Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome

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

[1]Department of Genetics, Stanford University, Stanford, CA, USA.

[2]Divisions of Hematology and Blood & Marrow Transplantation, Department of Medicine, Stanford University, Stanford, CA, USA.

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, compromising our understanding of the contribution of these elements to important bacterial phenotypes. Although long-read sequencing has been applied successfully to the assembly of contiguous bacterial isolate genomes, extraction of DNA of sufficient molecular weight, purity and quantity for metagenomic sequencing 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 are suitable for downstream long-read sequencing applications. We also present Lathe (www.github.com/bhattlab/lathe), a computational workflow for long-read basecalling, assembly, consensus refinement with long reads or Illumina short reads and genome circularization. Altogether, this protocol can yield high-quality contiguous or circular bacterial genomes from a complex human gut sample in approximately 10 d, with 2 d of hands-on bench and computational effort.

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 advances in metagenome sequencing, assembly and binning approaches, microbial genomes can be assembled de novo 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 human samples by using metagenomic sequencing, termed metagenome-assembled genomes[1–4]. Although these genomes have offered new insight into the diversity of microbes present in the human gut, even metagenome-assembled genomes that are considered high quality[5,6] can suffer from fragmented assemblies, lack of highly conserved genomic elements such as 16S ribosomal RNA sequences and absence of integrated mobile genetic elements such as insertion sequences and transposons. Such elements are integral to understanding genome plasticity, niche specificity and microbial evolution, as mobile genetic elements are often associated with biologically relevant phenotypes, such as antibiotic resistance, virulence and nutrient utilization[7]. However, as integrated mobile elements can range in size from hundreds to thousands of bases and can be duplicated within and across bacterial genomes[8], short-read-based assembly approaches often fail to assemble these elements and place them into proper genomic context, and independently assembled elements are often not binned with their bacterial hosts because they generally have very different coverage patterns.

Long-read sequencing approaches have greatly advanced genome assembly across fields, as longer reads are capable of bridging repeated regions to solve the structure of repetitive genomes. In addition, long reads are able to directly span single-nucleotide polymorphisms within a genome, allowing for improved detection of strain heterogeneity in a complex population. Several companies have now developed kits and protocols for extracting high-molecular-weight (HMW) DNA from sample types such as bacterial and mammalian cultures, blood, tissue, yeast and fungi. Novel HMW DNA extraction methods have been developed for other sample types, such as soil and sediment, plant matter, microbial mats, hot springs and various forensic specimens. These protocols often incorporate enzymatic or chemical lysis steps, followed by column, magnetic bead or phenol-choloroform purification and subsequent size selection, and many also include a preliminary physical lysis step using bead beating, grinding, blending or vigorous vortexing (the referenced kits and methods are described further in Supplementary Table 1). However, extraction of sufficient quantities of HMW DNA from stool samples has been a longstanding obstacle, as the vigorous mechanical lysis used to evenly and efficiently extract DNA from various organisms often yields overly fragmented DNA[9]. Therefore, we developed a refined protocol that uses a robust enzymatic lysis approach to extract microgram quantities of pure HMW DNA from stool samples.

Development of the protocol

This protocol integrates novel, multi-enzymatic lysis steps with DNA purification and size-selection approaches. Requiring <1 g of input sample, our approach can yield microgram

quantities of output DNA with fragment peak lengths in the tens of kilobases. 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. This method has been evaluated with bacterial mock communities and results in efficient DNA extraction from various Gram-positive and Gram-negative organisms[10]. As long-read platforms result in higher error rates than Illumina sequencing and require various unique processing steps and tools, we have also developed a downstream bioinformatics workflow for the basecalling, error-prone read assembly and circularization of long-read metagenomic data. This workflow includes optional polishing steps that use long reads or supplementary high-accuracy short reads to correct the homopolymer errors that are common in nanopore sequencing 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 and archaeal genomes[10]. 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.

Applications

We anticipate that this approach can be adapted for various sample types and platforms. Although our focus is on human stool sample extraction, this protocol has successfully been applied to canine and murine stool samples, as well as bacterial communities[10,11]. We expect that, with appropriate modifications, this method could be extended to other types of microbiomes, such as soil and aquatic communities. In particular, we anticipate that this extraction method will be useful for samples that contain complex matrices and for diverse microbial communities, as this protocol incorporates multiple purification steps and uses a cocktail of lytic enzymes that target a range of microbes and yields extracted DNA from bacteria, viruses, fungi and archaea. Researchers attempting to adapt this protocol to an alternative sample type should review any sample-specific protocols that currently exist (HMW or otherwise) and prepare the sample according to pre-lysis specifications (i.e., specific culturing, pelleting, resuspension or homogenization methods). In addition, the researcher should review existing protocols for additional purification steps that may target contaminants specific to their sample type and consider including those steps after lysis. Finally, although the library preparation steps of this protocol are specific to sequencing with Oxford Nanopore's MinION, our computational workflow can be used for long-read assembly of metagenomic or isolate long-read sequencing data from either the Oxford Nanopore or PacBio platforms, and offers multiple options for assembly and polishing tools[10]. Please refer to 'Experimental design' for additional explanation of these options.

Overview of the procedure

Here, we describe our protocol for the improved extraction of HMW DNA from human stool samples (Fig. 1), as well as our recommendations for sequencing approaches and our workflow for metagenomic assembly of long-read data (Fig. 2). Specifically, we describe our methods for enzymatic bacterial cell lysis (Steps 2–4), RNA and protein digestion (Step 7), sample purification (Steps 8 and 9) and DNA size selection (Steps 13–15) and

our recommendations for long-read assemblers, polishing methods and error correction. A condensed DNA extraction protocol can be found in Supplementary Note 1.

Advantages and limitations—This DNA extraction approach has been optimized for extraction of HMW DNA from human stool samples but has also been validated on mock microbial communities and bacterial isolates. We expect that this protocol can be adapted to alternative sample types, although modifications to pre- and post-lysis steps will probably be necessary to account for sample-specific preparation and contaminant cleanup. Mechanical lysis approaches such as bead beating remain the gold standard for consistent 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 consistent lysis from both Gram-positive and Gram-negative organisms[10]. In addition, although mechanical lysis approaches may be sufficient in circumstances with abundant input material, such that extensive size selection is allowable, our approach optimizes for high DNA yield and is advantageous in scenarios with limited input sample volume.

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[12,13]. We find that short-read polishing effectiveness suffers when short reads do not evenly cover the assembly (e.g., 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)[14]. We anticipate that with future improvements in long-read sequencing technology and basecallers, such as PacBio circular consensus sequencing[15] and neural network basecallers[16], the need for additional polishing with short reads will become less critical.

Alternative methods

Bead beating is the most common alternative to enzymatic lysis for DNA extraction from stool. Indeed, this mechanical lysis approach has been widely adopted for lysis of both Gram-positive and Gram-negative bacteria within complex matrices. Although the DNA obtained through bead beating is largely too sheared to produce long reads, subsequent stringent size-selection steps can yield higher-molecular-weight DNA. This method can be advantageous as bead beating effectively lyses a range of organisms. However, much of the DNA obtained after bead beating is sheared to very small sizes, and thus DNA yield after size selection is often limiting. This low yield can be overcome by performing many parallel extractions for samples with adequate biomass and quantity. One additional consideration when using bead beating is that size selection may be biased against organisms that are more easily lysed and have shorter, more sheared DNA. Additional approaches for HMW DNA extraction from various sample types can be found in Supplementary Table 1.

Experimental design

Sample lysis and contaminant digestion (Steps 1–7)—We recommend using samples that have been divided into aliquots and frozen as soon after collection as possible, because extended time at room temperature before freezing can lead to altered microbial abundance, as some taxa may die while others continue to divide[17,18]. When dividing stool samples into aliquots for DNA extraction, we prefer to use biopsy punches with plungers, which can precisely produce frozen stool aliquots 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 preparing aliquots, 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 preparation of stool aliquots and resuspension, bacteria are typically lysed through either an enzymatic or mechanical approach. Mechanical approaches such as vigorous bead beating remain the gold standard for unbiased lysis, as enzymatic approaches may show bias between Gram-positive and Gram-negative bacteria. However, vigorous bead beating can cause DNA shearing, and therefore is a feasible method for long-read sequencing only when DNA yield is sufficient for extensive size selection. Although gentle bead beating is also effective for increasing HMW DNA yield, it may leave some bacteria intact and result in extraction of DNA that is not representative of the sample. Enzymatic lysis is advantageous for HMW extraction, as it avoids the extensive shearing caused by mechanical lysis approaches. 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 with minimal shearing. 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, a lysyl endopeptidase that is effective in lysing Gram-positive bacteria (enzymatic lysis agents are detailed in Supplementary Table 2). Characterization of bias between this enzymatic approach and vigorous bead beating approaches has shown that enzymatic extraction can lead to some lysis bias, but that this bias does not systematically deplete Gram-positive organisms[10]. We follow enzymatic lysis with a nucleic acid precipitation step, and then we digest RNA and protein by using RNase A and proteinase K, respectively.

Genomic tip purification (Steps 8–12)—After 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 the column. 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, the Circulomics Short Read Eliminator kit or Solid Phase Reversible Immobilization (SPRI) beads. Although 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. The Circulomics Short Read Eliminator kit has not been evaluated in the context of this extraction protocol, but it is commonly used in HMW DNA extraction protocols and is worth consideration here. For applications where input sample is limited, we recommend size selection with SPRI beads, as they provide further sample purification and 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 detailed in Materials. 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 >15 kb and minimal mass <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. It is important to note that size selection may lead to bias in organismal relative abundance, as organisms that lyse more easily may have smaller DNA fragment sizes and be selected against during size selection.

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 by using the Qubit Broad Range double-stranded DNA (dsDNA) quantification kit, which has a quantitation range of 2–1,000 ng/µl. Concentration can also be measured with a NanoDrop spectrophotometer, but 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 long-read sequencing, as both Oxford Nanopore and PacBio platforms are sensitive to impurities. For example, contaminants can decrease the efficiency of enzymatic steps during library preparation and cause clogging of the sequencing pores in a nanopore flow cell (see https://community.nanoporetech.com/contaminants). DNA contamination should be assessed by using a NanoDrop spectrophotometer. The suggested sample purity is $A_{260}/A_{230} > 2.0$ and $A_{260}/A_{230} > 1.8$. We have found that A_{260}/A_{230} values as low as 1.3 can yield adequate sequencing runs, with sustained pore viability for more than 2 d and yields of >15 giga base pairs (Gbp) of sequencing data after basecalling. Finally, we recommend quantifying DNA size distribution by 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 by 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 >15 kb (or 7 kb, if using a Bioanalyzer), with minimal mass <2.5 kb (Supplementary Fig. 1). If considerable mass remains <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. Supplementary shotgun sequencing of short reads can be used to correct long-read assembly errors (see Metagenomic assembly and post-processing), so we recommend performing shotgun sequencing as well as long-read sequencing on the extracted DNA.

Library preparation and sequencing (Steps 19–22)—Although we generally recommend sequencing on the Oxford Nanopore MinION (or equivalent) platform because of its lower equipment costs, portability and ability to generate extremely long reads, the PacBio platform offers an excellent alternative in settings where researchers have access to a PacBio sequencer. Although generally quite comparable to one another, one potential advantage of the PacBio platform over the Oxford Nanopore MinION is that the PacBio system yields reads with relatively random error profiles, which are more easily corrected in assembly when compared to the frequent homopolymer errors of the Oxford Nanopore MinION platform[12,13]. For the purpose of this protocol, we recommend preparing DNA for nanopore sequencing by 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 Gbp of sequencing from a single MinION R9.4 flow cell. In addition, this protocol includes AMPure bead cleanup steps that improve sample purity before sample loading. To ensure maximum yield and an optimized ratio of available DNA ends to sequencing adapters, the input DNA should be adjusted on the basis of 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 min and uses transposome-mediated tagmentation to attach sequencing adapters to DNA. However, we find that libraries prepared with the Rapid Sequencing protocol have a lower yield of total sequencing data than do libraries prepared with 300–400 ng of input DNA by using the Genomic DNA by Ligation protocol.

After library preparation, we recommend immediate flow cell loading and sequencing by using Oxford Nanopore R9.4 SpotOn flow cells and the accompanying MinION sequencing devices. The MinION sequencing device is operated by 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 16 GB of random-access memory, 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 1–4 d 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 23–26)—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[19], Flye[20], miniasm[21] and wtdbg2[22], and hybrid assemblers, such as hybridSPADES[23] and OPERA-MS[24]. All of these approaches are in active use. On the basis of our experience, we favor using a long-read assembly approach followed by short-read or long-read polishing. For this purpose, we have developed Lathe, a workflow that combines basecalling, assembly and circularization steps into one workflow (Fig. 2)[10]. A tutorial for Lathe is available on GitHub (https://github.com/bhattlab/lathe). Basecalling is conducted with the Oxford Nanopore Guppy basecaller before 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. After assembly, contigs are polished with long reads by using Racon[25] and Medaka (https://nanoporetech.github.io/medaka/), shotgun sequencing short reads by using Pilon[26] or both long reads and short reads. For reference, our previous work has identified baseline error rates of 494 indels and 88 mismatches per 100 kbp in an assembly of a mock community, while long-read polishing removes 83% of indels and 71% of mismatches, and short-read polishing removes 93% of indels and 82% of mismatches[10]. Therefore, in most cases, short-read polishing is preferred for its increased accuracy and speed. When performing short-read polishing, Lathe subsamples reads to 50× coverage to limit computational time, as even coverage of 50× across the assembly is sufficient for even polishing. 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 by using Medaka, which uses neural networks to recognize and correct nanopore homopolymer errors. After polishing, Lathe evaluates candidate contigs >1.7 Mb for circularization. This threshold is set to include most whole genome-scale contigs while excluding smaller contigs to speed up the computationally intensive circularization steps. Contigs are evaluated for

over-circularization through end-alignment with nucmer[27] and trimmed. In addition, 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 read; 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 Note 2). These outputs can be used directly in downstream processing steps, such as binning and taxonomic classification.

Materials

Biological samples

- Stool sample, stored indefinitely at −80 °C without buffer **! CAUTION** All samples should be obtained with informed consent and in accordance with relevant guidelines.

Reagents

- Sterile PBS

- Lytic enzyme solution (Qiagen, cat. no. 158928)

- MetaPolyzyme (Millipore Sigma, cat. no. MAC4L-5MG)

- 20% 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)

- 3 M 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)

- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63881), with an aliquot prepared in custom buffer for size selection (https://www.protocols.io/view/dna-size-selection-3-4kb-and-purification-of-dna-u-n7hdhj6)

- 50% Polyethylene glycol (Fisher Scientific, cat. no. BP233-100)

- 1 M Tris-HCl, pH 8.0 (Thermo Fisher Scientific, cat. no. 15568025)

- 10% Tween 20 (Fisher Scientific, cat. no. BP337-500)

- 0.5 M 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)

- 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 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 and 07-764-705)

- Aerosol barrier pipette tips (Fisher Scientific, cat. nos. 02-707-439, 02-707-432, 02-707-430 and 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 terabyte of internal storage or an external solid-state drive

- Lathe workflow (https://github.com/bhattlab/lathe)

Reagent setup

For HMW DNA extraction, AMPure XP beads must be washed of their standard buffer and resuspended in a custom stock solution. To prepare the stock solution, mix 4.24 ml of Milli-Q water, 100 μl of 1M Tris-HCl, 20 μl of 0.5 M EDTA (pH 8), 2.2 ml of 50% (wt/vol) PEG, 200 μl of 10% (vol/vol) Tween 20 and 3.2 ml of 5 M NaCl. To make the custom bead solution, resuspend the SPRI beads by gently inverting the bottle, then pipette 40 μl of the resuspended SPRI beads in a microcentrifuge tube and place the tube on a magnetic bead rack. After the beads have pelleted, remove the original buffer. Add 1 ml of Milli-Q water and remove the microcentrifuge tube from the magnetic rack. Vortex to thoroughly mix the beads with the water. Place the tube back on the magnetic rack until the beads have pelleted against the magnet (typically ~30–60 s) and then pipette off the wash buffer (water). Repeat this wash process four times in total. After the final wash, pipette off the supernatant from the pelleted beads and resuspend the beads in 40 μl of custom stock solution.

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 distribute a 150-mg aliquot of stool into a 2-ml microcentrifuge tube. For lower biomass stool samples, prepare aliquots of up to 300 mg of stool.

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

! CAUTION Biopsy punches are sharp objects and can easily slip while preparing aliquots. Do not hold the sample tube by hand when preparing aliquots.

2. Suspend the sample in 500 μl of PBS and vortex for 3–4 s to mix. Add 5 μl of Qiagen lytic enzyme solution and 10 μl of MetaPolyzyme to the stool suspension. Mix by inverting six times slowly and gently. Incubate the mixture in a 37 °C heat block for 1 h.

3. In a fume hood, add 12 μl of 20% (wt/vol) SDS and 500 μl of phenol/chloroform (pH 8). Add ~100 μl of phase-lock gel to the microcentrifuge tube. Alternatively, add ~100 μl of phase-lock gel to the inside cap of the microcentrifuge tube rather than directly into the tube for ease of application.

   ▲ **CRITICAL STEP** Use of phase-lock gel is optional but makes pipetting in the next step much easier. However, using more phase-lock gel than specified will negatively affect yield.

4. Place tubes into the multi-position vortexer and vortex for 5 s at minimum speed. Centrifuge the tube for 5 min at 10,000g at room temperature (18–21 °C). Decant the aqueous phase into a fresh 2-ml microcentrifuge tube.

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

6. Spin the tube for 10 min at 10,000g at room temperature, making sure that the hinge is facing the outside edge. Being very careful not to disrupt the pellet, remove and discard the supernatant by using a P200 pipette. Wash the pellet twice with 100 μl of freshly prepared 80% (vol/vol) ethanol.

7. Add 1 ml of buffer G2 from the Genomic DNA Buffer Set (Qiagen), 4 μl of RNase A (100 mg/ml) and 25 μl of Proteinase K. Invert the tube three times slowly to mix. Incubate the mixture in a 56 °C heat block for 1 h. After 30 min 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.

9. Equilibrate the Genomic-tip 20/G with 1 ml of buffer QBT from the Genomic DNA Buffer Set (Qiagen) and allow buffer to flow through into a waste reservoir by gravity flow. 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 of 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 of prewarmed buffer QF.

   ▲ **CRITICAL STEP** The Genomic-tip should not be forced into a tube. It must sit above the waste container and collection tube. This can generally be done

by using tube racks and/or pipette tip boxes if no other appropriate set-up is available.

? TROUBLESHOOTING

10. Precipitate the genomic DNA by adding 700 µl (0.7 volumes) of room temperature isopropanol. Invert the tube gently to mix and incubate at room temperature for 10 min.

11. Centrifuge for 15 min at 10,000g at room temperature. Remove the supernatant with a P200 pipette and wash the pellet with 200 µl of 80% (vol/vol) ethanol. Pipette off the 80% ethanol.

   ▲ CRITICAL STEP The pellet may be small and will probably dislodge from the tube wall. Do not attempt to pipette off all of the 80% ethanol, as this will probably remove the pellet. Instead, leave a small pool of ethanol and the pellet.

12. Air dry the pellet and the remaining 80% ethanol by leaving the tube cap open for 10–20 min until the ethanol pool is   10 µl, but do not completely dry the pellet. Gently resuspend in 200 µl of 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 specified in 'Reagent setup'. 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 The 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 the beads. Wait for ~3 min or until the solution has become clear. Carefully remove the supernatant with a P200 pipette. Wash pelleted beads with 200 µl of freshly prepared 80% ethanol (vol/vol) and then pipette off the 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 s.

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

15. Remove the tube from the magnetic rack and resuspend the beads in 50 µl of nuclease-free water. If proceeding with the Rapid Sequencing library preparation protocol, resuspend in 15 µl of nuclease-free water instead. Incubate the suspension for 10 min at 37 °C. Pellet the beads on the magnetic rack for ~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 by using a Qubit. The suggested minimum concentration is 20 ng/μl.

? TROUBLESHOOTING

17. Quantify the DNA purity by using a nanodrop. The suggested purity is $A_{260}/A_{230} > 2$, $A_{260}/A_{280} > 1.8$.

? TROUBLESHOOTING

18. Quantify the DNA size distribution with an Agilent TapeStation or equivalent. The suggested size distribution is a major peak mean >15 kb, with minimal mass (<50 fluorescence units) <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; https://community.nanoporetech.com/protocols/gDNA-sqk-lsk109/v/ GDE_9063_v109_revT_14Aug2019), 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 h, 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 500 GB of storage, optionally an external solid state hard drive.

22. After the run has progressed and <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 d. For downstream assembly applications, we recommend generating 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 (https://docs.conda.io/en/latest/miniconda.html), Snakemake[28] and Singularity[29]. 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), short-read polishing by passing in path to short reads, both long-read and short-read polishing by using the polish_both parameter or no polishing by using the skip_polishing parameter.

25. Run the Lathe pipeline by 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. Postprocess the final Lathe assembly by using binning tools such as MetaBAT2[30] or DAStool[31]. Circular genomes can be found in the 3.circularization/3.circular_sequences directory, in FASTA format.

Troubleshooting

Troubleshooting advice can be found in Table 1.

Timing

Steps 1–18, DNA extraction: 8 h

Steps 19–22, library preparation and sequencing: 4 d

Steps 23–26, metagenomic assembly and post-processing: 5 d

Anticipated results

This protocol describes methods for extraction, sequencing, assembly and binning of HMW DNA from human stool samples. In our experience, we find that the DNA extraction method described here can yield 1–2 μg 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 (Supplementary Fig. 1). 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Data availability

No new data were generated or analyzed for this manuscript.

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). [PubMed: 30661755]

2. Almeida A et al. A new genomic blueprint of the human gut microbiota. Nature 568, 499–504 (2019). [PubMed: 30745586]

3. Nayfach S, Shi ZJ, Seshadri R, Pollard KS & Kyrpides NC New insights from uncultivated genomes of the global human gut microbiome. Nature 568, 505–510 (2019). [PubMed: 30867587]

4. Almeida A et al. A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. Nat. Biotechnol Forthcoming (2020).

5. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P & Tyson GW CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015). [PubMed: 25977477]

6. Bowers RM 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). [PubMed: 28787424]

7. Vandecraen J, Chandler M, Aertsen A & Van Houdt R The impact of insertion sequences on bacterial genome plasticity and adaptability. Crit. Rev. Microbiol 43, 709–730 (2017). [PubMed: 28407717]

8. Darmon E & Leach DRF Bacterial genome instability. Microbiol. Mol. Biol. Rev 78, 1–39 (2014). [PubMed: 24600039]

9. Yuan S, Cohen DB, Ravel J, Abdo Z & Forney LJ Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS ONE 7, e33865 (2012). [PubMed: 22457796]

10. Moss EL, Maghini DG & Bhatt AS Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat. Biotechnol 38, 701–707 (2020). [PubMed: 32042169]

11. Ribado JV The impact of environmental exposures on the human and mouse gut microbiome. Dissertation, Stanford University, 2019).

12. Rang FJ, Kloosterman WP & de Ridder J From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19, 90 (2018). [PubMed: 30005597]

13. Rhoads A & Au KF PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13, 278–289 (2015). [PubMed: 26542840]

14. Tamburini FB et al. Short- and long-read metagenomics of South African gut microbiomes reveal a transitional composition and novel taxa. Preprint at https://www.biorxiv.org/content/10.1101/2020.05.18.099820v2.

15. Wenger AM et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol 37, 1155–1162 (2019). [PubMed: 31406327]

16. Wick RR, Judd LM & Holt KE Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20, 129 (2019). [PubMed: 31234903]

17. Gorzelak MA et al. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. PLoS One 10, e0134802 (2015). [PubMed: 26252519]

18. Flores R et al. Collection media and delayed freezing effects on microbial composition of human stool. Microbiome 3, 33 (2015). [PubMed: 26269741]

19. 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). [PubMed: 28298431]

20. Lin Y et al. Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl Acad. Sci. USA 113, E8396–E8405 (2016). [PubMed: 27956617]

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

22. Ruan J & Li H Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17, 155–158 (2019). [PubMed: 31819265]

23. Antipov D, Korobeynikov A, McLean JS & Pevzner PA hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32, 1009–1015 (2016). [PubMed: 26589280]

24. 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). [PubMed: 31359005]

25. Vaser R, Sovi I, Nagarajan N & Šiki M Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746 (2017). [PubMed: 28100585]

26. Walker BJ et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963 (2014). [PubMed: 25409509]

27. Delcher AL, Salzberg SL & Phillippy AM Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinformatics Chapter 10, Unit 10.3 (2003).

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

29. Kurtzer GM, Sochat V & Bauer MW Singularity: scientific containers for mobility of compute. PLoS ONE 12, e0177459 (2017). [PubMed: 28494014]

30. Kang D et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359 (2019). [PubMed: 31388474]

31. Sieber CMK et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol 3, 836–843 (2018). [PubMed: 29807988]
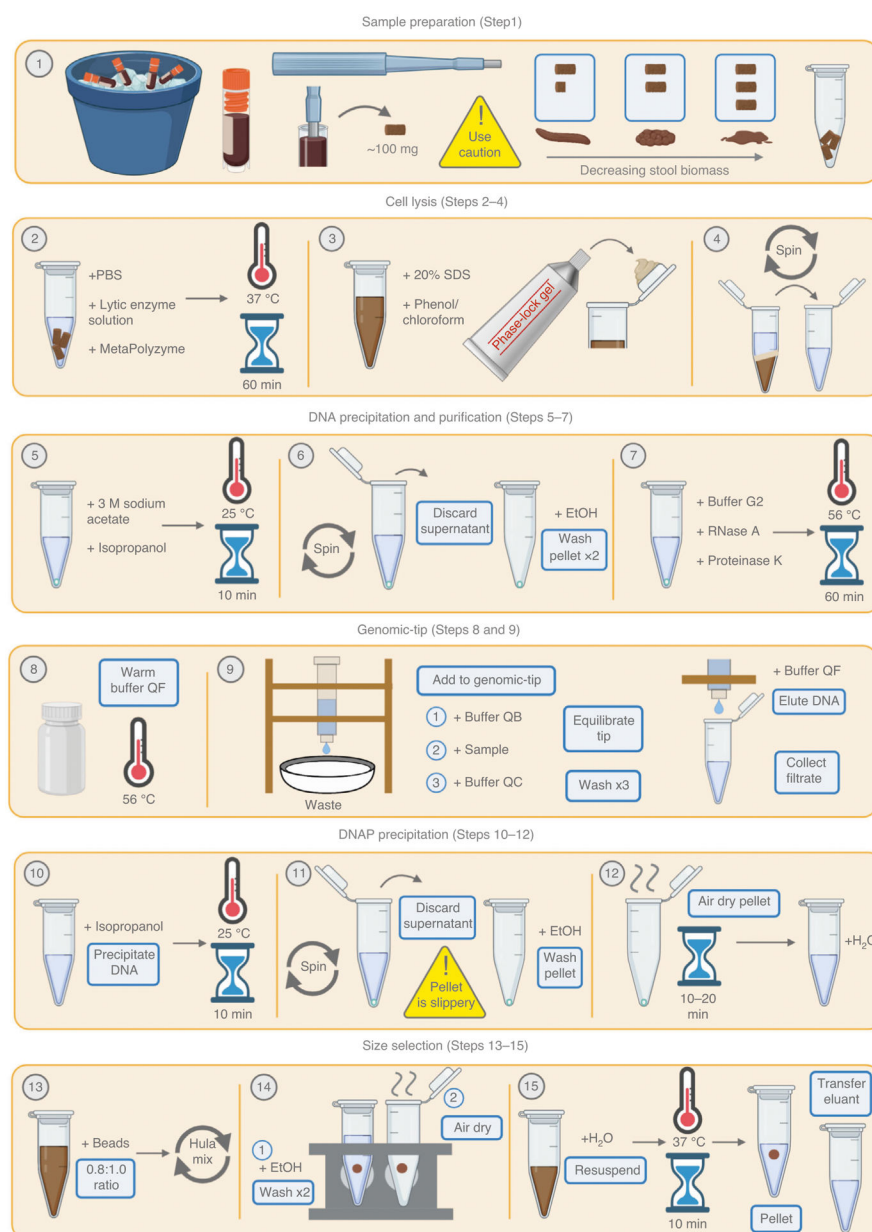
Fig. 1 |. High-molecular-weight DNA extraction workflow.
Frozen stool is best divided into aliquots by using a biopsy punch with a plunger, taking extreme caution with the blade. After dividing the stool into aliquots, the DNA extraction workflow proceeds with sample suspension, enzymatic and chemical lysis and DNA purification. The Qiagen Genomic-tip is a column-based purification method designed to minimize shearing of DNA. Allow all reagents to flow through the Genomic-tip column by gravity; do not use a syringe or vacuum to pull liquid through, as this will destroy the column. After DNA has been extracted, large fragments can then be selected for using SPRI beads prepared in a custom buffer. Extracted DNA should be assessed for concentration, contamination and size distribution by using a Qubit fluorometer, NanoDrop spectrophotometer and Agilent TapeStation, respectively (not depicted). After DNA has
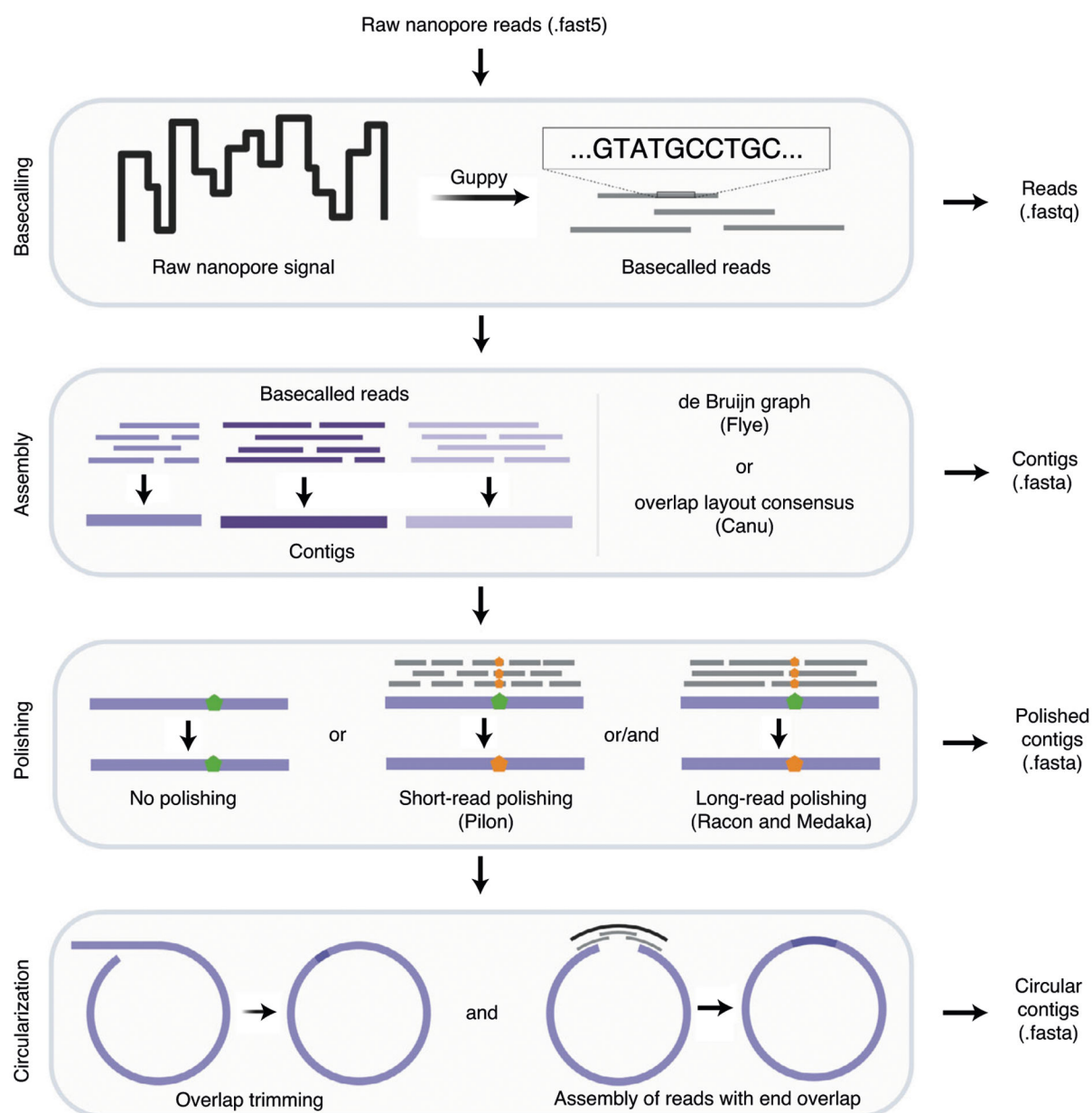
been extracted and meets quality thresholds, it can be carried through library preparation protocols for nanopore sequencing. EtOH, ethanol.

Fig. 2 |. Post-sequencing bioinformatic workflow.
After sequencing, the computational assembly workflow (Step 25) is used to perform basecalling of raw nanopore signal before performing assembly by using either Flye or Canu assemblers. The workflow then provides optional polishing steps by aligning short or long reads back to the assembly and correcting single nucleotide errors and indels (green) with consensus bases (orange) from the aligned reads. Finally, the workflow is used to perform circularization steps through self-alignment and trimming or assembly of endpoint contigs.

Table 1 |

Troubleshooting table

Step	Problem	Possible reason	Solution
7	The 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 min. If the pellet does not dislodge through inversion, gently dislodge the pellet with a pipette tip after the incubation is complete
9	The sample does not flow through the 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 or SPRI beads not properly resuspended before preparing aliquots	Ensure that 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, recombine the supernatant and eluate and attempt to size select again
17	DNA contamination levels are above recommended thresholds	Carryover of ethanol, phenol or isopropanol	Perform an additional SPRI bead selection by using a 1:1 ratio of beads to sample volume
18	DNA major peak <15 kb or high mass <2.5 kb	SPRI bead ratio too permissive	Perform an additional round of SPRI bead size selection by using a lower ratio of beads to sample volume (e.g., 0.75 volumes of bead suspension)
20	The sample pools on SpotOn port and does not enter the array	Sample entry relies on capillary action. If the SpotOn port is dry, the sample may not enter the 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 the priming port