Video Article

Efficient Nucleic Acid Extraction and 16S rRNA Gene Sequencing for Bacterial Community Characterization

Melis N. Anahtar[1], Brittany A. Bowman[1], Douglas S. Kwon[1]

[1]Ragon Institute of MGH, MIT, and Harvard, Massachusetts General Hospital

Correspondence to: Douglas S. Kwon at dkwon@mgh.harvard.edu

Abstract

There is a growing appreciation for the role of microbial communities as critical modulators of human health and disease. High throughput sequencing technologies have allowed for the rapid and efficient characterization of bacterial communities using 16S rRNA gene sequencing from a variety of sources. Although readily available tools for 16S rRNA sequence analysis have standardized computational workflows, sample processing for DNA extraction remains a continued source of variability across studies. Here we describe an efficient, robust, and cost effective method for extracting nucleic acid from swabs. We also delineate downstream methods for 16S rRNA gene sequencing, including generation of sequencing libraries, data quality control, and sequence analysis. The workflow can accommodate multiple samples types, including stool and swabs collected from a variety of anatomical locations and host species. Additionally, recovered DNA and RNA can be separated and used for other applications, including whole genome sequencing or RNA-seq. The method described allows for a common processing approach for multiple sample types and accommodates downstream analysis of genomic, metagenomic and transcriptional information.

Video Link

The video component of this article can be found at https://www.jove.com/video/53939/

Introduction

The human lower reproductive tract, gastrointestinal system, respiratory tract, and skin are colonized by complex bacterial communities that are critical for maintaining tissue homeostasis and supporting the health of the host[1]. For instance, certain lactobacilli create an inhospitable environment for pathogens by acidifying the vaginal vault, producing antimicrobial effectors and modulating local host immunity[2-4]. The growing appreciation for the bacterial microbiome's importance has also increased interest in characterizing bacterial communities in many clinical contexts. Here we describe a method to determine the composition of the bacterial microbiome from genital swabs. The protocol can be readily modified for stool samples and swabs collected from other anatomical locations and other host species.

Due to the inherent limitations in the number of samples that can be collected and stored from a given study participant, this protocol was designed to extract DNA, RNA, and potentially even protein from a single swab using an adapted phenol-chloroform based bead-beating method[5,6]. The combination of physical disruption of bacterial cell walls with bead-beating and chemical disruption with detergents allows rapid lysis of Gram-positive, Gram-negative, and acid-fast bacteria without additional enzymatic digestion steps. To obtain high quality RNA, it is recommended to use dry swabs that were kept at or below 4 °C immediately after collection and during transport to the laboratory (if applicable), and stored long-term at -80 °C.

To determine the bacterial microbiome within a given sample, this procedure utilizes 16S rRNA gene amplicon sequencing, which is currently the most cost-effective means to comprehensively assign bacterial taxonomy and perform relative quantification. Alternative methods include targeted qPCR[7], custom microarrays[8], and whole-genome sequencing[9]. The 16S rRNA gene contains nine hypervariable regions, and there is no consensus regarding the optimal V region to sequence for vaginal microbiome studies. Here we use the 515F/806R primer set and build on the pipeline designed by Caporaso *et al.*[10-12]. Caporaso *et al.*'s 515F/806R primer set enables multiplexing of hundreds of samples on a single sequencing run due to the availability of thousands of validated barcoded primers and compatibility with Illumina sequencing platforms. Unlike the Human Microbiome Project's 27F/338R primer set[13], 515F/806R also effectively amplifies *Bifidobacteriaceae* and thus accurately captures *Gardnerella vaginalis*, an important member of the vaginal microbial community in some women. Alternatively, a 338F/806R primer pair has been successfully used for pyrosequencing of vaginal samples[14] and a 515F/926R primer pair has recently become available for next-generation sequencing[12].

Finally, this protocol provides basic instructions to perform 16S amplicon analysis using the Quantitative Insights into Microbial Ecology (QIIME) software package[15]. Successful implementation of the QIIME commands described here yields a table containing bacterial taxonomic abundances for each sample. Many additional quality control steps, taxonomic classification methods, and analysis steps can be incorporated into the analysis, as described in detail on the QIIME website (http://qiime.org/index.html). If the analysis will be performed on an Apple

computer, the MacQIIME package[16] provides easy installation of QIIME and its dependencies. Alternative software packages for 16S rRNA gene sequence analysis include Mothur[17] and UPARSE[18].

Protocol

The study protocol was approved by and followed the guidelines of the Biomedical Research Ethics Committee of the University of KwaZulu-Natal (Durban, South Africa) and the Massachusetts General Hospital Institutional Review Board (2012P001812/MGH; Boston, MA).

1. Extraction of Total Nucleic Acid from Cervicovaginal Swabs

Note: Perform nucleic acid extractions in sets of 16 samples or fewer. The protocol as written below assumes samples are processed in sets of 12. If performing multiple rounds of extractions, serially number the extraction batches and record each sample's extraction batch number as well as other sample information (include metadata such as the participant's ID number, age, date/time of swab collection, hormonal contraceptive type, sexually transmitted infection testing results, *etc.*) in **Table 1**.

1. Preparation of reagents and fume hood
    1. Prepare a buffer comprised of 200 mM sodium chloride (NaCl), 200 mM Tris, and 20 mM edetic acid (EDTA) in 100 ml of nuclease-free water. Filter-sterilize the solution by passing it through a 0.22 µm filter. Chill an aliquot of 10 ml of buffer on wet ice.
    2. Adjust the pH of the phenol:chloroform:isoamyl alcohol (IAA) (25:24:1) to pH 7.9 by adding 65 µl of Tris alkaline buffer per 1 ml phenol, shaking the mixture for 2 min, and allowing the two phases to separate either naturally or by centrifugation at 10,000 x g for 5 min at RT.
    Caution: Phenol is toxic if swallowed, if inhaled, or in contact with skin and eyes. Do not breathe fumes. Wear impervious gloves, safety glasses with side-shields, and a lab coat.
    3. Filter-sterilize 25 ml of 20% sodium dodecyl sulfate (SDS) through a 0.22 µm filter. Make 5 ml aliquots of the sterilized SDS.
    4. Chill a 10 ml aliquot of isopropanol at -20 °C.
    5. Prepare a bead beating tube for each swab to be processed by weighing out 0.3 g of glass beads into a sterile 2 ml tube that is suitable for the bead beater.
    6. Obtain swabs by sampling the ectocervix with a sterile absorbent swab. Immediately after collection, place the swab into an empty and sterile cryovial, store at 4 °C for 1 to 4 hr during transport to the lab, and store for several months at -80 °C. Transfer the swabs (contained within individual tubes) to be processed to wet ice.
    7. Prepare the biological safety cabinet (BSC). Use a BSC with a "thimble" connected to the building exhaust to ensure proper removal of volatile chemicals.
        1. Remove all materials from the hood.
        2. Clean all surfaces of the hood with bleach, followed by a decontaminant that removes RNases, DNases, and DNA from surfaces. Clean all subsequent items brought into the hood using bleach followed by a nucleic acid decontaminant, including gloves. Use fresh RNase/DNase-free reagents, such as pipette tips, whenever possible.
        3. Tape a sterilized chemical biohazard bag to the rear of the hood. All dry waste containing phenol or chloroform should be placed into this bag for proper disposal.
        4. Place a sterile bottle into the hood to collect liquid waste containing phenol or chloroform.

2. Phenol-chloroform extraction.
    1. In the hood, to each bead beating tube, add 500 µl of buffer (from step 1.1.1), 210 µl of 20% sodium dodecyl sulfate, and 500 µl of phenol:chloroform:IAA (25:24:1, pH 7.9).
    2. Transfer the swab from the transport vial into the bead beating tube using a new pair of sterile forceps. Thoroughly rub the swab head against the internal walls of the bead beating tube for at least 30 sec. Re-cap the sample when done. If performing extractions from multiple swabs, change gloves between each sample.
    3. Chill the sample on ice for at least 10 min. Remove the swab from the bead beating tube by holding the swab handle with sterile tweezers while pressing the swab head against the internal tube wall using a clean P200 tip. Discard the swabs in the dry chemical waste bag. Note: The "squeegee" action (pressing the swab head) will liberate liquid from the absorbent swab and increase the nucleic acid recovery.
    4. Place the bead beating tube into the bead beater and homogenize for 2 min at 4 °C.
    5. Centrifuge the bead beating tube for 3 min at 6,000 x g *and* 4 °C to pellet debris and separate the aqueous and phenol phases.
    6. Transfer the aqueous phase (~ 500 - 600 µl) to a sterile 1.5 ml tube. Add an equal volume of phenol:chloroform:IAA. Mix by inversion and brief vortexing.
    7. Centrifuge the tube for 5 min at 16,000 x g and 4 °C.
    8. Transfer the aqueous phase to a new sterile 1.5 ml tube. Be conservative and do not transfer material from the interphase layer or the underlying phenol phase. Note the volume of the transferred aqueous phase. Save the phenol phase for future protein isolation.
    9. Add 0.8 volume of isopropanol and 0.1 volume of 3M sodium acetate (pH 5.5). Mix thoroughly by inversion and briefly vortexing.
    10. Precipitate the nucleic acid by chilling the tube at -20 °C for at least 2 hr (up to O/N).

3. Isopropanol precipitation and ethanol wash
    1. Centrifuge the tube for 30 min at approximately 16,000 x g and 4 °C. Carefully use a pipette to remove the supernatant, leaving the pellet intact.
    2. Add 500 µl of 100% ethanol. Dislodge the pellet with gentle vortexing or pipetting without touching the pellet. Centrifuge for 5 min at 16,000 x g and 4 °C.
    3. Carefully discard the ethanol supernatant. Use a P10 pipet to remove as much ethanol as possible without disturbing the pellet.
    4. Air dry the pellet at RT for 15 min.

5. Resuspend the pellet in 20 µl of ultra-pure 0.1x Tris-EDTA buffer. Allow the sample to chill on ice for 10 min and pipette repeatedly to ensure full resuspension. If the pellet does not dissolve, transfer the tube to a 40 °C heat block for up to 10 min to aid dissolution.
6. Measure the nucleic acid concentration using a spectrophotometer[19].
7. If desired, separate DNA from RNA using a column clean-up kit, following the manufacturer's protocol[20].
8. Store the nucleic acid at -80 °C or continue.

2. PCR Amplification of the 16S *rRNA* Gene V4 Hypervariable Region

Note: Perform the PCR amplification in sets of 12 samples or fewer to minimize the risk of contamination and human error. If performing multiple rounds of amplification, serially number the amplification batches and record each sample's amplification batch number in **Table 1**.

1. Preparation of the reagents and PCR hood
   1. Add the PCR amplification set information to **Table 1**, which will serve as the basis of the mapping file at the sequence analysis stage.
   2. Remove all materials from a PCR hood and clean the internal surfaces thoroughly with bleach followed by a decontaminant that removes RNases, DNases, and DNA. Be sure to decontaminate every reagent and piece of equipment (*e.g.,* pipettes) before placing them in the hood. Wear fresh gloves cleaned with a nucleic acid decontaminant prior to working in the hood.
   3. If necessary, dilute the nucleic acid template to 50 - 100 ng/µl using DNA-free and nuclease-free water.
   4. Thaw aliquots of the 5x high-fidelity (HF) buffer, dNTPs, and primers in the clean PCR hood. Gently vortex and centrifuge all solutions after thawing. To minimize freeze-thaw cycles and the risk of stock contamination, prepare aliquots of the 5x HF buffer, dNTPs, and primers.
   5. Place a clean benchtop cooler rack for microcentrifuge tubes and a PCR plate cooler into the hood.

2. For PCR reaction, prepare the master mix by combining 15.5 µl of ultra-pure water, 5 µl of 5x HF buffer, 0.5 µl of dNTPs, 0.5 µl of 515F forward primer, 0.75 µl of 3% DMSO, and 0.25 µl of Polymerase for each reaction. Assemble all reaction components in the cooler and add the polymerase last. Mix thoroughly by pipetting. Add two extra samples to the reaction count when preparing the master mix, to account for pipetting error.
3. PCR reaction setup:
   Note: Perform amplifications in triplicate, meaning each sample is amplified in three separate 25 µl reactions. Run a no-template water control with each primer pair. Work quickly but carefully, avoiding introduction of any contamination.
   1. Label an 8-well strip with individual caps and place into a PCR cooler.
   2. Pipette 90 µl of master mix into the first well.
   3. Add 2 µl of the reverse primer (**Supplemental File 1**). Be sure to carefully note the reverse primer barcode used with each sample in **Table 1**.
   4. Mix well and transfer 23 µl of master mix to the fourth well (the no-template control).
   5. Add 2 µl of water to the fourth well.
   6. Add 6 µl of the appropriate sample to the first well. Mix well and transfer 25 µl to the second well. Change tips and transfer another 25 µl from the first well to the third well. Firmly cap every well, making sure not to touch the inside of the wells or cap in the process.
   7. Repeat for each sample.

4. Perform PCR amplification
   1. Transfer the strip tubes to a thermocycler and run the following program: 30 sec at 98 °C, followed by 30 cycles of 10 sec at 98 °C, 30 sec at 57 °C, and 12 sec at 72 °C, followed by a 10 min hold at 72 °C and final hold at 4 °C.
   2. Perform the following steps on a clean lab bench. Quickly spin the tubes to collect liquid from the walls. Combine triplicate PCR reactions from each sample, with a total volume of 75 µl, into a sterile labeled tube. Also transfer 25 µl of each no-template control into a separate sterile tube. Do not combine amplicons from different samples yet.

5. Validation of successful PCR amplification of samples by gel electrophoresis.
   1. Prepare a 1.5% agarose gel (1.5 g agarose powder in 100 ml of 1x TAE buffer) with enough wells to hold each amplicon, water control, and ladder[21].
   2. While the gel hardens (about 30 min), prepare the sample for electrophoresis: Add 1 µl of 6x loading dye to a new, labeled tube. To that tube, add 5 µl of the amplicon and mix by pipetting.
   3. When the gel has set, remove the combs, place the gel in the electrophoresis tank, and fill the tank with 1x TAE buffer.
   4. To the first well, add 5 µl of DNA ladder.
   5. Load 5 µl of the sample amplicon to another well. Load 5 µl of the no-template amplicon to a separate well. Continue as needed for each sample.
   6. When all samples have been loaded, slide the tank lid in place and turn on the power source to 120 V. Allow the gel to run for 30 - 60 min.
   7. View the gel under UV light.
      1. Verify successful amplification of each sample by noting a single strong band around 380 bp. If there is a double band, re-amplify the sample with a different reverse barcode (Step 2.3). If there is no band at all, re-amplify the sample using either the same reverse barcode or a new reverse barcode (Step 2.3). If re-amplification is unsuccessful, PCR inhibitors may be present in the sample, in which case, perform a column-based DNA cleanup to remove PCR inhibitors.
         Note: Successful amplification may not be possible if the bacterial DNA concentration in the original sample is insufficient (<5 ng/µl).
      2. Verify the lack of reagent contamination by noting the absence of a band in the no-template control.

8. Store the remaining 70 µl of amplicon at -20 °C. Discard the remaining 20 µl of the no-template control, assuming it did not yield a band.

3. Library Pooling and High-Throughput Sequencing

1. Create the amplicon pool by combining an equal volume (2 - 5 µl) of each amplicon into a single sterile tube. If the band from a sample looked particularly weak, add twice the volume relative to the rest of the samples.
2. Remove the PCR primers from the amplicon pool using a PCR Clean-up kit, following the manufacturer's instructions[22]. Perform the clean-up with multiple columns if the amplicon pool volume is over 100 µl. Note: Each column has a 100 µl capacity.
    1. Store the library at -20 °C or proceed to the next step.
3. If applicable, combine the primer-free amplicon pools to create the final library. Determine the DNA concentration of the library using a spectrophotometer or a fluorometric system[23]. A 260/280 ratio between 1.8 - 2.0 is indicative of pure DNA.
4. Dilute the library to 20 nM. Confirm the quality of the library by visualizing a single band around 400 bp using an electrophoresis instrument. Confirm the concentration of the library using a fluorometric system[23].
5. Perform a final 1:10 dilution in water to dilute the library to 2 nM. Then, store the library at 20 °C indefinitely.
6. Send an aliquot of the final library with the three required sequencing primers (Read 1, Read 2, and Index; see Tables of Materials/ Equipment) to be sequenced on an Illumina sequencer. If fewer than 300 samples have been multiplexed for sequencing, use a single-end 300 bp run and with a 12 bp index read on a MiSeq, with a final library concentration of 5 pM and a 10% denatured PhiX spike-in. See the supplemental materials of Caporaso *et al.* ISME J, 2012[10] for detailed sequencing instructions.

4. Sequence Analysis

Note: Outlined here is a basic pipeline for sequence analysis using the QIIME 1.8.0 software package. For simplicity, the provided commands assume that the mapping file is called mapping.txt, the 12 bp index read file is called index.fastq, and the 300 bp sequencing read file is called sequences.fastq. Install QIIME or MacQIIME[16] and familiarize yourself with the basics of UNIX to execute these commands. Read the complete guide to QIIME at:

1. Complete the mapping file for the experiment (**Table 1**). Include as much metadata as possible. Note which samples have been extracted or amplified in the same batch, to determine whether there are batch effects.
2. Save the mapping file as a text file, *e.g.,* mapping.txt. Validate the formatting of the mapping file by executing the following command:validate_mapping_file.py -m mapping.txt -o mapping_output
Note: This command uses the built-in "validate_mapping_file.py" QIIME script that makes a new folder, called "mapping_output", containing an .html file indicating the mapping file errors, if any.
3. Check the quality of the sequencing reads using a high-throughput sequence data quality checking program, such as FastQC (http:// www.bioinformatics.babraham.ac.uk/projects/fastqc/). **Figure 5** demonstrates the per base sequence quality that can be expected from a successful run.
Note: The sequencer assigns each nucleotide base a Phred quality score, which corresponds to the probability that the base has been erroneously called. A Phred quality score of 10 indicates that there is a 10% chance that the nucleotide has been incorrectly assigned, 20 indicates a 1% chance, 30 indicates a 0.1% chance, and 40 (the highest possible score) indicates a 0.01% chance[24].
4. Using the mapping file as a key, demultiplex, quality filter the sequencing data, and save the results to a folder (in this case, called "sl_out") by executing this command[25]: split_libraries_fastq.py --rev_comp_mapping_barcodes -i sequences.fastq -o sl_out/ -b index.fastq -m mapping.txt -q 29
Note: The q flag denotes the maximum unacceptable Phred quality score, *e.g.,* "-q 29" filters out any sequences with Phred scores below 30, ensuring 99.9% accuracy of the base calls.
5. Using the Greengenes 16S operational taxonomic unit (OTU) reference database[26] (http://qiime.org/home_static/dataFiles.html), perform open-reference OTU picking by executing this command[27]: pick_open_reference_otus.py -i sl_out/seqs.fna -r 97_otus.fasta -o ucrss/ -s 0.1
Note: The -s flag indicates the fraction of sequences that failed to align to the reference database that will be included in the *de novo* clustering. "-s 0.1" includes 10% of the failed sequences in the *de novo* clustering. Use the -a flag to parallelize the OTU picking process and reduce the processing time from days to hours if multiple cores are available.
6. Create a user-friendly taxonomic abundance table by merging OTUs at the species level by executing this command[28]: summarize_taxa.py -i ucress/otu_table_mc2.biom -o summarized_otuSpecies/ -L 7
Note: The resulting table can be easily viewed in any spreadsheet software. Note that 16S rRNA sequencing does not reliably provide species level resolution.
7. Determine the ecological diversity within each sample by computing several alpha diversity metrics with the QIIME script alpha_diversity.py. Then, determine the diversity between pairs of samples using the QIIME script beta_diversity.py.
8. Visualize the data, *e.g.,* by using an EMPeror[29] principal coordinates plot or heatmap.
9. Perform formal statistical comparisons of mapping file categories, *e.g.,* with QIIME's compare_catagories.py script[30].

Representative Results

The general overview of the protocol, which enables the determination of relative bacterial abundances from a swab using 16S rRNA gene sequencing, is shown in **Figure 1**.The protocol has been optimized for human vaginal swabs, but can be easily adapted for most mucosal sampling sites and other hosts. **Figure 2** demonstrates the high-quality DNA and RNA that can be isolated using the bead-beating protocol. **Figure 3** illustrates a successful PCR amplification of 12 samples, where each amplification with a sample yielded a single strong band of the correct size and each water control did not yield a band. **Figure 4** illustrates the quantification of the final library pool prior to sequencing. **Figure 5** shows a typical sequence quality profile after a single-end 300 bp MiSeq run.
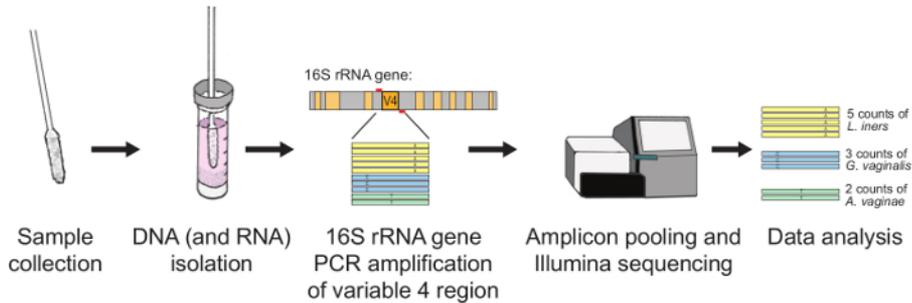
Figure 1. Schematic Overview of the Protocol. First, nucleic acid is extracted from a swab by bead-beating in a buffered solution containing phenol, chloroform, and isoamyl alcohol. Variable region 4 of the 16S rRNA gene is then amplified from the resulting nucleic acid using PCR. PCR amplicons from up to hundreds of samples are then combined and sequenced on a single run. The resulting sequences are matched to a reference database to determine relative bacterial abundances. The entire protocol can be performed in approximately three days. Please click here to view a larger version of this figure.
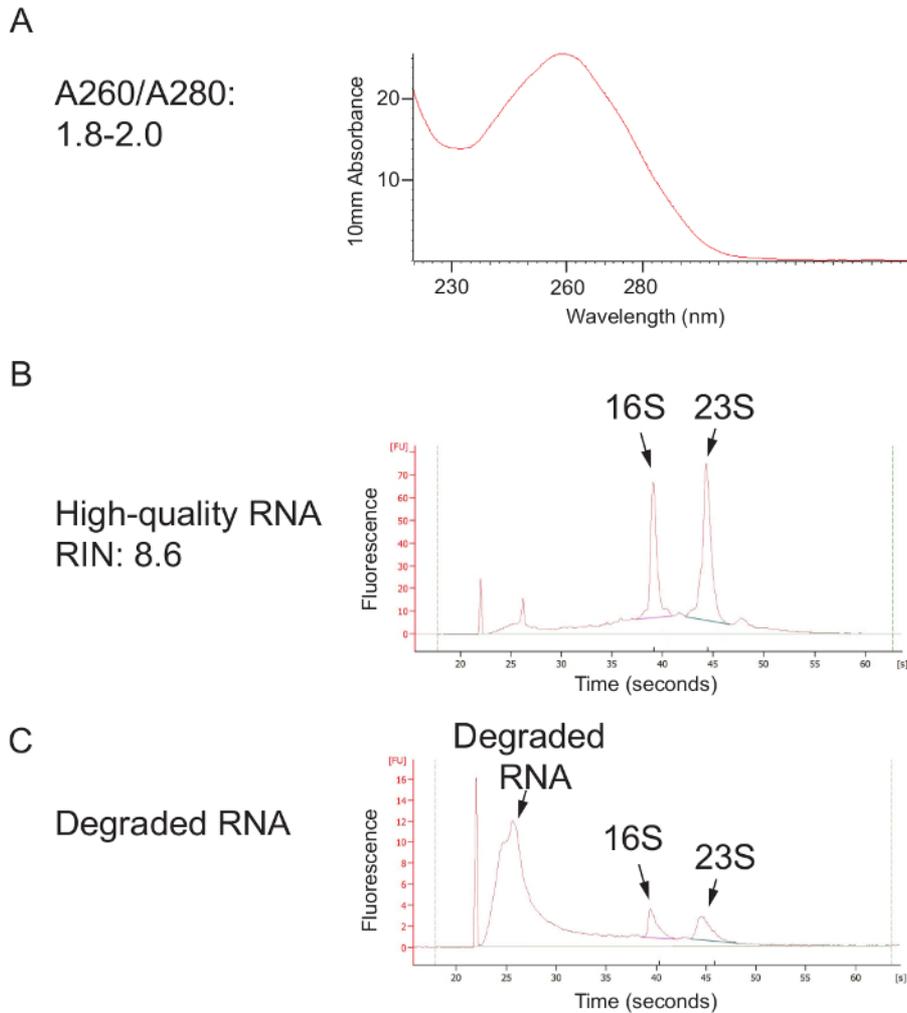
Figure 2. High-quality Nucleic Acid Extracted Using the Phenol:Chloroform Bead Beating Method. (A) DNA quality, as assessed using a spectrophotometer. An A260/A280 ratio between 1.8 and 2.0 indicates pure nucleic acid that is not contaminated with phenol or protein. **(B)** After a column clean-up, this protocol can yield high-quality RNA, indicated by strong 16S and 23S rRNA peaks. **(C)** RNA degradation can occur if the sample is not kept cold after collection (during transport and storage) or if RNases are present during processing. Please click here to view a larger version of this figure.
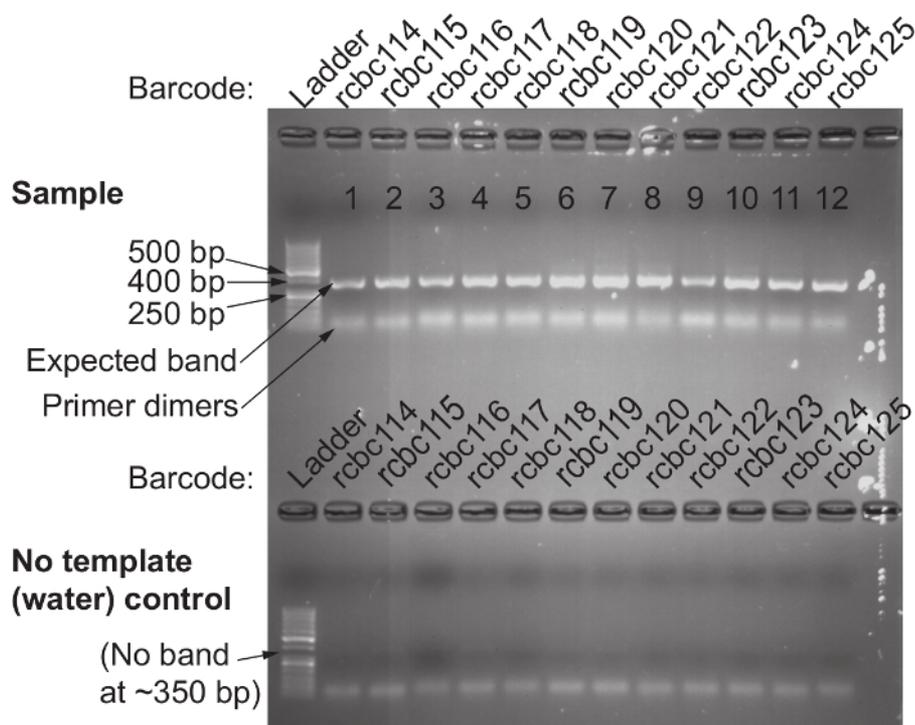
Figure 3. Confirmation of Successful 16S rRNA Gene Amplification Using the 515F and Barcoded 806R Primer Set. Top) Gel electrophoresis is used to confirm the presence of a single band around 380 base pairs in every sample that was amplified with template. The absence of a band indicates unsuccessful amplification; this is usually due to human error and the PCR reaction from that sample should be repeated. **Bottom**) No template (water) controls run in parallel with the same primer pair should not have a band present. The presence of a band in the water control indicates contaminated reagents; discard the reagents that may be contaminated and re-do the PCR amplifications of both the template and water control for that primer pair. Please click here to view a larger version of this figure.
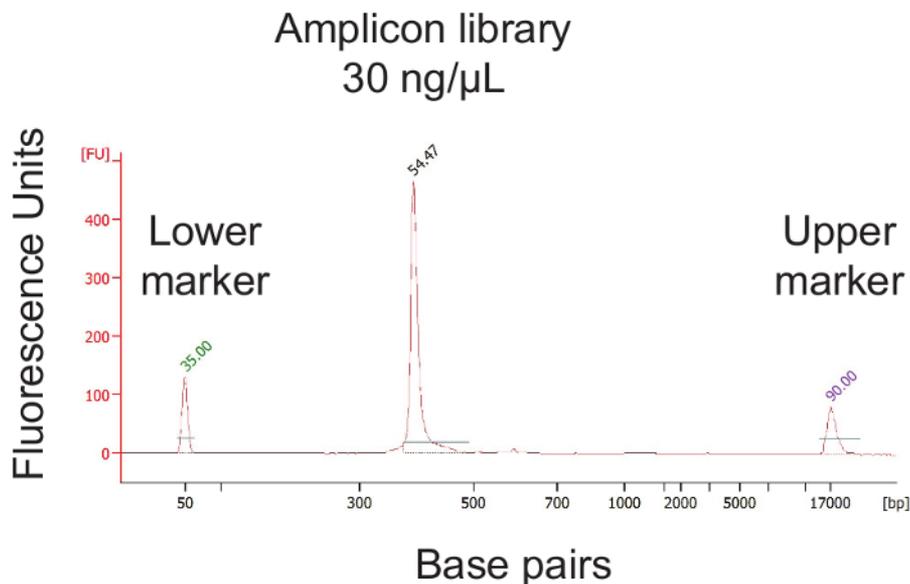
Figure 4. Quantification of the Final Library Pool Concentration and Validation of the Library Size. After pooling the individual sample amplicons, the concentration of the final library pool must be determined. The library pool must then be further diluted to achieve a 2 nM concentration. Please click here to view a larger version of this figure.
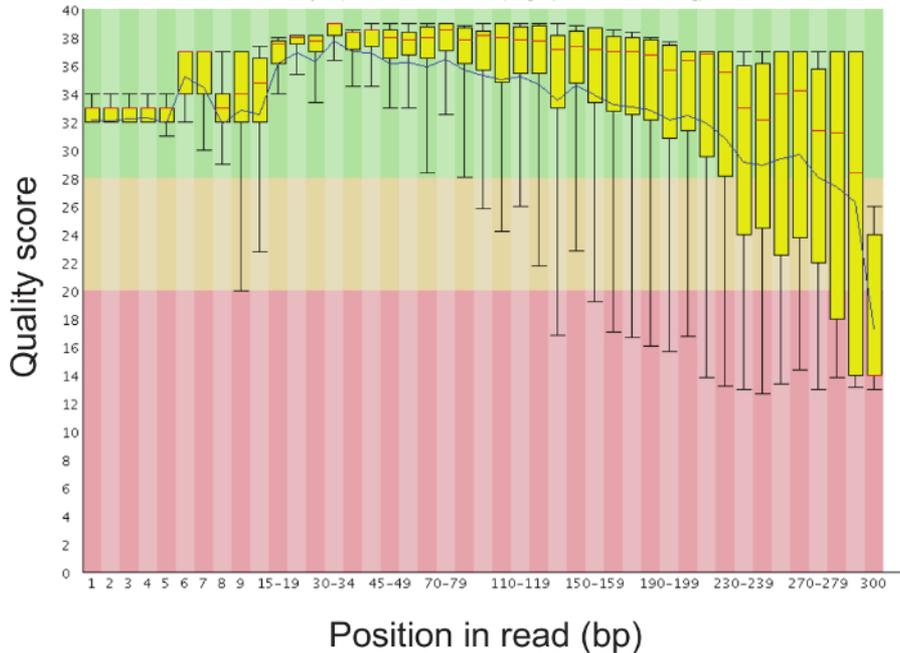
Figure 5. Representative Bar Plot of the Sequence Quality Scores at Each Position of the Read. It is normal for the sequence quality to drop after 200 base pairs, but the average quality score should remain above 30. Please click here to view a larger version of this figure.

#SampleID	Barcode Sequence	LinkerPrimer Sequence	rcbcPrimer	SampleType	Extraction Batch	Amplification Plate	Description
#An example mapping file can be found at: http://qiime.org/_static/Examples/File_Formats/Example_Mapping_File.txt							
AG2350	TCCCTTGTCTCC	CCGGACTACHVGGGTWTCTAAT	rcbc000	Cervical Swab	1	A	

Table 1. Mapping File Template. Creating an accurate and thorough mapping file is critical for successfully executing the protocol. The mapping file is not only required for executing QIIME, but it also enables the researcher to maintain the link between the sample barcode and metadata, to analyze the data for any systematic biases (*e.g.*, batch-to-batch variation), and to determine interesting correlations between the metadata and bacterial populations. A bare-bones mapping file is provided, but users are encouraged to add as many columns containing metadata as possible. Examples of additional metadata for a vaginal swab includes the participant's age, date/time of swab collection, hormonal contraceptive type (if applicable), sexually transmitted infection testing results, etc.

Supplemental File 1. List of Barcoded Reverse Primer Sequences[10]. The first three columns can be used to complete the mapping file, and the last column provides the entire primer sequence for ordering purposes. Please click here to download this file.

Discussion

Here we describe a protocol for the identification and characterization of relative bacterial abundances within a human vaginal swab. This protocol can easily be adapted for other sample types, such as stool and swabs of other body sites, and for samples collected from a wide variety of sources. The extraction of nucleic acid by bead-beating in a buffered solution of phenol and chloroform allows for isolation of both DNA and RNA, which is particularly important when working with precious samples collected through clinical studies. The isolated bacterial DNA is excellent for bacterial taxonomic identification and genomic assembly, while the simultaneous collection of RNA provides the opportunity to determine functional bacterial, host, and viral contributions through RNA-seq. The described protocol uses a validated one-step primer set that has been successfully deployed on a wide range of sample types, including human, canine, and environmental samples[10]. The availability of thousands of barcoded primers enables multiplexing of samples and tremendous savings on sequencing costs. The complete cost (including all reagents, a single sequencing run, and primers but not equipment) is about $20 per sample when 200 samples are multiplexed. Additionally, there is very high reproducibility when multiple swabs from the same sample site are processed independently through the entire pipeline. Overall, the protocol is cost efficient, flexible, reliable, and repeatable.

The nucleic acid extraction portion of this protocol is limited by the safety precautions required when working with phenol and chloroform, and the challenges of automating the pipeline to a high-throughput, 96-well plate format. Additionally, the vigorous bead beating used for mechanical lysis shears the bacterial DNA to approximately 6 kilobase fragments; if longer DNA fragments are required for downstream applications, the duration of bead beating should be shortened. The limitations of the bacterial identification portion of this protocol are inherent to any method that relies on 16S rRNA gene sequencing. 16S rRNA sequencing is ideal for bacterial identification to the genus and even species level, but rarely provides strain level identification. While the V4 variable region of the 16S rRNA gene provides robust discrimination amongst most bacterial species[11], additional computational methods such as Oligotyping[31] may need to be used to precisely identify certain species, such as *Lactobacillus crispatus*. Finally, information about the precise bacterial functional capabilities within a particular sample cannot be determined by 16S rRNA gene sequencing alone, though this protocol enables extraction of whole genome DNA and RNA that can be used towards this purpose.

The most critical step to ensuring success with this protocol is taking great care to prevent contamination during sample collection, nucleic acid extraction, and PCR amplification. Ensure sterility at the time of sample collection by wearing clean gloves and using sterile swabs, tubes, and scissors. To assess for contamination of the collection materials, collect negative control swabs by placing additional unused swabs directly into transport tubes at the time of sampling. In the lab, perform all pre-amplification steps in a sterilized hood containing only decontaminated supplies and using only molecular grade, DNA-free reagents. During nucleic acid extraction, prevent cross-contamination by using new sterile forceps and fresh gloves with each sample, and keeping all tubes closed unless in use. Processing unused swabs in parallel ensures sterility of both the sample collection and nucleic acid extraction; the unused swabs should not yield a pellet after isopropanol precipitation and ethanol washing. If a pellet does appear, perform 16S rRNA gene amplification to determine a possible source of the contamination (*e.g.,* the presence of *Streptococcus* or *Staphylococcus* would indicate skin contamination). Additionally, perform PCR amplifications with no template control reactions in parallel to ensure that the PCR reagents and reactions have not been contaminated. If a band appears in a no template control, discard the reagents and repeat the amplification with fresh reagents. Taking these precautions will ensure successful sequencing of the bacteria of interest.

The PCR amplification step tends to require the most troubleshooting. Amplifying in sets of twelve samples provides a balance between efficiency and consistency. The complete absence of bands across all samples in a given amplification set indicates a systematic failure, *e.g.,* forgetting to add a reagent or incorrectly programming the thermocycler. The absence of a band from a few samples is usually due to human error, and the amplifications should be re-run with the same pairing of sample and reverse primer. In the case of continued absence of a band, the sample can be re-amplified using a reverse primer with a different barcode. Repeated amplification failures with multiple reverse primers may indicate an inhibitor present in the sample. In that case, cleaning the DNA with a column will often remove inhibitors without significantly altering relative bacterial abundances. If multiple bands result after amplification, re-amplify the sample with a different reverse primer barcode.

In addition to preventing environmental contamination and ensuring amplification of a single specific product, successful sequencing relies on care when preparing the library pool. The goal is to combine equimolar amounts of each sample's amplicons to ensure approximately the same number of sequencing reads per sample. If the nucleic acid concentrations prior to amplification are comparable, simply adding equal volumes of each sample's amplicons is sufficient when creating the library pool. However, if the nucleic acid concentrations are vastly different and added in equal volume, the sample with the low nucleic acid concentration will be poorly represented with a low number of reads. In this case, it is possible to add a higher volume of the amplicons from the low concentration sample based upon the relative intensity of the gel band. Alternatively, it is possible to more rigorously remove primers from the individual amplicons, quantify individual sample's amplicon concentration using a fluorometric dsDNA quantification kit, and precisely combine equimolar amounts of each sample.

Once a well-balanced amplicon pool is generated, it becomes critical to carefully measure the pool's concentration. Subsequent careful dilution and spike-in with PhiX to increase the read complexity is critical for achieving optimal sequencing results. High-throughput sequencers that use sequencing by synthesis are very sensitive to the cluster density on the flow cell. Loading a library pool that is too concentrated will result in overclustering, with lower quality scores, lower data output, and inaccurate demultiplexing[32]. Loading a library pool that is too dilute will also result in low data output. Carefully quantifying the library pool prior to sequencing will ensure optimal results.

16S rRNA gene sequencing provides a comprehensive assessment of the bacteria present within a given sample and is an absolutely critical first step in hypothesis generation. The presence of a rich set of metadata further enables the researcher to test associations between particular bacterial species and important biological factors. Furthermore, the same 16S information can be used to infer the bacterial functions using with tools such as PICRUSt[33]. The ultimate goal is to use 16S characterization to identify novel associations that can be further tested and validated in model systems, adding to our growing understanding of the impact of the bacterial microbiome on human health and disease.

Disclosures

The authors have nothing to disclose.

Acknowledgements

References

1. Huttenhower, C. Structure, function and diversity of the healthy human microbiome. *Nature*. **486**, 207-214 (2012).
2. O'Hanlon, D. E., Moench, T. R., & Cone, R. A. Vaginal pH and microbicidal lactic acid when lactobacilli dominate the microbiota. *PloS one*. **8**, e80074 (2013).
3. Aldunate, M. *et al.* Vaginal concentrations of lactic acid potently inactivate HIV. *The Journal of antimicrobial chemotherapy* **68**, 2015-2025 (2013).
4. Anahtar, M. N. *et al.* Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract. *Immunity* **42**, 965-976 (2015).
5. Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L., & Gordon, J. I. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proceedings of the National Academy of Sciences of the United States of America*. **110**, 20236-20241 (2013).
6. Chomczynski, P., & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*. **162**, 156-159 (1987).
7. Srinivasan, S. *et al.* Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. *PloS one* **5**, e10197 (2010).
8. Dols, J. A. *et al.* Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. *American journal of obstetrics and gynecology* **204**, 305 e301-307 (2011).
9. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **9**, 811-814 (2012).
10. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal* **6**, 1621-1624 (2012).
11. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1**, 4516-4522 (2011).
12. Earthmicrobiome Project. *16S rRNA Amplification Protocol.* (2015).
13. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1**, 4680-4687 (2011).
14. Srinivasan, S. *et al.* Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PloS one* **7**, e37818 (2012).
15. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335-336 (2010).
16. Werner, J. *MacQIIME.* Source: http://www.wernerlab.org/software/macqiime (2015).
17. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537-7541 (2009).
18. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods*. **10**, 996-998 (2013).
19. Thermo Fisher Scientific. *NanoDrop 2000/2000c Spectrophotometer, V1.0 User Manual.* Source: http://www.thermoscientific.com/content/dam/tfs/ATG/CAD/CAD Documents/Product Manuals & Specifications/Molecular Spectroscopy/UV Visible Spectrophotometers/Spectrophotometer Systems/NanoDrop/NanoDrop-2000-User-Manual-EN.pdf (2009).
20. Qiagen. *AllPrep DNA/RNA Mini Kit.* Source: https://http://www.qiagen.com/us/shop/sample-technologies/rna-sample-technologies/dna-rna-protein/allprep-dnarna-mini-kit/ - orderinginformation (2015).
21. Lee, P. Y., Costumbrado, J., Hsu, C. Y., & Kim, Y. H. Agarose gel electrophoresis for the separation of DNA fragments. *Journal of visualized experiments: JoVE.* (2012).
22. MoBio. *UltraClean PCR Clean-Up Kit Instruction Manual.* Source: http://www.mobio.com/images/custom/file/12500(1).pdf (2013).
23. Invitrogen. *Quant-iT PicoGreen dsDNA Reagent.* Source: https://tools.thermofisher.com/content/sfs/manuals/mp07581.pdf (2008).
24. Illumina. *Quality Scores for Next-Generation Sequencing.* Source: <http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf> (2011).
25. QIIME. *split_libraries_fastq.py.* Source: http://qiime.org/scripts/split_libraries_fastq.html (2015).
26. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* **72**, 5069-5072 (2006).
27. QIIME. *pick_open_reference_otus.py.* Source: http://qiime.org/scripts/pick_open_reference_otus.html (2015).
28. QIIME. *summarize_taxa.py.* Source: http://qiime.org/scripts/summarize_taxa.html (2015).
29. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A., & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience.* **2**, 16 (2013).
30. QIIME. *Comparing categories.* Source: http://qiime.org/tutorials/category_comparison.html (2015).
31. Eren, A. M. *et al.* Exploring the diversity of Gardnerella vaginalis in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PloS one* **6**, e26732 (2011).

32. Illumina. *Diagnosing and preventing flow cell overclustering on the MiSeq system.* Source: http://support.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf (2015).

33. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology* **31**, 814-821 (2013).