asbhatt@stanford.edu

Abstract

Short read metagenomic sequencing and *de novo* genome assembly of the human gut microbiome can yield draft bacterial genomes without isolation and culture. However, bacterial genomes assembled from short read sequencing are often fragmented. Furthermore, these metagenome-assembled genomes often exclude repeated genomic elements, such as mobile genetic elements, which compromises our understanding of the contribution of these elements to important bacterial phenotypes. While long read sequencing has been applied successfully to the assembly of contiguous bacterial isolate genomes, extraction of DNA of sufficient quality, length, and quantity from stool samples can be challenging. Here, we present a protocol for the extraction of microgram quantities of high molecular weight DNA from human stool samples that is sufficient for downstream long read sequencing applications. We also present a Lathe, a computational workflow for long read basecalling, assembly, and genome circularization.

Altogether, this protocol can yield highly contiguous or circular bacterial genomes from a complex human gut sample.

Introduction

Metagenomic sequencing of the mixture of bacteria, viruses, and eukaryotes in the human gut microbiome has offered new insight into the composition and functions of these organisms. With *de novo* genome assembly, microbial genomes can be assembled directly from metagenomic sequence data, offering a route to investigate genome structure and function without the need for isolation and culture. Recently, such *de novo* genome assembly approaches have led to the curation of vast databases of microbial genomes assembled directly from metagenomic sequencing, termed metagenome-assembled genomes (MAGs)[1–4]. While these genomes have offered new insight into the diversity of microbes present in the human gut, even MAGs that are considered 'complete'[5,6] can suffer from fragmented assemblies, lack of highly conserved genomic elements such as 16S ribosomal RNA sequences, and absence of mobile genetic elements. Such elements are integral to understanding genome plasticity and microbial evolution, as mobile genetic elements often are associated with biologically relevant phenotypes such as antibiotic resistance, virulence, and nutrient utilization. However, as mobile elements can range in size from hundreds to thousands of bases and can be duplicated within and across bacterial genomes, short read based assembly approaches often fail to assemble these elements and place them into proper genomic context.

Long read sequencing approaches have greatly advanced genome assembly across fields, as longer reads are capable of spanning repeated regions to solve repetitive genomes. However, extraction of sufficient quantities of high molecular weight DNA from stool is challenging as standard mechanical lysis approaches, which were developed to evenly and

efficiently extract DNA from a variety of organisms, can yield overly fragmented DNA. To overcome this challenge, we have developed a protocol that uses an enzymatic lysis approach to extract microgram quantities of pure, high molecular weight DNA from stool samples. This method has been evaluated with bacterial mock communities and results in efficient DNA extraction from a variety of Gram-positive and Gram-negative organisms. As nanopore sequencing approaches do not result in genome-scale DNA sequencing reads, we have also developed a downstream bioinformatics workflow for the basecalling, assembly, polishing, and circularization of long read metagenomic data. We have recently applied these methods to human stool samples, yielding many highly contiguous draft genomes as well as 20 single-contig, fully circular bacterial genomes[7]. The contiguity of these sequences has allowed us to identify multiple types of mobile genetic elements in bacterial genomes that have evaded circularization despite being found in high abundance, illustrating the utility of this approach for investigating bacterial genome structure and evolution.

Our protocol has integrated existing methods for bacterial DNA extraction with DNA purification and size selection approaches. Requiring less than gram quantities of input sample, our approach can yield microgram quantities of output DNA with peak lengths in the tens of kilobases (kb). This yield is sufficient for direct use with both Oxford Nanopore and PacBio sequencing applications, without the need for whole genome amplification or PCR-based amplification approaches. While our focus is on human stool sample extraction, this approach has successfully been applied to canine and murine stool samples, and with appropriate modifications, may be extended for other types of microbiomes, such as soil and ocean communities. Our computational workflow can be used for long read assembly of metagenomic or isolate long read sequencing data, and offers multiple options for assembly and polishing tools.

Here, we describe our protocol for the extraction of high molecular weight DNA from human stool samples (Figure 1), as well as our recommendations for sequencing approaches and our workflow for metagenomic assembly of long read data (Figure 2). Specifically, we describe our methods for enzymatic bacterial cell lysis, RNA and protein digestion, sample purification, and DNA size selection, as well as our choices of long read assemblers, polishing methods, and error correction. A condensed DNA extraction protocol can be found in Supplementary Note 1.

Experimental Design

Sample lysis and contaminant digestion (steps 1-7)

When aliquoting stool samples for DNA extraction, we prefer to use biopsy punches with plungers (see Equipment), as they can precisely aliquot frozen stool and limit freeze-thaw cycles. Care should be taken to avoid injury when using these sharp tools, and we recommend placing the sample tube in a rack rather than holding the tube by hand during the punching procedure. When aliquoting, one should consider the biomass of the sample. If a stool sample has a lower biomass and is more watery in consistency, a greater total mass is recommended for extraction input. Alternatively, multiple extractions can be performed in parallel and pooled on the column purification step (Step 9). After stool aliquoting and resuspension, bacteria are typically lysed through either an enzymatic or mechanical approach. Enzymatic lysis is advantageous for high molecular weight extraction, as it avoids the extensive shearing caused by mechanical lysis approaches, such as bead beating. However, vigorous bead beating remains the gold standard for unbiased lysis, as enzymatic approaches may show bias between Gram-positive and Gram-negative bacteria. While gentle bead beating is also effective for increasing high molecular weight DNA yield, it may leave some bacteria intact and result in extraction of DNA that is not representative of the sample. In general, we recommend enzymatic lysis using a combination of lytic enzyme solution (Qiagen 158928) and MetaPolyzyme

(Millipore Sigma MAC4L-5MG) for effective lysis of a range of microbes. The MetaPolyzyme enzyme mixture includes lyticase and chitinase, which disrupt glucan and chitin in cell walls, as well as lysozyme, mutanolysin, and lysostaphin, which disrupt linkages in peptidoglycans, and achromopeptidase, which is effective in lysing Gram-positive bacteria. We follow enzymatic lysis with a nucleic acid precipitation step, and then we digest RNA and protein using RNase A and proteinase K.

Genomic tip purification (steps 8-12)

Following lysis and RNA and protein digestion, we apply the sample to QIAGEN Genomic-tip columns for additional purification. When applied to the Genomic-tip column, DNA binds to the column resin while proteins, RNA, low molecular weight DNA, and other contaminants flow through. At this step, multiple extractions from the same sample can be combined into one column to increase yield. As the column operates through gravity flow, the column tip should be placed above, rather than in, a collection tube during the elution step and the sample should flow unassisted.

Size selection (steps 13-15)

After DNA has been extracted, we recommend additional size selection to deplete shorter DNA fragments and enrich for longer fragments, as long fragments and their resulting long reads are critical for assembling contiguous genomes. There are several methodological choices for additional size selection, including size selection with the Sage Science BluePippin or Solid Phase Reversible Immobilization (SPRI) bead selection steps. While BluePippin size selection is effective at accurately and thoroughly eliminating DNA below a desired threshold, we find that the total mass lost with this protocol is high, necessitating a higher input mass to ensure adequate yield. Thus, for applications where input sample is limited, we recommend size selection with SPRI beads, as they provide additional sample purification, reasonable yield, and

the supernatant can be retained for additional selection steps. Typically, SPRI beads are used for sample clean-up and size selection. As the ratio of beads to sample is increased, the binding of smaller fragments to the beads becomes more efficient. Conversely, a lower ratio of beads to sample will lead to more stringent selection for longer DNA fragments. However, as standard SPRI beads are typically intended for size selection within a range of 150-800 bp, preparation of the SPRI beads for fragment selection >2.5 kb requires a custom buffer, as previously detailed[8]. Given the variable nature of a custom buffer preparation, a range of bead to sample volume ratios should be tested with a non-precious DNA sample, such as DNA extracted from an abundant stool sample, to determine an appropriate bead to sample ratio to achieve peak DNA fragment lengths of greater than 15 kb and minimal mass under 2.5 kb. After size selection and an initial size distribution quantification with an Agilent TapeStation (see DNA quality assessment), additional rounds of size selection can be performed with lower bead to sample ratios to increase selection stringency if quantification shows retention of fragments below 2.5 kb. Conversely, the supernatant of each selection step can be retained for repooling in the event that the original selection was too stringent. The nuclease-free water volume for the final bead resuspension step can be altered to yield the desired input volume for downstream sequencing applications. We recommend resuspending the sample in 50 µl for downstream library preparation with the Oxford Nanopore Genomic DNA by Ligation kit.

DNA quality assessment (steps 16-18)

After size selection, the resultant DNA should be assessed for concentration, contamination, and size distribution. We recommend evaluating DNA concentration with a Qubit Fluorometer using the Qubit Broad Range dsDNA quantification kit, which has a quantitation range of 2-1000 ng/µl. Concentration can also be measured with a NanoDrop Spectrophotometer; however, we prefer the Qubit quantification because of its specificity for detecting dsDNA and its sensitivity at low concentrations. The Oxford Nanopore Genomic DNA by Ligation kit requires an input of 50

μL and suggests a minimum DNA concentration of 20 ng/μl, for a total mass of 1,000 ng. However, we have found that concentrations as low as 6 ng/μl (for a total of 300 ng) have yielded libraries sufficient for sequencing. It is important to assess contaminant levels when continuing to downstream nanopore sequencing, as contaminants can cause clogging of the sequencing pores in a nanopore flow cell. DNA contamination should be assessed using a NanoDrop Spectrophotometer. The suggested sample purity is A260/A230 > 2.0 and A260/A280 > 1.8. We have found that A260/A230 values as low as 1.3 can yield adequate sequencing runs. Finally, we recommend quantifying DNA size distribution using an Agilent TapeStation and accompanying Agilent Genomic DNA ScreenTape, which has a sizing range of 200-60,000 bp and rapid analysis time. Alternatively, size distribution can be assessed using the Agilent Bioanalyzer 2100 system and accompanying Agilent High Sensitivity DNA kit, which can detect a size range of 50-7,000 bp. Our suggested size distribution is a major peak mean greater than 15 kb (or 7 kb, if using a Bioanalyzer), with minimal mass below 2.5 kb (Supplementary Figure 1). If considerable mass remains below 2.5 kb, consider an additional round of SPRI bead size selection with a lower ratio of beads to sample. At this point, extracted DNA can be used for library preparation. As shotgun sequencing short reads can be used to correct long read assembly errors (see Metagenomic assembly and post-processing), we recommend performing shotgun sequencing as well as long read sequencing on the extracted DNA.

Library preparation and sequencing (steps 19-21)

We recommend preparing DNA for nanopore sequencing using the Oxford Nanopore Genomic DNA by Ligation library preparation kit which incorporates steps for DNA repair, DNA end preparation, and sequencing adapter attachment. This protocol is intended for direct DNA sequencing and we have found that it can yield up to 30 gigabase pairs (Gbp) of sequencing from a single MinION R9.4 flow cell. Additionally, this protocol includes AMPure bead cleanup

steps that improve sample purity prior to sample loading. To ensure maximum yield and an optimized ratio of available DNA ends to sequencing adapters, the input DNA should be adjusted based on the peak size and total mass to 100-200 fmol, as instructed in the extended Genomic DNA by Ligation protocol. The Oxford Nanopore Rapid Sequencing protocol is an acceptable alternative protocol for DNA extractions with lower total mass, as the protocol's suggested input is 400 ng. This protocol can be performed in 10 minutes, and uses transposome-mediated tagmentation to attach sequencing adapters to DNA. However, we find that libraries prepared with the Rapid Sequencing protocol have lower yield of total sequencing data compared to libraries prepared with 300-400 ng of input DNA using the Genomic DNA by Ligation protocol.

After library preparation, we recommend immediate flow cell loading and sequencing using Oxford Nanopore R9.4 SpotOn flow cells and the accompanying MinION sequencing devices. The MinION sequencing device is operated using the Oxford Nanopore MinKNOW software, which provides an interactive graphical user interface for controlling sequencing experiments. The MinKNOW software and MinION device require a laptop or desktop computer with at least 16 GB of RAM, an i7 CPU, and a USB3 port. When setting up a sequencing run in MinKNOW, we suggest deactivating live basecalling, as basecalling is incorporated into the downstream Lathe workflow. Flow cell loading incorporates two priming steps before applying the sample, and should be performed as instructed in the Genomic DNA by Ligation protocol. We typically find that a single sequencing run can continue to generate data for one to four days before all sequencing pores are depleted, depending on the quality of the library. After a flow cell has been used, it can be discarded or washed for re-use by applying a nuclease to digest DNA occupying sequencing pores, as instructed in the Oxford Nanopore Flow Cell Wash protocol (EXP-003).

Metagenomic assembly and post-processing (steps 22-25)

Once sequencing data have been collected, the next step is pre-processing and basecalling followed by metagenomic assembly. Various assemblers are appropriate for the assembly of long read metagenomic data. These include long read assemblers such as Canu[9], Flye[10], miniasm[11], and wtdbg2[12], and hybrid assemblers such as hybridSPADES[13] and OPERA-MS[14]. All of these approaches are in active use. Based on our experience, we favor using a long read assembly approach followed by short read or long read polishing. For this purpose, we have developed a Lathe, a workflow that combines basecalling, assembly, and circularization steps into one workflow (Figure 2)[7]. Basecalling is conducted with the Oxford Nanopore Guppy basecaller prior to assembly with either Canu or Flye. We recommend implementing Canu when aiming for highly contiguous or closed genomes while maximizing structural variant detection sensitivity. Alternatively, we recommend implementing Flye when prioritizing speed and cost of assembly. Following assembly, contigs are polished with either nanopore long reads using Racon[15] and Medaka[16], shotgun sequencing short reads using Pilon[17], or both long reads and short reads. We find that when short read coverage is relatively even across the assembly, short reads alone are sufficient for polishing steps. However, when short read coverage is uneven, such as in cases when short reads are produced from a separate extraction or from AT-rich genomes, polishing can be improved by the addition of long read polishing steps. In cases where short reads are unavailable, long reads alone can be used for error correction. After polishing, Lathe identifies candidate contigs for circularization based on contig length. Contigs are evaluated for over-circularization through end-alignment with nucmer[18] and trimmed. Additionally, Lathe collects reads aligning to either end of candidate contigs, assembles these reads with Canu, and aligns this spanning contig to the candidate contig to attempt circularization. Finally, Lathe also incorporates misassembly detection steps by identifying points in assemblies that are spanned by either zero or one long reads; Lathe then breaks contigs at these putative misassembly points. The final outputs of Lathe include

basecalled FASTQ files, the full assembly, and a folder of circularized genomes (Supplementary Figure 2). These outputs can be used directly in downstream processing steps, such as binning and taxonomic classification.

Limitations of the Approach

This DNA extraction approach has been optimized for extraction of high molecular weight DNA from human stool samples. Other types of material, such as other stool sources and bacterial isolates, may not perform optimally with this protocol, especially when stool contains undigested material that is not dense enough to be removed with centrifugation. However, it is likely that this protocol can be adapted in combination with other methods to apply to other stool types, bacterial isolates, and environmental samples. It should also be noted that mechanical lysis approaches such as bead beating remain the gold standard for more even lysis and downstream relative abundance classification, as mechanical lysis is considered less biased than enzymatic approaches. However, we have previously shown that this enzymatic approach is capable of relatively even lysis from both Gram-positive and Gram-negative organisms.

Our downstream computational workflow for basecalling, assembly, and circularization is designed for error-prone, long read sequencing data generated from nanopore or PacBio sequencing. As nanopore sequencing incurs a high error rate in homopolymer regions and PacBio sequencing has a high, but relatively random error rate, polishing with short reads is still recommended for indel correction and high quality assembly. We find that short read polishing effectiveness suffers when short reads do not evenly cover the assembly, in cases where short reads were produced from a different DNA extraction or when polishing low GC regions that are biased against in some short read library preparation methods. We anticipate that with future improvements in long read sequencing technology and basecallers, such as PacBio circular

consensus sequencing[19] and neural network basecallers[20], the need for additional polishing with short reads will become less critical.

Materials

Biological Samples

- Stool sample, stored at -80°C without buffer

  ! CAUTION All samples should be obtained with informed consent and in accordance with relevant guidelines

Reagents

- Sterile phosphate buffered saline (PBS)
- Lytic enzyme solution (Qiagen, cat. no. 158928)
- MetaPolyzyme (Millipore Sigma, cat. no. MAC4L-5MG)
- 20% Sodium Dodecyl Sulfate (SDS; Thermo Fisher Scientific, cat. no. AM9820)
- Phenol/Chloroform pH 8.0 (Sigma-Aldrich, cat. no. 77617-100ML)

  ! CAUTION Phenol can be absorbed through the skin and can cause burns. Chloroform is an irritant and possible carcinogen. Use appropriate safety measures.
- Phase lock gel (Fisher Scientific, cat. no. 14-635-5D)
- 3M sodium acetate (Fisher Scientific, cat. no. BP333-500)
- Absolute ethanol (Fisher Scientific, cat. no. BP2818500)

  ! CAUTION Ethanol is flammable. Store according to appropriate guidelines.
- Genomic DNA buffer set (Qiagen, cat. no. 19060)
- RNAse A (Qiagen, cat. no. 19101)
- Proteinase K (Qiagen, cat. no. 19133)
- Isopropanol (Fisher Scientific, cat. no. AC326960010)

  ! CAUTION Isopropanol is flammable. Store according to appropriate guidelines.
- Nuclease-free water (Invitrogen, cat. no. AM9937)

- Solid phase reversible immobilization (SPRI) beads (Fisher Scientific, cat. no. 09-981-123), prepared in custom buffer[8]
- 50% Polyethylene glycol (Sigma-Aldrich, cat. no. 202444-500G)
- 1 M Tris-HCl pH 8.0 (Fisher Scientific, cat. no. 15-567-027)
- 10% Tween-20 (Fisher Scientific, cat. no. BP337-500)
- 10 mM Tris-HCl pH 8.0 (Fisher Scientific, cat. no. 15-567-027)
- 0.5 M Ethylenediaminetetraacetic acid (EDTA; Fisher Scientific, cat. no. S311-100)
- 5 M Sodium chloride (Fisher Scientific, cat. no. S640-3)
- Qubit broad range reagents (Thermo Fisher Scientific, cat. no. Q32853)
- Agilent TapeStation genomic DNA reagents (Agilent Technologies, cat. no. 5067-5366)
- Agilent TapeStation ScreenTape (Agilent Technologies, cat. no. 5067-5365)
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63881)
- Ligation sequencing kit (Oxford Nanopore Technologies, cat. no. SQK-LSK109)

  CRITICAL We recommend the ligation sequencing kit for its higher total sequencing output. Other kits, such as the Oxford Nanopore Rapid Sequencing (SQK-RAD004) are also appropriate in cases of lower total DNA mass, but may result in lower total sequencing yield.
- NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing (NEB cat. no. E7180S)
- Flow cell (Oxford Nanopore Technologies, cat. no. FLO-MIN106D)

Equipment

- Dry ice
- Ice
- LowBind tubes, 1.5 ml (Fisher Scientific, cat. no. 13-698-791)
- LowBind tubes, 2 ml (Fisher Scientific, cat. no. 13-698-792)
- PCR tubes, 0.2 ml (Fisher Scientific, cat. no. AM12230)
- Pipettors (Fisher Scientific, cat. nos. 07-764-700, 07-764-701, 07-764-702, 07-764-704, 07-764-705)

- Aerosol barrier pipette tips (Fisher Scientific, cat. nos. 02-707-439, 02-707-432, 02-707-430, 02-707-404)
- Analytical Scale (Fisher Scientific, cat. no. S72710)
- Microcentrifuge (Fisher Scientific, cat. no. 07-203-954)
- Mini vortex (Fisher Scientific, cat. no. 14-955-151)
- Multi-head benchtop vortex (Benchmark Scientific, cat. no. BV1005)
- Benchtop centrifuge (Beckman Coulter, cat. no. 392244)
- Thermal cycler (Thermo Fisher Scientific, cat. no. A37835)
- Heat block (Fisher Scientific, cat. no. 88-870-001)
- Hula mixer (Thermo Fisher Scientific, cat. no. 15920D)
- Biopsy punch or similar for weighing stool (Fisher Scientific, cat. no. 12-460-410)

  ! CAUTION Biopsy punches are sharp and should be handled with care.
- Genomic-tip 20/G kit (Qiagen, cat. no. 10223)
- Magnetic microcentrifuge tube rack (Thermo Fisher Scientific, cat. no. 12321D)
- Qubit Fluorometer (Thermo Fisher Scientific, cat. no. Q33327)
- Agilent TapeStation or equivalent (Agilent Technologies, cat. no. G2992AA)
- Agilent TapeStation loading tips (Agilent Technologies, cat. no. 5067-5598)
- Nanodrop (Thermo Fisher Scientific, cat. no. ND-2000)
- Oxford Nanopore Technologies MinION
- Computer with Windows 7,8, or 10, OSX Sierra, High Sierra, or Mojave, or Linux Ubuntu 16.04 or 18.04, USB3 ports, and 1 TB internal storage or external SSD

Procedure

DNA Extraction (Timing: 8 h)

1. Keeping the frozen stool sample on dry ice as much as possible to maintain sample integrity. Place the sample tube in a tube rack and use a biopsy punch to aliquot 150 mg stool into a 2 ml microcentrifuge tube. Suspend the sample in 500 µl PBS and vortex for 3-4 seconds to mix. For lower biomass stool samples, aliquot up to 300 mg stool.

   ! CAUTION All samples should be obtained with informed consent and in accordance with relevant guidelines.

   ! CAUTION As biopsy punches are sharp objects and can easily slip while aliquoting. Do not hold the sample tube by hand when aliquoting.

2. Add 5 µL Qiagen lytic enzyme solution and 10 µl Metapolyzyme to the stool suspension. Mix by inverting six times slowly and gently. Incubate the mixture in a 37°C heat block for 1 hour.

3. In a fume hood, add 12 µl 20% SDS and 500 µl Phenol/Chloroform pH 8. Add approximately 100 µl of phase-lock gel to the microcentrifuge tube. Alternatively, add approximately 100 µl of phase-lock gel to the inside cap of the microcentrifuge tube rather than directly into the tube for ease of application.

4. Place tubes into the multi-position vortexer and vortex for 5 seconds at minimum speed. Centrifuge the tube for 5 min at 10,000*g* at room temperature. Decant the aqueous phase into a fresh 2 ml microcentrifuge tube.

5. Add 90 µl 3M sodium acetate and 500 µl isopropanol. Invert the tube thrice slowly to mix. Incubate the mixture at room temperature for 10 min.

6. Spin the tube for 10 min at 10,000*g* at room temperature, making sure that the hinge is facing the outside edge. While being very careful not to disrupt the pellet, remove and discard the supernatant using a P200 pipette. Wash the pellet twice with 100 µl freshly prepared 80% ethanol.
7. Add 1 mL buffer G2 from the Genomic DNA Buffer Set (Qiagen), 4 µl RNase A (100 mg/ml), and 25 µl Proteinase K. Invert the tube thrice slowly to mix. Incubate the mixture in a 56°C heat block for 1 hr. After 30 minutes of incubation, dislodge the pellet by inverting gently once or twice.

   ? TROUBLESHOOTING
8. Prewarm 1 mL per column of buffer QF from the Genomic DNA Buffer Set (Qiagen) at 56°C. Equilibrate the Genomic-tip 20/G with 1 mL buffer QBT from the Genomic DNA Buffer Set (Qiagen), and allow buffer to flow through into a waste reservoir by gravity flow.
9. Invert the sample twice gently and apply to the equilibrated Genomic-tip. Allow to enter resin by gravity flow. Wash the Genomic-tip three times with 1 mL buffer QC from the Genomic DNA Buffer Set (Qiagen). Place the Genomic-tip above a 2 mL collection tube. Elute the genomic DNA into a collection tube with 1 mL prewarmed Buffer QF.

   ? TROUBLESHOOTING
10. Precipitate the genomic DNA by adding 700 µL (0.7 volumes) of room temperature isopropanol. Invert the tube gently to mix, incubate at room temperature for 10 min, and centrifuge for 15 min at 10,000*g* at room temperature.
11. Remove the supernatant with a P200 pipette and wash the pellet with 200 µL 80% ethanol. Pipette off the 80% ethanol. The pellet may be small, and will likely dislodge from the tube wall. Do not attempt to pipette all alcohol, as this will likely remove the pellet. Instead, leave a small pool of ethanol and the pellet.

12. Air dry the pellet and the remaining ethanol by leaving the tube cap open for 10-20 min until the ethanol pool is less than 10 µL, but do not completely dry the pellet. Gently resuspend in 200 µL nuclease-free water.

    PAUSE POINT The extracted DNA can be stored at 4°C for several months

13. Prepare beads in a custom buffer as has been described[8]. Add 0.8 volumes (160 µL) of the custom bead suspension to the tube and gently flick to mix. Incubate the tube for 10 min on a Hula mixer at room temperature.

    CRITICAL STEP Bead suspension to sample ratio will vary with each preparation of the custom buffer. Test the selection stringency of each bead preparation with a non-precious sample to ensure proper selection.

14. Spin the tube down briefly and place the tube on a magnetic rack to pellet beads. Wait for approximately 3 min, or until the solution has become clear. Carefully remove the supernatant with a P200 pipette. Wash pelleted beads with 200 µL freshly prepared 80% ethanol, then pipette off ethanol. Repeat the wash step once more. Remove the tube from the magnetic rack, spin it down quickly, place the tube back on the magnetic rack, and pipette off any residual ethanol. Air dry the beads for 30 seconds.

    CRITICAL STEP Do not overdry the beads, as this may negatively impact DNA recovery and can lead to irreversible binding of DNA to the beads.

15. Remove the tube from the magnetic rack and resuspend beads in 50 µL nuclease-free water. If proceeding with the Rapid Sequencing library preparation protocol, resuspend in 15 µL nuclease-free water instead. Incubate the suspension for 10 min at 37°C. Pellet the beads on the magnetic rack for approximately 3 min, or until the solution has become clear, and transfer the eluent to a fresh microcentrifuge tube.

    PAUSE POINT The extracted DNA can be stored at 4°C for several months

16. Quantify the DNA concentration using a Qubit. The suggested minimum concentration is 20 ng/μL.

    ? TROUBLESHOOTING

17. Quantify the DNA purity using a nanodrop. The suggested purity is A260/A230 > 2, A260/A280 > 1.8.

    ? TROUBLESHOOTING

18. Quantify the DNA size distribution with an Agilent TapeStation or equivalent. The suggested size distribution is a major peak mean greater than 15 kb, with minimal mass (<50 fluorescence units) below 2.5 kb.

    PAUSE POINT The extracted DNA can be stored at 4°C for several months

    ? TROUBLESHOOTING

Library Preparation and Sequencing (Timing: 4 d)

CRITICAL This protocol is for library preparation and subsequent sequencing with an Oxford Nanopore MinION. For sequencing on other Oxford Nanopore or PacBio platforms, follow protocols designed for those platforms.

19. Prepare DNA for sequencing following the Oxford Nanopore Technologies protocol for Genomic DNA by Ligation (SQK-109)[21], modifying the instructions as described in the following steps. In adapter ligation and clean-up steps, incubate the ligation reaction on a Hula mixer for all incubation steps. Wash beads with Long Fragment Buffer to enrich for fragments longer than 3 kb. For the final elution incubation step, incubate at 37°C rather than room temperature.

20. Load a RevD R9.4 MinION flow cell (or similar) into a MinION sequencing device, check flow cell quality, and load the prepared sample as described in the Genomic DNA by Ligation protocol.

? TROUBLESHOOTING

21. Start the sequencing run as instructed, with the following modifications. Set the runtime to 96 hrs, as the flow cell may still be viable after the default run duration has elapsed. Deactivate live basecalling, as basecalling is integrated into the Lathe workflow (detailed below). Set the data output path to a location with at least 500 Gb of storage, optionally an external solid state hard drive.

22. After the run has progressed and fewer than 10 pores remain active, stop the sequencing run.

    CRITICAL STEP Depending on the sample purity and the original number of active pores in the flow cell, this step can take 1-4 days. For downstream assembly applications, we recommend generating at least 6 Gbp of long read data. Assembly contiguity will improve with increased depth of coverage.

Metagenomic Assembly and Post-processing (Timing: 5 d)

CRITICAL As basecalling, assembly, polishing, and circularization are resource-intensive processes, we recommend performing all computational analysis in a high performance computing environment.

23. Install miniconda3[22], Snakemake[23] and Singularity[24]. Clone the Lathe GitHub repository from https://github.com/bhattlab/lathe, and copy the config.yaml file into a working directory.

24. Edit the config.yaml file with desired parameters, as described on the Lathe GitHub repository. In particular, select either Canu or Flye for assembly, and configure Lathe for long read polishing (default), or short read polishing by passing in path to short reads, or

both long and short read polishing using the polish_both parameter, or no polishing using the skip_polishing parameter.

25. Run the Lathe pipeline using snakemake, as described in the Lathe GitHub repository. If basecalling has previously been conducted, one can bypass this step by preloading a final basecalled FASTQ file, as instructed in the GitHub repository.
26. Assembly output can be found in FASTA format in the 5.final folder in the sample subdirectory. For metagenomic samples, postprocess the final Lathe assembly using binning tools such as MetaBAT2[25] or DAStool[26]. Circular genomes can be found in the 3.circularization/3.circular_sequences directory, in FASTA format.

Timing

Steps 1-18, DNA extraction: 8 h

Steps 19-21, library preparation and sequencing: up to 4 d

Steps 22-25, metagenomic assembly and post-processing: 5 d

Troubleshooting

Step	Problem	Possible reason	Solution
7	Pellet does not dislodge by inversion	Small pellets may adhere to the side of the tube and be difficult to dislodge	If the pellet does not dislodge, attempt to dislodge the pellet again after 20 minutes. If the pellet does not dislodge through inversion, gently dislodge the pellet with a pipet tip after the incubation is complete.
9	Sample does not flow through column	DNA is highly concentrated	Use a disposable syringe to slowly depress air into the column to gently encourage sample flow
16	Low DNA yield (< 10 ng/μL)	Poor recovery of DNA from SPRI beads, SPRI bead ratio too stringent, SPRI beads not properly resuspended before aliquoting	Ensure SPRI beads are thoroughly resuspended. Increase the ratio of SPRI beads to sample volume (e.g. 0.85 or 0.9 volumes of bead suspension). Optionally combine supernatant and eluate and reselect.
17	DNA contamination levels are above recommended thresholds	Carryover of ethanol, phenol, or isopropanol	Perform an additional SPRI bead selection using a 1:1: ratio of beads to sample volume.
18	DNA major peak < 15 kb, or high mass below 2.5 kb	SPRI bead ratio too permissive	Perform an additional round of SPRI bead size selection using a lower ratio of beads to sample volume (e.g. 0.75 volumes of bead suspension).
20	Sample pools on SpotOn port and does not enter array	Sample entry relies on capillary action. If SpotOn port is dry, sample may not enter array.	Ensure that liquid bubbles through SpotOn port during second priming step of flow cell loading. If sample pooling occurs, load an additional 200 μL of priming mix through priming port.

Anticipated Results

This protocol describes methods for extraction, sequencing, and assembly of high molecular weight DNA from human stool samples. In our experience, we find that the DNA extraction method described here can yield 1-2 micrograms of DNA from an initial input of 300-500 mg of stool. This DNA has a size distribution peak of 15-50 kb, which is sufficient for library preparation without PCR amplification, and subsequent sequencing on an Oxford Nanopore MinION sequencer. We find that these methods are capable of generating 6-30 Gbp of long read data on MinION R9.4 flow cells. In our experience, the Lathe workflow is capable of

producing at least one circular bacterial genome from a complex gut metagenome with 6 Gbp of long read data. However, these results may vary with coverage, gut complexity, DNA fragment size, and bacterial genomic structure.

Author Contributions

E.L.M. and A.S.B. conceived the study. E.L.M., D.G.M., and A.S.B. performed all experiments and data analysis. D.G.M. and A.S.B. wrote the paper with input from all authors. All authors read and approved the final manuscript.

Acknowledgments

We thank all members of the Bhatt laboratory for experimental advice and discussions. In particular, we thank Brayon Fremin for making suggestions for the abbreviated DNA extraction protocol, and Matthew Grieshop and Summer Vance for helpful comments on the manuscript. D.G.M. was supported by the Stanford Graduate Fellowships in Science and Engineering program. E.L.M. was supported by the National Science Foundation Graduate Research Fellowship no. DGE-114747. This work was supported by the Damon Runyon Clinical investigator award, grant no. NIH R01AI148623 and NIH R01AI143757 to the Bhatt lab, and grant no. NIH P30 AG047366, which supports the Stanford ADRC. Computational work was supported by NIH S10 Shared Instrumentation grant no. 1S10OD02014101 and by NIH grant no. P30 CA124435, which supports the Genetics Bioinformatics Service Center, a Stanford Cancer Institute Shared Resource.

Competing Interests

The authors declare no competing interests.

References

1. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
2. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
3. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
4. Almeida, A. *et al.* A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. *bioRxiv* 762682 (2019) doi:10.1101/762682.
5. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
6. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
7. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0422-6.
8. Nagar, R. & Schwessinger, B. DNA size selection (>3-4kb) and purification of DNA using an improved homemade SPRI beads solution. v1 (protocols.io.n7hdhj6).

doi:10.17504/protocols.io.n7hdhj6.

9. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
10. Lin, Y. *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E8396–E8405 (2016).
11. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
12. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* (2019) doi:10.1038/s41592-019-0669-3.
13. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).
14. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).
15. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
16. Medaka — Medaka 0.12.1 documentation. https://nanoporetech.github.io/medaka/index.html.
17. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
18. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* **Chapter 10**, Unit 10.3 (2003).
19. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
20. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

21. Genomic DNA by Ligation (SQK-LSK109). *Nanopore Community* https://community.nanoporetech.com/protocols/gDNA-sqk-lsk109/v/GDE_9063_v109_revT_14Aug2019.
22. Conda — conda 4.8.3.post14+07a113d8 documentation. https://conda.io/projects/conda/en/latest/.
23. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
24. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).
25. Kang, D. *et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies*. https://peerj.com/preprints/27522/ (2019) doi:10.7287/peerj.preprints.27522v1.
26. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**, 836–843 (2018).

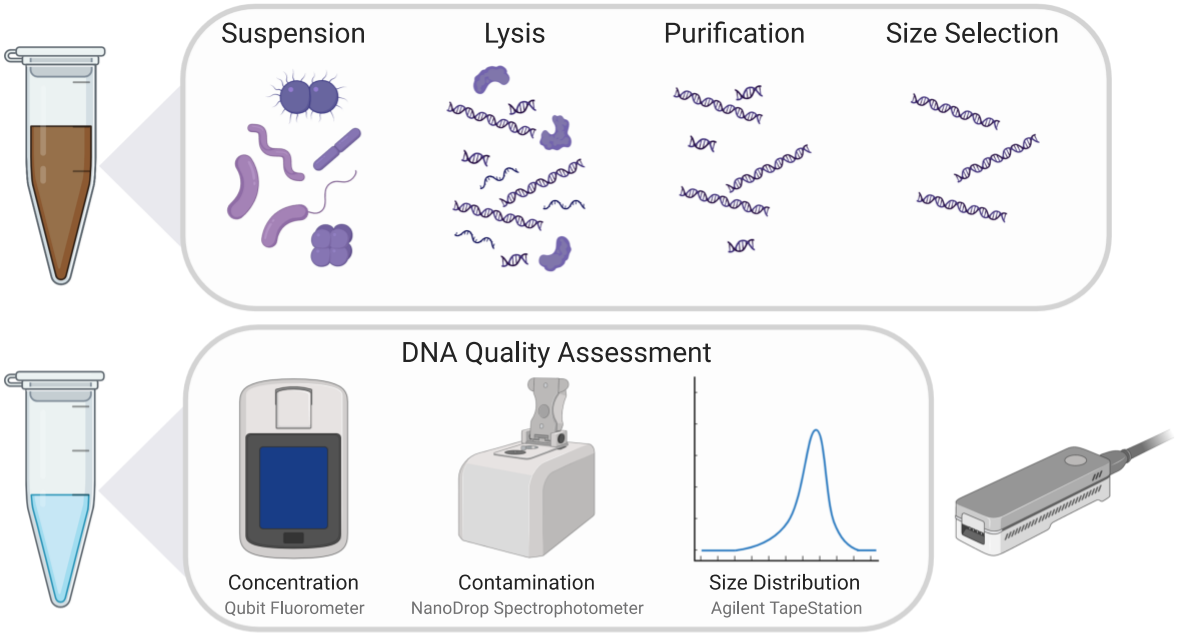


Figure 1. High molecular weight DNA extraction workflow

After aliquoting a stool sample for processing, the DNA extraction workflow proceeds with sample suspension, enzymatic bacterial lysis, and DNA purification. Large DNA fragments can then be selected for using SPRI beads prepared in a custom buffer. After selection, DNA should be assessed for concentration, contamination, and size distribution using a Qubit fluorometer, NanoDrop spectrophotometer, and Agilent TapeStation, respectively. After DNA has been extracted and meets quality thresholds, it can be carried through library preparation protocols for nanopore sequencing.

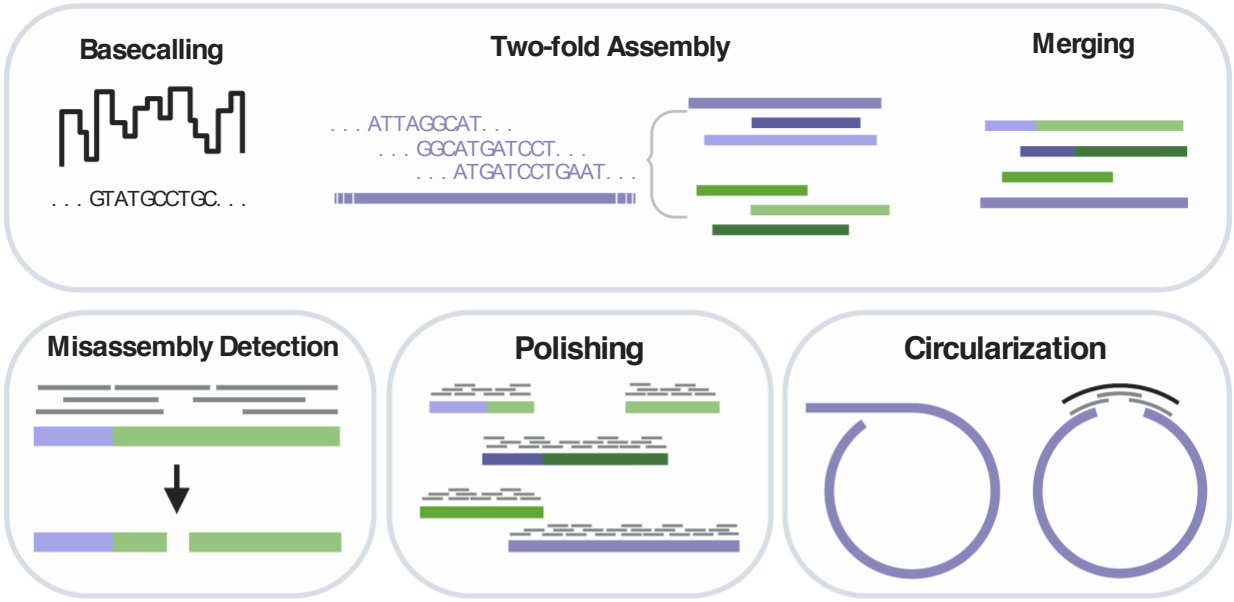


Figure 2. Post-sequencing bioinformatic workflow

After sequencing, the computational assembly workflow performs basecalling of raw nanopore signal before performing a two-fold assembly with two estimated genome size parameters. The two assemblies are then merged. The workflow also performs misassembly detection steps by breaking assembly points with low coverage, polishing steps by aligning short or long reads back to the assembly, and circularization steps through self-alignment and trimming or assembly of endpoint contigs.

Supplementary Note 1: Abbreviated high molecular weight extraction protocol

Here, we provide an abbreviated protocol for high molecular weight DNA extraction. This protocol substitutes the Qiagen Genomic-tip purification with a MicroSpin S-400 HR column purification, which is a faster column protocol. Additionally, the RNase A and proteinase K digestion steps are incorporated into other incubation steps, and the sample is incubated in SDS for more effective denaturation. Altogether, this protocol can be performed in approximately 4 hours.

Alternative reagents:

- MicroSpin S-400 HR column (Millipore Sigma cat. No. 27-5140-01)

Protocol:

1. Keeping stool sample on dry ice as much as possible to maintain sample integrity, use a biopsy punch to aliquot 150 mg stool into a 2 ml microcentrifuge tube. Suspend the sample in 500 µl PBS and vortex for 3-4 seconds to mix. For lower biomass stool samples, aliquot up to 300 mg stool.
2. Add 5 µL Qiagen lytic enzyme solution and 2 µl Metapolyzyme to the stool suspension. Mix by inverting six times slowly and gently. Incubate the mixture in a 37°C heat block for 30 min.
3. Add 2 µL RNAse A and invert to mix. Incubate for 30 minutes at 37C. Add 12 µL 20% SDS and 2 µL Proteinase K and invert to mix. Incubate for 30 minutes at 56C.
4. In a fume hood, add 500 µl Phenol/Chloroform pH 8. Add approximately 100 µl of phase-lock gel to the microcentrifuge tube. Alternatively, add approximately 100 µl of phase-lock gel to the inside cap of the microcentrifuge tube rather than directly into the tube for ease of application.
5. Place tubes into the multi-position vortexer and vortex for 5 seconds at minimum speed. Centrifuge the tube for 5 min at 10,000*g* at room temperature. Decant the aqueous phase into a new 2 mL microcentrifuge tube.
6. Add 90 µl 3M sodium acetate and 500 µl isopropanol. Invert the tube thrice slowly to mix. Incubate the mixture at room temperature for 10 min.
7. Spin the tube for 10 min at 10,000*g* at room temperature, making sure that the hinge is facing the outside edge. While being very careful not to disrupt the pellet, remove and discard the supernatant and allow the pellet to air dry. Resuspend pellet in 110 µl nuclease free water.
8. Break bottom off of MicroSpin S-400 HR column and loosen cap by a quarter turn. Spin at 1 min at 500*g*.
9. Place the column in a fresh microcentrifuge tube and apply all of the sample to the column. Spin for 2 min at 500*g*.
10. Prepare beads in a custom buffer as has been described. Add 0.8 volumes (80 µL) of the custom bead suspension to the tube and gently flick to mix. Incubate the tube for 10 min on a Hula mixer at room temperature.

CRITICAL STEP Bead suspension to sample ratio will vary with each preparation of the custom buffer. Test the selection stringency of each bead preparation with a non-precious sample to ensure proper selection.

11. Spin the tube down briefly and place the tube on a magnetic rack to pellet beads. Wait for approximately 3 min, or until the solution has become clear. Carefully remove the supernatant with a pipette. Wash pelleted beads with 200 µL freshly prepared 80% ethanol, then pipette off ethanol. Repeat the wash step once more. Remove the tube from the magnetic rack, spin it down quickly, place the tube back on the magnetic rack, and pipette off any residual ethanol. Air dry the beads for 30 seconds.

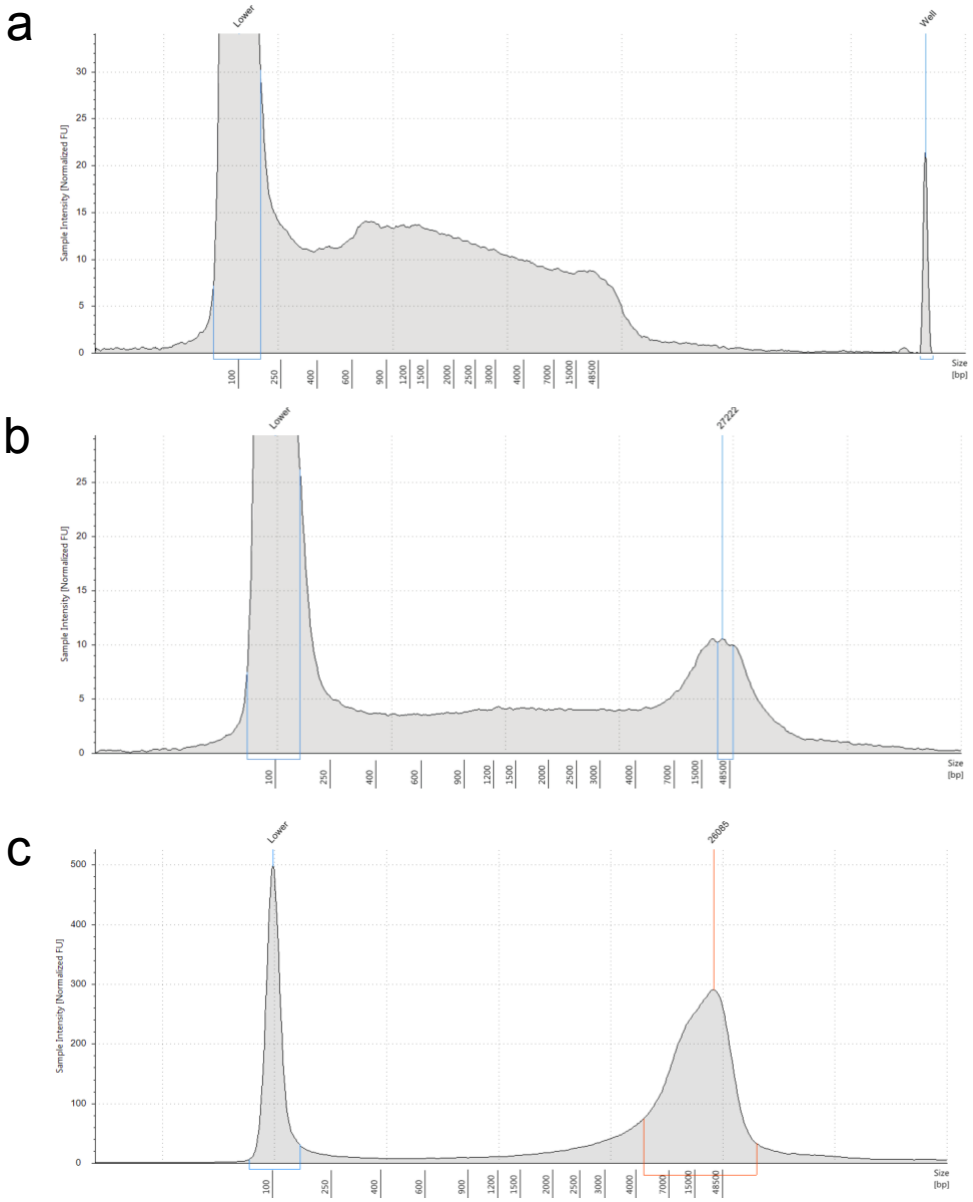
    CRITICAL STEP Do not overdry the beads, as this may negatively impact DNA recovery.

12. Remove the tube from the magnetic rack and resuspend beads in 50 µL nuclease-free water. Optionally resuspend in 15 µL nuclease-free water if proceeding with the Rapid Sequencing library preparation protocol. Incubate the suspension for 10 min at 37°C. Pellet the beads on the magnetic rack and transfer the eluent to a fresh microcentrifuge tube.

    PAUSE POINT The extracted DNA can be stored at 4°C for several months

13. Quantify the DNA concentration using a Qubit. The suggested minimum concentration is 20 ng/µL.
14. Quantify the DNA purity using a nanodrop. The suggested purity is A260/A230 > 2, A260/A280 > 1.8.
15. Quantify the DNA size distribution with a TapeStation. The suggested size distribution is a major peak mean greater than 15 kilobases, with minimal mass (<50 fluorescence units) below 2.5 kilobases.

    PAUSE POINT The extracted DNA can be stored at 4°C for several months



Supplementary Figure 1. Example DNA fragment length distributions
TapeStation traces of DNA fragment lengths for three samples, where the x-axis represents fragment size, the y-axis represents relative fluorescence at that length, and the peak centered at 100 bp indicates a molecular weight standard. A) A sample that is not recommended for nanopore sequencing, with no clear peak in DNA fragment length and excessive mass below 2.5 kb. B) A sample that could be used directly for sequencing or optionally size selected once more, which has a clear peak above 15 kb but includes mass below 2.5 kb. C) A sample that can be used directly for sequencing, with a clear peak above 15 kb and minimal mass below 2.5 kb.

0.basecall
-- *sample.fq*
-- data_links
-- nanoplots
1.assemble
-- assemble_100m (if specified)
-- assemble_250m (if specified)
-- sample_merged.fasta
-- sample_raw_assembly.fa
-- sample_raw_assembly.fa.amb
-- sample_raw_assembly.fa.ann
-- sample_raw_assembly.fa.bwt
-- sample_raw_assembly.fa.fai
-- sample_raw_assembly.fa.pac
-- sample_raw_assembly.fa.sa
2.polish
-- sample_polished.fasta
-- sample_polished.fasta.bam
-- sample_polished.fasta.bam.bai
-- sample_polished.fasta.fai
-- pilon (if specified)
-- racon (if specified)
-- medaka (if specified)

3.circularization
-- 1.candidate_genomes
-- 2.circularization
-- *3.circular_sequences*
-- 4.sample_circularized.corrected.fasta
-- 4.sample_circularized.fasta
-- 4.sample_circularized.fasta.bam
-- 4.sample_circularized.fasta.bam.bai
-- 4.sample_circularized.fasta.fai
-- 4.sample_circularized.fasta.misassemblies.tsv
5.final
-- *sample_final.fa*

Supplementary Figure 2. Lathe workflow output directory structure
Output directory of the Lathe workflow includes folders with output files from basecalling, assembly, polishing, and circularization. Subfolders are indicated in blue. Files commonly used for downstream applications include the basecalled reads (sample.fq), the circularized genomes (3.circular_sequences), and the final assembly (sample_final.fa), which are indicated with italics.