Journal of Visualized Experiments

Evaluating the Impact of Hydraulic Fracturing on Stream Using Microbial Molecular Signatures

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE61904R3
Full Title:	Evaluating the Impact of Hydraulic Fracturing on Stream Using Microbial Molecular Signatures
Corresponding Author:	Jeremy Ryan Chen See Juniata College Huntingdon, Pennsylvania UNITED STATES
Corresponding Author's Institution:	Juniata College
Corresponding Author E-Mail:	chensej@juniata.edu
Order of Authors:	Jeremy Ryan Chen See
	Olivia Wright
	Lavinia Unverdorben
	Nathan Heibeck
	Regina Lamendella
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Open Access (US$4,200)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Huntingdon, PA, United States
Please confirm that you have read and agree to the terms and conditions of the author license agreement that applies below:	I agree to the Author License Agreement
Please specify the section of the submitted manuscript.	Environment
Please provide any comments to the journal here.	

TITLE:

Evaluating the Impact of Hydraulic Fracturing on Stream Using Microbial Molecular Signatures

AUTHORS & AFFILIATIONS:

Jeremy R. Chen See[1,2], Olivia Wright[1], Lavinia V. Unverdorben[1,2], Nathan Heibeck[1], Regina Lamendella[1,2]

[1]Department of Biology, Juniata College, Huntingdon, PA, USA
[2]Wright Labs, LLC, Huntingdon, PA, USA

Corresponding Author:
Regina Lamendella (lamendella@juniata.edu)

Email Addresses of Co-Authors:
Jeremy R. Chen See (chensej@juniata.edu)
Olivia Wright (wrighog18@juniata.edu)
Lavinia V. Unverdorben (unverlv16@juniata.edu)
Nathan Heibeck (heibens18@juniata.edu)

KEYWORDS:

hydraulic fracturing, natural gas, sampling, bacteria, streams, genetic analysis

SUMMARY:

Here, we present a protocol to investigate the impacts of hydraulic fracturing on nearby streams by analyzing their water and sediment microbial communities.

ABSTRACT:

Hydraulic fracturing (HF), commonly called "fracking", uses a mixture of high-pressure water, sand, and chemicals to fracture rocks, releasing oil and gas. This process revolutionized the U.S. energy industry, as it gives access to resources that were previously unobtainable and now produces two-thirds of the total natural gas in the United States. Although fracking has had a positive impact on the U.S. economy, several studies have highlighted its detrimental environmental effects. Of particular concern is the effect of fracking on headwater streams, which are especially important due to their disproportionately large impact on the health of the entire watershed. The bacteria within those streams can be used as indicators of stream health, as the bacteria present and their abundance in a disturbed stream would be expected to differ from those in an otherwise comparable but undisturbed stream. Therefore, this protocol aims to use the bacterial community to determine if streams have been impacted by fracking. To this end, sediment, and water samples, from streams near fracking (potentially impacted) and upstream or in a different watershed of fracking activity (unimpacted) must be collected. Those samples are then subjected to nucleic acid extraction, library preparation, and sequencing to investigate microbial community composition. Correlational analysis and machine learning models can subsequently be employed to identify which features are explanative of variation in the community, as well as identification of predictive biomarkers for fracking's impact. These

methods can reveal a variety of differences in the microbial communities among headwater streams, based on the proximity to fracking, and serve as a foundation for future investigations on the environmental impact of fracking activities.

INTRODUCTION:

Hydraulic fracturing (HF), or "fracking", is a method of natural gas extraction, which has become increasingly prevalent as the demand for fossil fuels continues to rise. This technique consists of using high-powered drilling equipment to inject a blend of water, sand, and chemicals into methane-rich shale deposits, usually to release trapped gasses[1].

Because these unconventional harvesting techniques are relatively new, it is important to investigate the effects of such practices on nearby waterways. Fracking activities mandate the clearing of large swaths of land for equipment transportation and well pad construction. Approximately 1.2-1.7 hectares of land must be cleared for each well pad[2], potentially impacting runoff and water quality of the system[3]. There is a lack of transparency surrounding the exact chemical composition of fracking fluid, including what biocides are used. Additionally, fracking wastewater tends to be highly saline[2]. Furthermore, the wastewater may contain metals and naturally occurring radioactive substances[2]. Therefore, the possibility of leaks and spills of fracking fluid due to human error or equipment malfunction is concerning.

Stream ecosystems are known to be very sensitive to changes in surrounding landscapes[4] and are important for maintaining biodiversity[5] and proper nutrient cycling[6] within the entire watershed. Microbes are the most abundant organisms in freshwater streams and thus, are essential to nutrient cycling, biodegradation, and primary production. Microbial community composition and function serve as great tools to gain information on the ecosystem due to their sensitivity to perturbance, and recent research has shown distinct shifts in observed bacterial assemblages based on proximity to fracking activity[7,8]. For example, *Beijerinckia*, *Burkholderia*, and *Methanobacterium* were identified as enriched in streams near fracking while *Pseudonocardia*, *Nitrospira*, and *Rhodobacter* were enriched in the streams not near fracking[7].

Next generation sequencing of the 16S ribosomal RNA (rRNA) gene is an affordable method of determining bacterial community composition that is faster and cheaper than whole genome sequencing approaches[9]. A common practice within the field of molecular ecology is to use the highly variable V4 region of the 16S rRNA gene for sequencing resolution, often down to the genus level with a wide scope of identification[9], as it is ideal for unpredictable environmental samples. This technique has been implemented widely in published studies and has been successfully utilized to identify the impact of fracking operations on aquatic environments[7,8]. However, it is worth noting that bacteria have varying copy numbers of the 16S rRNA gene, which affects their detected abundances[10]. There are a few tools to account for this, but their efficacy is questionable[10]. Another practice that is quickly growing in prevalence and lacks this weakness is metatranscriptomic sequencing, in which all RNA is sequenced, allowing researchers to identify both active bacteria and their genes expression.

Therefore, in contrast to methods in previously published studies[7,8,11,12], this protocol also covers sample collection, preservation, processing, and analysis for investigating microbial community function (metatranscriptomics). The steps detailed herein allow researchers to see what impact, if any, fracking has had on the genes and pathways expressed by microbes in their streams, including antimicrobial resistance genes. Moreover, the level of detail presented for sample collection is greater. Although several of the steps and notes may seem obvious to experienced researchers, they could be invaluable to those just starting research.

Herein, we describe methods for sample collection and processing to generate bacterial genetic data as a means to investigate the impact of fracking on nearby streams based on our labs' several years of experience. These data can be used in downstream applications to identify differences corresponding to fracking status.

PROTOCOL:

1. Collection of sediment samples for nucleic acid extraction

1.1. Submerge a sterile 50 mL conical tube into the stream water. Wear gloves during sample collection to avoid introducing unwanted human contamination. Perform this step either from the shore or facing upstream if in the water.

1.2. While the conical tube is submerged, remove the cap, and use it to scoop approximately 3 mL of sediment from a depth of 1 to 3 cm into the conical tube.

1.3. Remove the conical tube from the water and dump out all water, except for a thin layer covering the sediment sample (approximately 1 mL).

1.4. Using a 1000 µL pipette and appropriate pipette tips, add 3 mL of DNA/RNA preservative (see **Table of Materials** for the preservative specifications) to the collected sample. Keep the pipette tips in a sterile pipette tip box and only attach them immediately before use and discarded after use. Swirl the capped conical tube for 5 s to ensure the preservative and sample are thoroughly mixed.

NOTE**:** Step 1.4 is not necessary, but it is strongly recommended if RNA is to be extracted from the sediments later.

1.5. Place the samples on ice for the rest of sample collection. Upon returning from collection, store in a freezer at -20 ˚C if the samples are to be used for 16S analysis (DNA), or -70 ˚C, if they are to be used for metatranscriptomics analysis (RNA).

2. Filter collection for nucleic acid extraction

2.1. Remove the cap of a sterile 1 L bottle. While facing upstream or from the shore, fill the bottle with stream water to the top and then dump it out. Repeat this process two more times to condition the bottle. Fill the entire bottle a fourth time and cap it.

NOTE: If reusing a 1 L bottle, it can be sterilized by rinsing with 10% bleach for 2 min, followed by rinsing three times with deionized water and then once with 70% ethanol, and finally autoclaving with settings: 30 min exposure time at 121.1 ˚C and 15 min drying time. During autoclaving, the cap on the bottle should be very loose to avoid the bottle being compressed in the process.

2.2. Once on a stable surface, use a sterile Luer lock syringe and draw up a full volume. Then connect the syringe to a sterile and DNA/RNA-free 1.7 cm diameter polyethersulfone filter with a pore size of 0.22 µm and push the entire volume through the filter by pressing the plunger all the way down. Repeat this process until the total volume collected in the bottle (1 L) is pushed through the filter.

NOTE: The volume of the syringe can be variable, if, the total amount of water pushed through the filter is tracked. However, generally, 60 mL is preferred. While 1 L is ideal, anecdotally, a volume of at least 200 mL would likely still collect enough biomass (assuming ~20,000 cells per mL) for the extraction of DNA and RNA.

2.3. Remove excess water from the filter by drawing up roughly 20 mL worth of air into the syringe and pushing it through the filter.

NOTE: This will help prevent loss of the preservative if step 2.4 is performed.

2.4. Using a P1000 micropipette, add 2 mL of a DNA/RNA preservative by discharging it through the filter's larger opening (where it was attached to the syringe) while holding the filter horizontally. The tip of the pipette should be within the barrel of the filter when the pipette is depressed to ensure the preservative enters the filter. Change the tip after each use.

NOTE: As with the sediment collection, this step is not necessary, but it is strongly recommended for increased nucleic acid yield later, especially for RNA.

2.5. Peel off one square of paraffin film and wrap it tightly around each opening/end of the filter to seal. Place the paraffin film wrapped filter into a sterile sample bag and then place the entire bag on ice during collection.

NOTE: Ensure that the side used to wrap the filter is sterile, i.e., not previously exposed to the environment.

2.6. Upon return from sampling, store filters at -20 ˚C for 16S or -70 ˚C for meta-transcriptomics.

3. Nucleic acid extraction and quantification

3.1. Clean the work area with 10% Bleach and 70% Ethanol before beginning sample transfer.

3.2. For sediment (from step 1.5), generally, use ~0.25 g of sample. Flame sterilize a metal tool by dipping it in a beaker of 70% ethanol and burning the ethanol off between samples.

3.3. For filters (from step 2.6), move the filter paper into a sterile tube for extraction. To do so follow the steps below.

3.3.1. Create a sterile, DNA and RNA free-surface by folding aluminum foil so that the inner part of the fold is not exposed to the outside environment and autoclaving the folded piece with the settings: 121.1 ˚C and 5 min drying time.

3.3.2. Sterilize a vise-grip with 70% ethanol and an open flame. Then use the vise-grip to break open the filter casing on the sterile surface and remove the core from the casing.

3.3.3. Use a sterile scalpel to cut the filter paper away from the core by slicing at the top and bottom and then along the seam. Fold the filter paper using sterile tweezers and then cut the filter into small pieces using the scalpel.

3.3.4. Place the filter pieces in a microcentrifuge tube for extraction. Make sure that the filter paper does not come into contact with any surfaces which are not sterilized or that could have nucleic acid present, as this would lead to unwanted contamination of the sample.

3.4. Perform DNA isolation as described previously[13] or by using a commercially available column-based kit (see **Table of Materials**). The steps for the commercial kit listed are briefly described below.

3.4.1. Lyse the cells within the sample by transferring it to a bead tube and subjecting it to a cell disruptor at high speed for at least 5 min. Centrifuge and transfer the supernatant to a sterile microcentrifuge tube.

3.4.2. Add lysis buffer to the supernatant (1:1 volume) and transfer to the provided filter (yellow). Centrifuge the filter.

3.4.3. Transfer the filter to a new sterile microcentrifuge tube. Add the preparation buffer (400 µL), centrifuge, and discard the flow through.

3.4.4. Add wash buffer (700 µL), centrifuge, and discard the flow through. Then add wash buffer (400 µL), centrifuge, and discard the flow through again.

3.4.5. Transfer the filter to a new sterile microcentrifuge tube. Elute with 50 µL of DNase/RNase free water and let sit for 5 min at room temperature before centrifuging.

3.4.6. During that in cubation period, prepare the III-HRC filter by placing it in a collection tube and adding the HRC prep solution (600 µL) to it, followed by a centrifugation step of 3 min at 8,000 x *g*.

3.4.7. Move the prepared filter onto a sterile microcentrifuge tube. Transfer the eluted DNA from step 3.4.5 to this filter and centrifuge at 16,000 x *g* for 3 min. The flow through contains the extracted DNA.

3.5. Store DNA extracts for both sediments and filters at -20 ˚C.

NOTE: DNA extracts can be stored for around 8 years at -20 ˚C assuming stable temperature, limited light exposure, and no harmful contaminants[14].

3.6. Perform RNA isolation as per the manufacturer's protocol. Store RNA extracts at -80 ˚C.

3.6.1. Lyse the cells within the sample by transferring it to a bead tube and subjecting it to a cell disruptor at high speed for at least five minutes. Centrifuge and transfer the supernatant to a sterile microcentrifuge tube.

3.6.2. Add lysis buffer to the supernatant (1:1 volume) and transfer to the provided column (yellow). Centrifuge the column.

3.6.3. Add an equal volume of 95-100% ethanol to the flow through and mix by pipetting up and down five times.

3.6.4. Place the IICG Column (green) on a sterile microcentrifuge tube. Transfer the mixed solution to the column and centrifuge.

3.6.5. Add wash buffer (400 µL), centrifuge, and discard the flow through.

3.6.6. Add 5 µL of DNase I and 75 µL of DNA digestion buffer to the column and incubate at room temperature for 15 minutes.

3.6.7. Add prep buffer (400 µL), centrifuge, and discard the flow through.

3.6.8. Add wash buffer (700 µL), centrifuge, and discard the flow through. Then add wash buffer (400 µL), centrifuge, and discard the flow through again.

3.6.9. Transfer the column to a new sterile microcentrifuge tube. Elute with 50 µL of DNase/RNase free water and let sit for 5 min before centrifuging.

3.6.10. During that incubation period, prepare the III-HRC filter by placing it in a collection tube and adding the HRC prep solution (600 µL) to it, followed by a centrifugation step of 3 min at 8,000 x *g*.

3.6.11. Move the prepared filter onto a sterile microcentrifuge tube. Transfer the eluted DNA from step 3.4.5 to this filter and centrifuge at 16,000 x *g* for 3 min. The flow through contains the extracted RNA.

NOTE: RNA extracts can only be stored for one year before they start to degrade[15]. Both DNA and RNA extracts are degraded by repeated freeze-thawing. Some protocols allow for the extraction of both DNA and RNA from the same sample[16,17].

3.7. Quantify the extracted DNA and RNA samples using a fluorometer or a spectrophotometer. See **Table 1** for example fluorometer DNA concentration values. For an example spectrophotometer quantification protocol, see reference[18]. Sediment DNA concentration values with the kit listed in **Table of Materials** generally range from 1 to 40 ng/µL, while filter DNA concentration values tend to range from 0.5 to 10 ng/µL. Sediment RNA concentration values with the kit listed in **Table of Materials** generally range from around 1 to 20 ng/µL, while filter RNA concentration values tend to be lower, typically ranging from 0.5 to 5 ng/µL.

4. DNA 16S rRNA library creation

4.1. Clean the work area with 10% Bleach and 70% Ethanol. The work area should be an enclosed space capable of producing laminar flow conditions (laminar flow hood).

4.2. Use the DNA extracts (from step 3.5) and prepare samples for 16S rRNA amplicon sequencing with a standard PCR protocol, such as the one described on the Earth Microbiome's website that amplifies the V4 hypervariable region of 16S rRNA[19] under laminar flow conditions.

4.3. Prepare a 2% agarose gel as described previously and let it solidify[17]. Mix 7 µL of PCR product and 13 µL of DNase free water. Add a gel loading dye to a final concentration of 1x. Once agarose is solidified, load this PCR products mix on a 2% agarose gel.

NOTE: Alternatively, a pre-cast gel can be used instead, as these gels run faster and come pre-made.

4.4. Run the gel at 90 V for 60-90 min to check for the band size of 386 as successful amplification for 16S rRNA V4 amplicons, using the Earth Microbiome's protocol.

5. DNA 16S rRNA library purification

5.1. Pool 10 µL of PCR products for the samples that yielded bright bands and 13 µL for the samples that yielded faint bands in an appropriately sized sterile microcentrifuge tube.

5.2. Check the concentration of the resulting pool using a fluorometer or spectrophotometer and prepare a 2% agarose gel as before. Ideally, the pool should have a concentration of at least 10 ng/µL, and most samples should have had a concentration of around 25 ng/µL.

5.3. Concentration and volume permitting, load around 150-200 ng in a well of 2% agarose gel.

5.4. Run the gel for 60-90 min at 90 volts.

5.5. Purify the pooled library by running a 2% agarose gel.

5.5.1. Excise the 386 bp DNA band from the gel and purify the pooled library using a commercially available kit as described previously[20]. Elute the purified DNA with 30 µL of 10 mM Tris-Cl (pH 8.5). Perform this step in a different area than DNA or RNA extraction to prevent future contamination, as cutting the gel will spread PCR amplicons onto both the experimenter and the surrounding area.

5.6. Check the concentration of the purified pool using a fluorometer or spectrophotometer. If purification went well, its concentration should be at least half of the unpurified pool's. Generally, the final concentration should range from 5 to 20 ng/µL.

5.7. Send the purified libraries for next generation sequencing. Ensure that they are kept cold during transport by including dry ice in the shipping container.

6. RNA library creation and purification

6.1. Several commercial kits can be utilized to create RNA libraries. For whichever one is used, follow the manufacturer's protocol as written while working in a sterile laminar flow environment. A very summarized version of the protocol for kit in the **Table of Materials** is presented below[21].

6.1.1. Make the first strand cDNA synthesis master mix (8 µL of nuclease-free water and 2 µL of First Strand Synthesis Enyzme Mix) and add it to the sample. Place the sample in the thermocycler with the conditions specified in the protocol.

6.1.2. Make the second strand cDNA synthesis master mix (8 µL of Second Strand Synthesis Reaction Buffer, 4 µL Second Strand Synthesis Enzyme Mix, and 48 µL of nuclease-free water) on ice and add it to the sample. Place in a thermocycler set to 16 ˚C for one hour.

6.1.3. Purify the reaction by adding the provided beads (144 µL) and performing two 80% ethanol washes (200 µL).

6.1.4. Elute with the provided TE buffer (53 µL) and transfer 50 µL of the supernatant to a clean PCR tube. Place the PCR tube on ice.

6.1.5. Make the end prep master mix (7 µL of End Prep Reaction Buffer and 3 µL of End Prep Enzyme Mix) on ice and add it to the PCR tube. Place the PCR tube in a thermocycler with the conditions specified in the protocol.

6.1.6. Mix the Diluted Adaptor (2.5 µL), Ligation Master Mix (30 µL) and Ligation Enhancer (1 µL) solutions on ice. Add the mixed solutions to the sample and place in a thermocycler for 15 min at 20 ˚C.

6.1.7. Purify the reaction by adding the provided beads (87 µL) and performing ethanol washes (200 µL) and elution as before, except only add 17 µL of TE.

6.1.8. Add indices (10 µL) and the Q5 Master Mix (25 µL) solution and place in a thermocycler with the conditions described in the protocol.

6.1.9. Purify the reaction by adding the provided beads (45 µL) and performing an addition two ethanol washes (200 µL) and elute with 23 µL of TE. Transfer 20 µL to a clean PCR tube.

6.2. Check the libraries for detectable concentrations of RNA using a Bioanalyzer, fluorometer, or spectrophotometer.

6.3. Pool the metatranscriptomic libraries in a roughly equimolar ratio.

6.4. Purify the library following the same protocol for the 16S library purification, except excise fragments between 250 and 400 bp. Whereas the 16S library had a distinct band representing the amplified region, the result here is a smear.

6.5. Check the concentration of the purified library as before.

6.6. Ship the purified library with dry ice to a sequencing facility.

NOTE: Alternatively, RNA extracts can be sent to a university or private company for library preparation and sequencing.

7. Microbial community analysis

7.1. Once sequencing is complete, access the sample data. Download it to a usable computer.

NOTE: Ideally, the device should have at least 16 gigabytes of RAM. For a discussion of computing requirements (for Qiime2), see https://forum.qiime2.org/t/recommended-specifications-to-run-qiime2/9808.

7.2. Use the software to analyze 16S rRNA data, e.g, mothur, QIIME2, and R. See here https://docs.qiime2.org/2020.8/tutorials/moving-pictures/ for an example Qiime2 16S analysis tutorial.

7.3. For metatranscriptomics (RNA) data, use HUMAnN2 and ATLAS to determine which genes and pathways are present in the samples.

NOTE: An example metatranscriptomics pipeline culminating in diversity and random forest analysis is presented in the **Supplemental Information file**. All commands are run through command line, e.g., Terminal for Mac users.

REPRESENTATIVE RESULTS:

The success of DNA and RNA extractions can be evaluated using a variety of equipment and protocols. Generally, any detectable concentration of either is considered sufficient to conclude that the extraction was successful. Examining **Table 1** then, all extractions, except for one, would be dubbed successful. Failure at this step is often due to low initial biomass, poor sample preservation, or human error during extraction. In the case of filters, extraction may have been successful even if the concentration is below detection. If those extracts do not yield bands for PCR (if doing 16S) or a detectable concentration after library preparation (metatranscriptomics), then they likely did truly fail.

If the 16S protocol is followed, bright bands following PCR amplification, as seen in wells 4 and 6 in **Figure 1,** indicate success, while a lack of bands, as seen in the other wells in the top row, indicates failure. Moreover, a bright band in the gel lane that contains a negative PCR control would also indicate a failure since it would be risky to assume that the contamination impacting the negative control(s) did not affect the samples.

For both 16S and metatranscriptomics, the success of sequencing can be evaluated by looking at the number of sequences obtained (**Figure 2**). 16S samples should have a minimum of 1,000 sequences, with at least 5,000 being ideal (**Figure 2A**). Likewise, metatranscriptomics samples should have a minimum of 500,000 sequences, with at least 2,000,000 being ideal (**Figure 2B**). Samples with fewer sequences than those minimums should not be used for analyses, as they may not accurately represent their bacterial community. However, samples that fall between the minimum and ideal can still be used though results should be interpreted more cautiously if many samples fall in that range.

The success of subsequent downstream analysis can be determined simply on the basis of whether the expected output files were obtained or not. At any rate, programs, such as Qiime2 and R (**Figure 3**), should allow for the evaluation of potential significant differences among the bacterial communities based on fracking. The data for **Figure 3** was obtained by collecting sediment samples from twenty-one different sites at thirteen different streams for 16S and metatranscriptomics analysis. Of those twenty-one sites, twelve of them were downstream of fracking activity and classified as HF+, and nine of them were either upstream of fracking activity or in a watershed where fracking was not occurring; these streams were classified as HF-. Besides the presence of fracking activity, the streams were otherwise comparable.

Those differences could take the form of consistent compositional shifts based on fracking status. If that were the case, HF+ and HF- samples would be expected to cluster apart from each other in a PCoA plot, as is the case in **Figure 3A** and **Figure 3B**. To confirm that those apparent shifts are not just an artifact of the ordination method, further statistical analysis is needed. For

example, a PERMANOVA[22] test on the distance matrix that **Figure 3A** and **Figure 3B** are based on revealed significant clustering based on fracking status, meaning that the separation observed in the plot is consistent with differences among the samples' bacterial communities, instead of an artifact of ordination. A significant PERMANOVA or ANOSIM result is a strong indication of consistent differences between HF+ and HF- samples, which would indicate that the HF+ samples were impacted by fracking, while a high p-value would indicate that the samples were not impacted. Metatranscriptomic data can likewise be visualized and evaluated using the same methods.

Examining differential features (microbes or functions) can reveal evidence that samples have been impacted too. One method of determining differential features is to create a random forest model. The random forest model can be used to see how well the samples' fracking status can be correctly classified. If the model performs better than expected by chance, that would be additional evidence of differences dependent on fracking status. Moreover, the most important predictors would reveal which features were most important for correctly differentiating samples (**Figure 3C**). Those features also then would have had consistently different values based on fracking status. Once those differential features are determined, the literature can be reviewed to see if they have been previously associated with fracking. However, it may be challenging to find studies that determined differential functions, as most have only used 16S rRNA compositional data. Therefore, for evaluating the implications of differential functions, one possible method would be to see if they have been previously associated with potential resistance to biocides commonly used in fracking fluid or if they could aid in tolerating highly saline conditions. Furthermore, examining the functional profile of a taxon of interest could reveal evidence of fracking's impact (**Figure 3D**). For example, if a taxon is identified as differential by the random forest model, its antimicrobial resistance profile in HF+ samples could be compared to its profile in HF- samples and if they differ greatly, that could suggest that fracking fluid containing biocides entered the stream.

FIGURE AND TABLE LEGENDS:

Table 1: Example DNA concentrations based on Fluorometer 1x DS DNA high sensitivity assay. Extractions for all these samples, except for 14, would be considered successful due to having detectable amounts of DNA.

Figure 1: Example e-gel with PCR products. The gel was pre-stained and visualized under a UV light, causing any DNA present on it to glow. PCR worked for the samples in wells 4 and 6 in the first row, as they both had one single bright band of the expected size (based on the ladder). PCR for the samples in the other six wells failed, as they did not produce any bands. The positive control (first well, second row) had a bright band, indicating that PCR was performed properly, and the negative controls (wells 6 and 7, second row) did not have any bands, indicating that samples were not contaminated. If a negative had a band as bright as the samples, PCR would have been considered a failure since it would be risky to assume that the samples had amplicons that were not just the result of contamination.

Figure 2: Example sequence counts. (**A**) 16S example sequence counts. Nearly all these 16S samples had over 1,000 sequences. The very few that had less than 1,000 sequences should be excluded from downstream analyses, as they had insufficient sequences to accurately represent their bacterial communities. Several sequences had between 1,000 and 5,000 sequences; while not ideal, they would still be usable since they exceed the bare minimum, and the majority of samples exceed the ideal minimum of 5,000 as well. (**B**) Metatranscriptomics example counts. All samples exceeded both the minimum (500,000) and ideal minimum (2,000,000) number of sequences. Therefore, sequencing was successful for all of them, and they could all be used in downstream analysis.

Figure 3: Example analysis. (**A**) PCoA plot based on coordinates calculated with a Weighted Unifrac distance matrix created and visualized through Qiime2. (**B**) PCoA plot based on coordinates calculated with the Weighted Unifrac distance matrix exported from Qiime2. The coordinates were visualized using the Phyloseq and ggplot2 packages in R. Metadata vectors were fitted to the plot using the Vegan package. Each point represents a sample's bacterial community, with closer points indicating more similar community compositions. Clustering based on fracking status for these 16S sediment samples was observed (PERMANOVA, p=0.001). Furthermore, the vectors reveal that the HF+ samples tended to have higher levels of Barium, Bromide, Nickel, and Zinc, which corresponded to different bacterial community composition compared to the HF- samples. (**C**) Plot of best predictors for a random forest model that tested where bacterial abundances could be used to predict fracking status among the samples. The random forest model was created through R using the randomForest package. The top 20 predictors are shown as well as the resulting decreases in impurity (measure of the number of HF+ and HF- samples grouped together) in the form of Mean Decrease in Gini Index when they are utilized to separate samples. (**D**) Pie chart showing the antimicrobial resistance profile of the Burkholderiales profile based on metatranscriptomic data. Sequences were first annotated with Kraken2 to determine which taxa they belonged to. BLAST was then used with those annotated sequences and the MEGARes 2.0 database to determine which antimicrobial resistance genes (in the form of "MEG_#") were being actively expressed. Antimicrobial resistance genes expressed by members of Burkholderiales were then extracted to see which ones were most prevalent among that taxa. While more costly and time-consuming, metatranscriptomics does allow for functional analyses, such as this which cannot be done with 16S data. Notably, Kraken2 was used for this example analysis, instead of HUMAnN2. Kraken2 is faster than HUMAnN2; however, it only outputs compositional information, instead of composition, contribution, and functions (genes) and pathways like HUMAnN2 does.

Supplementary File: An example metatranscriptomics pipeline.

DISCUSSION:

The methods described in this paper have been developed and refined over the course of several studies published by our group between 2014 and 2018[7,8,10] and have been employed successfully in a collaborative project to investigate the impacts of fracking on aquatic communities in a three year project that will soon submit a paper for publication. These methods will continue to be utilized over the course of the remainder of the project. Additionally, other

current literature investigating the impact of fracking on streams and ecosystems describe similar methods for sample collection, processing, and analysis[7,8,10,11]. However, none of those papers utilized metatranscriptomic analysis, making this paper the first to describe how those analyses can be used to elucidate fracking's impact on nearby streams. Furthermore, the methods presented here for sample collection are more detailed, as are the steps taken to avoid contamination.

One of the most important steps of our protocol is initial sample collection and preservation. Field sampling and collection comes with certain challenges, as maintaining an aseptic or sterile environment during collection can be difficult. During this step, it is vital to avoid contaminating samples. To do this, gloves should be worn, and only sterile containers and tools should be allowed to come into contact with samples. Samples should also be immediately placed on ice after collection to mitigate nucleic acid degradation. Adding a commercial nucleic acid preservative upon collection can also increase nucleic acid yield and allow samples to be stored for longer periods of time after collection. Whenever nucleic acid extraction is performed, it is important to use the appropriate amount of sample, too much can clog spin filters used for extraction (for those protocols that make use of them) but too little can result in low yields. Be sure to follow the instructions for whichever kit is used.

Similar to field collection, avoiding or minimizing contamination is also important during nucleic acid extraction and sample preparation, especially when working with low nucleic acid yield samples, such as suboptimal sediment samples (samples containing a large amount of gravel or rocks) or water samples. Therefore, as with sample collection, gloves should be worn during all these steps to reduce contamination. Additionally, all work surfaces used during lab procedures should be sterilized beforehand by wiping with a 10% bleach solution, followed by a 70% ethanol solution. For pipetting steps (3-6), filter tips should be used to avoid contamination due to the pipette itself. All tools used for lab work, including pipettes, should be wiped down before and after with the bleach and ethanol solutions. Filter tips should also be used as an extra precaution to avoid contamination, with tips being changed every time they touch a non-sterile surface. To evaluate contamination, extraction blanks and negatives (sterile liquid) should be included during every set of nucleic acid extractions and PCR reactions. If quantification after extractions reveals a detectable amount of DNA/RNA in the negatives, extractions can be repeated if there is sufficient sample left. If negative samples for PCR show amplification, troubleshooting should be performed to determine the source and then the samples should rerun. To account for low levels of contamination, it is recommended that extraction blanks and PCR negatives be sequenced so that the contaminants can be identified and removed, if necessary, during computational analysis. Conversely, PCR amplification could also fail due to a variety of causes. For environmental samples, inhibition of the PCR reaction is often the culprit, which can be due to a variety of substances interfering with Taq polymerase[23]. If inhibition is suspected, PCR grade water (see **Table of Materials**) can be used to dilute the DNA extracts.

This protocol has a few notable limitations and potential difficulties. Sample collection can be challenging for both water and sediment samples. In order to get enough biomass, ideally 1 L of stream water needs to be pushed through a filter. The pores of the filter need to be small to

capture microbes but can also trap sediment. If a lot of sediment is in the water due to recent rainfall, the filter can clog making it difficult to push the entire volume through the filter. For sediment collection, it can be challenging to estimate the depth of sediment during collection. Furthermore, it is important to ensure that the sediment collected is predominantly soil, as pebbles and rocks will lead to lower nucleic acid yield and may not be an accurate representation of the microbial community. Lastly, it is vital as well that samples are kept on ice after collection, especially if a preservative is not used.

Though this protocol covers both metatranscriptomics and 16S lab protocols, it should be emphasized that these two methods are very different in both process and in the type of data they provide. The 16S rRNA gene is a commonly targeted region, highly conserved in bacteria and archaea, and useful for characterizing the bacterial community in a sample. Although a targeted and specific approach, species level resolution is often unattainable, and characterizing newly diverged species or strains is difficult. Contrarily, metatranscriptomics is a broader approach that captures all the active genes and microbes present within a sample. Whereas 16S provides only data for identification, metatranscriptomics can provide functional data such as expressed genes and metabolic pathways. Both are valuable and when combined, they can reveal which bacteria are present and which genes they are expressing.

This paper describes methods for field collection and sample processing for both 16S rRNA and metatranscriptomic analyses in the context of studying fracking. Additionally, it details collection methods for high quality DNA/RNA from low biomass samples and for long-term storage. The methods described here are the culmination of our experiences with sample collection and processing in our efforts to learn how fracking impacts nearby streams through examining the structure and function of their microbial communities. Microbes respond quickly to disturbances, and consequently, which microbes are present and the genes they express can provide information about the effects of fracking on ecosystems. Overall, these methods could be invaluable in our understanding of how fracking impacts these important ecosystems.

ACKNOWLEDGMENTS:

The authors would like to acknowledge the funding sources for the projects that led to the development of these methods, with those sources being: the Howard Hughes Medical Institute (http://www.hhmi.org) through the Precollege and Undergraduate Science Education Program, as well as by the National Science Foundation (http://www.nsf.gov) through NSF awards DBI-1248096 and CBET-1805549.

DISCLOSURES:

The authors have nothing to disclose.

REFERENCES

1. US EPA. The process of unconventional natural gas production. *US EPA*. <https://www.epa.gov/uog/process-unconventional-natural-gas-production> (2013).

2. Brittingham, M. C., Maloney, K. O., Farag, A. M., Harper, D. D., Bowen, Z. H. Ecological risks of shale oil and gas development to wildlife, aquatic resources, and their habitats. *Environmental*

614 *Science & Technology*. **48** (19), 11034–11047 (2014).
615 3. McBroom, M., Thomas, T., Zhang, Y. Soil erosion and surface water quality impacts of natural
616 gas development in East Texas, USA. *Water*. **4** (4), 944–958 (2012).
617 4. Maloney, K. O., Weller, D. E. Anthropogenic disturbance, and streams: land use and land-use
618 change affect stream ecosystems via multiple pathways. *Freshwater Biology*. **56** (3), 611–626
619 (2011).
620 5. Meyer, J. L. et al. The contribution of headwater streams to biodiversity in river networks1.
621 *JAWRA Journal of the American Water Resources Association*. **43** (1), 86–103 (2007).
622 6. Alexander, R. B., Boyer, E. W., Smith, R. A., Schwarz, G. E., Moore, R. B. The role of headwater
623 streams in downstream water quality. *Journal of the American Water Resources Association*. **43**
624 (1), 41–59 (2007).
625 7. Ulrich, N. et al. Response of aquatic bacterial communities to hydraulic fracturing in
626 Northwestern Pennsylvania: A five-year study. *Scientific Reports*. **8** (1), 5683 (2018).
627 8. Chen See, J. R. et al. Bacterial biomarkers of Marcellus shale activity in Pennsylvania. *Frontiers*
628 *in Microbiology*. **9**, 1697 (2018).
629 9. Rausch, P. et al. Comparative analysis of amplicon and metagenomic sequencing methods
630 reveals key features in the evolution of animal metaorganisms. *Microbiome*. **7** (1), 133 (2019).
631 10. Louca, S., Doebeli, M., Parfrey, L. W. Correcting for 16S rRNA gene copy numbers in
632 microbiome surveys remains an unsolved problem. *Microbiome*. **6** (1), 41 (2018).
633 11. Trexler, R. et al. Assessing impacts of unconventional natural gas extraction on microbial
634 communities in headwater stream ecosystems in Northwestern Pennsylvania. *Frontiers in*
635 *Microbiology*. **5**, 522 (2014).
636 12. Mumford, A. C. et al. Shale gas development has limited effects on stream biology and
637 geochemistry in a gradient-based, multiparameter study in Pennsylvania. *Proceedings of the*
638 *National Academy of Sciences*. **117** (7), 3670–3677 (2020).
639 13. JoVE Core Biology DNA Isolation. *Journal of Visualized Experiments*.
640 <https://www.jove.com/cn/science-education/10814/dna-isolation>.
641 14. Oxford Gene Technology DNA Storage and Quality. *OGT*.
642 <https://www.ogt.com/resources/literature/403_dna_storage_and_quality> (2011).
643 15. ThermoFisher SCIENTIFIC Technical Bulletin #159: Working with RNA. *Thermoscientific*.
644 <//www.thermofisher.com/us/en/home/references/ambion-tech-support/nuclease-
645 enzymes/general-articles/working-with-rna.html>.
646 16. QIAGEN AllPrep DNA/RNA Mini Kit. *Qiagen*.
647 <https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-
648 purification/multianalyte-and-virus/allprep-dnarna-mini-kit/#orderinginformation> (2020).
649 17. ZymoBIOMICS DNA/RNA Miniprep Kit. *Zymo Research*.
650 <https://www.zymoresearch.com/products/zymobiomics-dna-rna-miniprep-kit> (2020).
651 18. Desjardins, P., Conklin, D. NanoDrop microvolume quantitation of nucleic acids. *Journal of*
652 *Visualized Experiments*. (45), e2565 (2010).
653 19. Earth microbiome project 16S Illumina amplicon protocol: Earth microbiome project.
654 <https://earthmicrobiome.org/protocols-and-standards/16s/> (2018).
655 20. Gel Purification: Binding, washing and eluting a sample | Protocol. *Journal of Visualized*
656 *Experiments.* <https://www.jove.com/v/5063/gel-purification>.
657 21. New England Biolabs protocol for the use with NEBNext Poly(A) mRNA magnetic isolation

module (E7490) and NEBNext Ultra II RNA library prep kit for Illumina (E7770, E7775). <https://www.neb.com/protocols/2017/03/04/protocol-for-use-with-purified-mrna-or-rrna-depleted-rna-and-nebnext-ultra-ii-rna-library-prep-ki> (2020).
22. Anderson, M. J. Permutational multivariate analysis of variance (PERMANOVA). *Wiley StatsRef: Statistics Reference Online*. 1–15 (2017).
23. Schrader, C., Schielke, A., Ellerbroek, L., Johne, R. PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology*. **113** (5), 1014–1026 (2012).

Figure 1

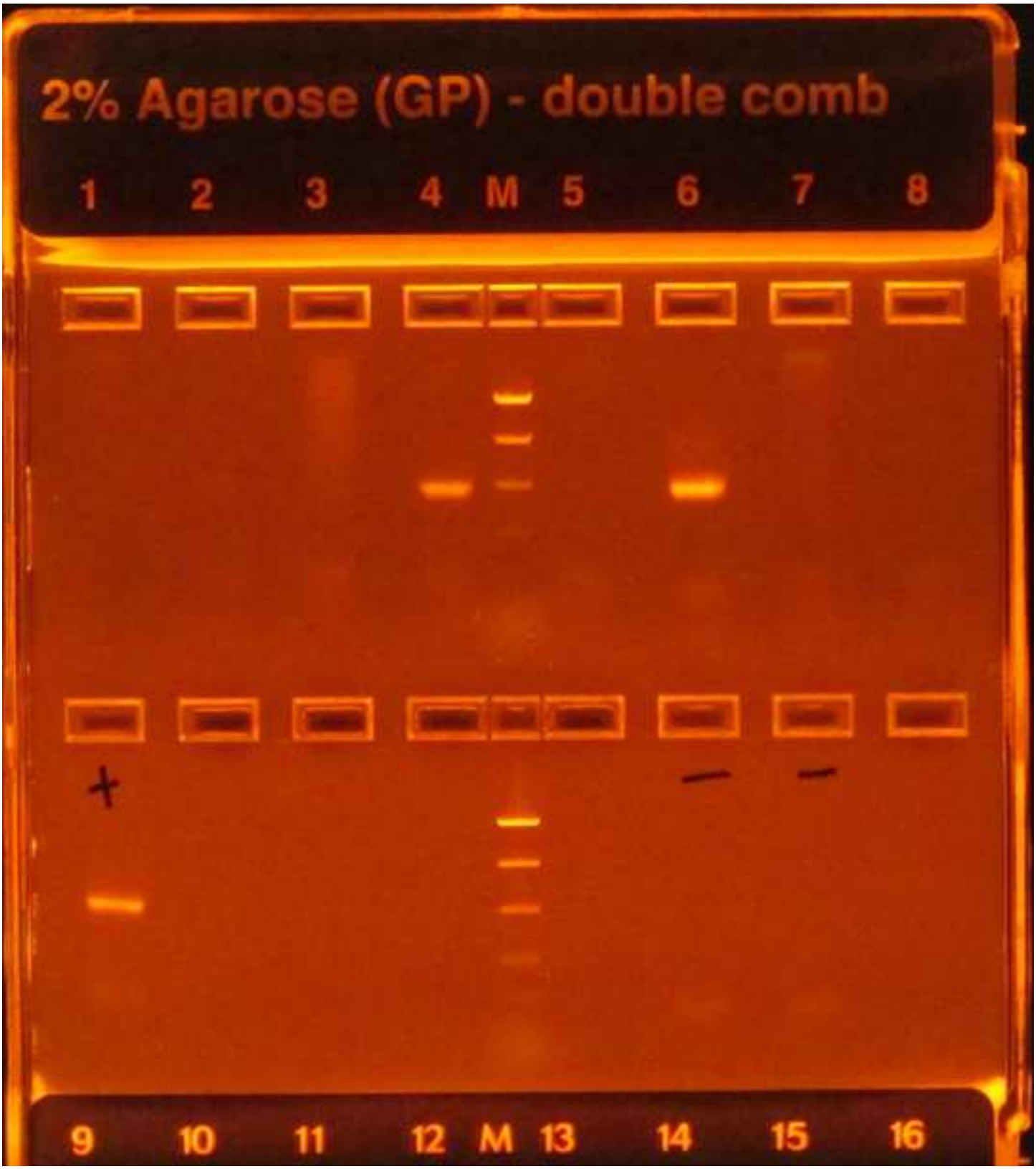

Figure 2

Click here to access/download;Figure;Figure-2_seq_finished.png

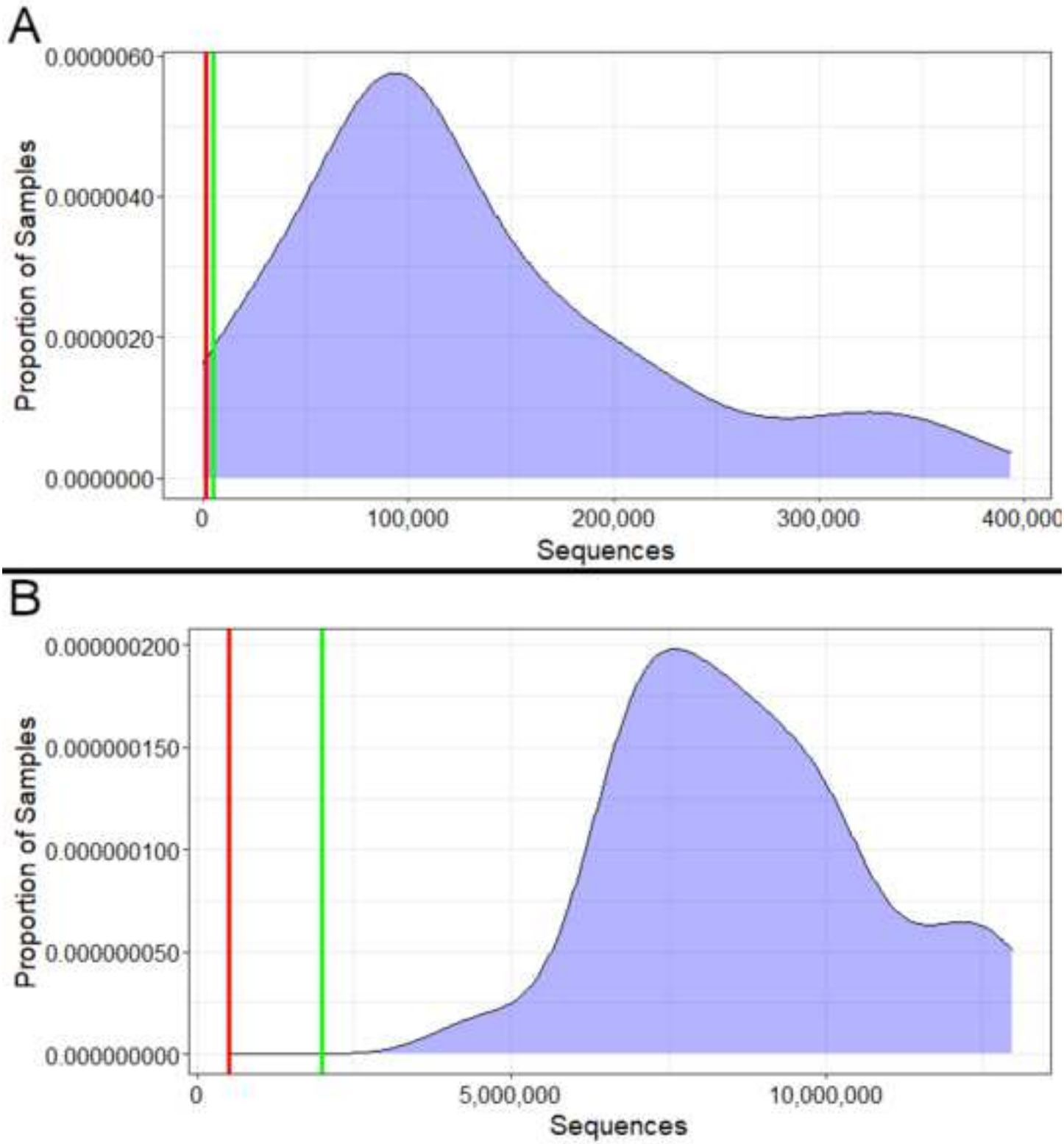

Figure 3

Click here to access/download;Figure;figure-3_FINAL.png

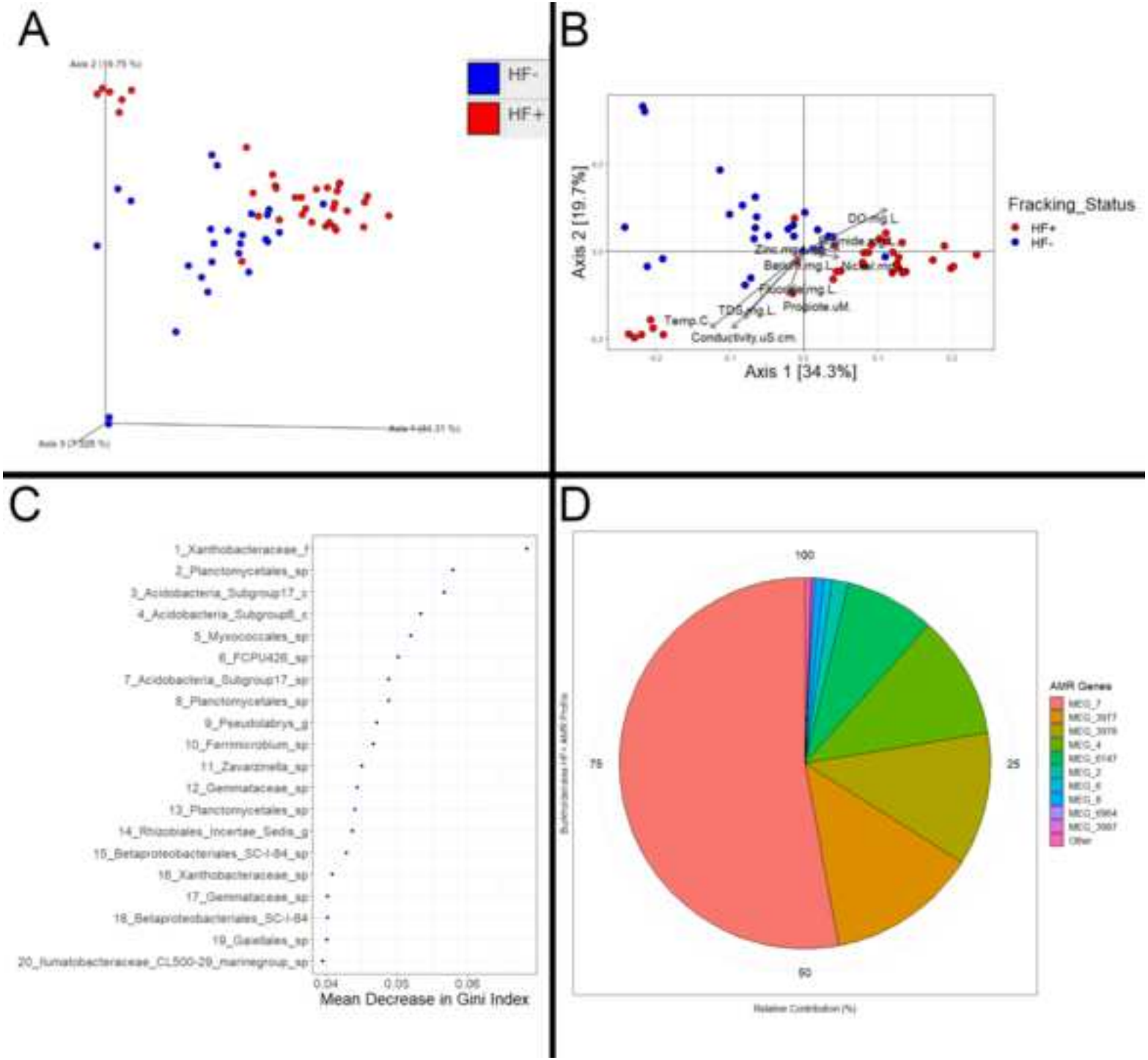

Table 1. Example DNA concentration values. For both DNA and RNA extractions, any detectable concentration g

SampleID	Concentration (ng/µL)
1	1.5
2	1.55
3	0.745
4	0.805
5	7.82
6	0.053
7	0.248
8	0.945
9	1.82
10	0.804
11	0.551
12	1.69
13	4.08
14	Below_Detection
15	7.87
16	0.346
17	2.64
18	1.15
19	0.951

;enerally indicates success, while a concentration below detection suggests extraction may have failed tł

nough for very low biomass samples, like filters, that might not necessarily be the case.

Name of Chemical/Solution
200 Proof Ethanol
Agarose
Disinfecting Bleach
DNA gel loading dye
DNA ladder
DNA/RNA Shield (2x)
Ethidium bromide
Forward Primer
Isopropanol
PCR-grade water
Platinum Hot Start PCR Master Mix (2x)
Reverse Primer
TBE Buffer (Tris-borate-EDTA)

Name of Kits/Equipment
1 L bottle
1.5 mL Microcentrifuge tubes
2% Agarose e-gel
50 mL Conicals
500 mL Beaker
Aluminum foil
Autoclave
Centrifuge
Cooler
Disruptor Genie
Electrophoresis chamber
Electrophoresis power supply
Freezer (-20 C)
Freezer (-80 C)
Gloves
Heat block
Lab burner
Laminar Flow Hood
Library purification kit
Magnet Plate
Microcentrifuge
Micropipette (1000 µL volume)
Micropipette (2 µL volume)
Micropipette (20 µL volume)
Micropipette (200 µL volume)
NEBNext Ultra II RNA Library Prep with Sample Purification Beads
Parafilm
PCR Tubes
Pipette tips (for 1000 µL volume)

Pipette tips (for 20 µL volume)
Pipette tips (for 200 µL volume)
PowerWulf ZXR1+ computer cluster
Qubit fluorometer starter kit
Scoopula
Sterile blades
Sterivex-GP Pressure Filter Unit
Thermocycler
Vise-grip
Vortex-Genie 2
WHIRL-PAK bags
ZymoBIOMICS DNA/RNA Miniprep kit

Company	Catalog Number
Thermo Fisher Scientific	A4094
Thermo Fisher Scientific	BP1356-100
Walmart (Clorox)	No catalog number
Thermo Fisher Scientific	R0611
MilliporeSigma	D3937-1VL
Zymo Research	R1200-125
Thermo Fisher Scientific	BP1302-10
Integrated DNA Technologies (IDT)	51-01-19-06
MilliporeSigma	563935-1L
MilliporeSigma	3315932001
Thermo Fisher Scientific	13000012
Integrated DNA Technologies (IDT)	51-01-19-07
Thermo Fisher Scientific	B52

Company	Catalog Number
Thermo Fisher Scientific	02-893-4E
MilliporeSigma	BR780400-450EA
Thermo Fisher Scientific	G401002
CellTreat	229421
MilliporeSigma	Z740580
Walmart (Reynolds KITCHEN)	No number
Gettinge	LSS 130
MilliporeSigma	EP5404000138-1EA
ULINE	S-22567
Bio-Rad	3591456
Bio-Rad	1664000EDU
Bio-Rad	1645050
K2 SCIENTIFIC	K204SDF
K2 SCIENTIFIC	K205ULT
Thermo Fisher Scientific	19-020-352
MilliporeSigma	Z741333-1EA
Sterlitech	177200-00
AirClean Systems	AC624LFUV
Qiagen	28704
Alpaqua	A001219
Thermo Fisher Scientific	75004061
Pipette.com	L-1000
Pipette.com	L-2
Pipette.com	L-20
Pipette.com	L-200R
New England BioLabs Inc.	E7775S
MilliporeSigma	P7793-1EA
Thermo Fisher Scientific	AM12230
Pipette.com	LF-1000

Pipette.com	LF-20
Pipette.com	LF-250
PSSC Labs	No number
Thermo Fisher Scientific	Q33239
Thermo Fisher Scientific	14-357Q
AD Surgical	A600-P10-0
MilliporeSigma	SVGP01050
Bio-Rad	1861096
Irwin	2078500
MilliporeSigma	Z258415-1EA
ULINE	S-22729
Zymo Research	R2002

Comments/Description
400 mL need to be added to Buffer PE (see Qiagen QIAQuck Gel Extraction kit protocol) and 96 mL needs to be added to the DNA/RNA Wash Buffer (see ZymoBIOMICS DNA/RNA Miniprep kit protocol).
100 g per bottle. 0.6 g of agarose would be needed to make one 2% 30 mL gel.
Use a 10% bleach solution for cleaning the work area before and after lab procedures
Each user-made (i.e. non-e-gel) should include loading dye with all of the samples in the ratio of 1 µL dye to 5 µL s
A ladder should be run on every gel/e-gel
3 mL per sediment sample (50 mL conical) and 2 mL per water sample (filter)
Used for staining user-made e-gels
0.5 µL per PCR reaction
Generally less than 2 mL per library. Volume needed varies by mass of excised gel fragment (see Qiagen QIAQuick
13 µL per PCR reaction (assuming 1 µL of sample DNA template is used)
10 µL per PCR reaction
0.5 µL per PCR reaction
1 L of 10x TBE buffer (30 mL of 1x TBE buffer would be needed to make one 30 mL gel)

Comments/Description
One needed per stream (the same bottle can be used for multiple streams if it is sterilized between uses)
5 microcentrifuge tubes are needed per DNA extraction and an additional 3 are needed to purify RNA (see ZymoB
Each gel can run 10 samples (so 9 with a PCR negative and 8 if the extraction negative is run on the same gel)
1 50 mL conical needed per sediment samples
Only 1 needed (for flame sterilization)
Aluminum foil can be folded and autoclaved. The part not exposed to the environment can then be used as a sterile, DNA and RNA free surface for processing filters
Only one needed
Only 1 needed
Just about any cooler can be used. This one is listed due to being made of foam, making it lighter and thus easier t
Only one needed
Only 1 needed
Only 1 needed
One needed to store DNA extracts
One needed to store RNA extracts
The catalog number is for Medium gloves.
Only one needed
Only one needed
Only 1 needed
One kit has enough for 50 reactions
Only one needed
Only one needed
Only 1 needed
Only 1 needed
Only 1 needed
Only 1 needed
One kit has enough reagents for 24 samples.
2 1" x 1" squares are needed per filter
One tube needed per reaction
Pack of 576 tips

Pack of 960 tips
Pack of 960 tips
This is just an example of a supercomputer powerful enough to perform metatranscriptomics analysis in a timely r
Comes with a Qubit 4 fluorometer, enough reagent for 100 DNA assays, and 500 Qubit tubes
Only one needed
One needed per filter
1 filter needed per water sample
Only one needed
Only one needed (for cracking open the filters)
Only 1 needed
1 needed per filter
One kit has enough reagents for 50 samples.

ample

Gel Extraction kit protocol).

IOMICS DNA/RNA Miniprep kit protocol)

:o take along for field sampling.

manner. Only one needed.

Responses to Comments

Editorial comments:

Changes to be made by the Author(s):

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. Please check spaces and headings.

2. Please provide an email address for each author.

Jeremy Chen See - chensej@juniata.edu
Olivia Wright - wrighog18@juniata.edu
Lavinia Unverdorben - lvunverdorben@gmail.com
Nathan Heibeck - heibens18@juniata.edu
Regina Lamendella - lamendella@juniata.edu

3. Please rephrase the Summary to clearly describe the protocol and its applications in complete sentences between 10-50 words: “Here, we present a protocol to …”

Thank you for the comment. See below for our revised Summary.

> Here, we present a protocol to investigate the impacts of hydraulic fracturing on nearby streams through analyzing their water and sediment microbial communities.

4. Please rephrase the Abstract to more clearly state the goal of the protocol.

Thank you for this comment. We added a sentence to the Abstract to state the protocol’s goal. See below

> ...Therefore, this protocol aims to use the bacterial community to determine if streams have been impacted by fracking...

5. Please revise the Introduction to include all of the following:

a) The advantages over alternative techniques with applicable references to previous studies

Thank you for the comment. We added the following paragraph to the introduction to highlight the advantages of the methods presented here compared to previous studies.

> Therefore, in contrast to protocols in previously published studies[7,8,10,11], this one also covers sample collection, preservation, processing, and analysis for investigating microbial community function (metatranscriptomics). The steps detailed herein allow researchers to see what impact, if any, fracking has had on the genes and pathways expressed by microbes in their streams, including antimicrobial resistance genes. Moreover, the level of detail presented for sample collection is greater. Although several of the steps and notes may seem obvious to experienced researchers, they could be invaluable to those just starting research.

b) Please adjust the numbering of the Protocol to follow the JoVE Instructions for Authors. For example, 1 should be followed by 1.1 and then 1.1.1 and 1.1.2 if necessary. Please refrain from using bullets or dashes.

Thank you for this comment.

The numbering has been adjusted.

6. Under 2, step 3: what is the volume of the sterile Luer lock syringe?

Thank you for this comment. The volume is 60 mL. We added a note below 2.3 with this information. See below.

> Note: The volume of the syringe can be variable, as long as the total amount of water pushed through the filter is tracked. However, generally, 60 mL is preferred.

7. Please use µ symbol instead of uL throughout the manuscript.

Thank you for this comment.

We made the change as requested.

8. Please note that your protocol will be used to generate the script for the video and must contain everything that you would like shown in the video. Please add more details to your protocol steps. Please ensure you answer the "how" question, i.e., how is the step performed? For instance, more details would be helpful for the (part 6) RNA library creation and purification and (part 7) microbial community analysis. Alternatively, add references to published material specifying how to perform the protocol action. Please add more specific details (e.g. button clicks for software actions, numerical values for settings, etc) to your protocol steps. There should be enough detail in each step to supplement the actions seen in the video so that viewers can easily replicate the protocol.

Thank you for this comment.

The following steps and notes have been added throughout the paper.

1.4.1. To draw up the preservative, first set the volume to 1000 µL and then attach the tip to the micropipette by firmly pressing the micropipette onto the tip while it is in the tip rack.

1.4.2. Once the tip is attached, press down on the micropipette to the first stop and put the tip slightly below the preservative's meniscus. Release the first stop.

1.4.3. Open the conical containing the sediment sample and put the pipette tip inside of it and depress to the second stop

1.4.4. Repeat steps 1.4.1 thru 1.4.3 two additional times for a total of 3000 µL being added to the sample

Note: Change tips between samples and after any time the tip touches a non-sterile surface.

1.4.5. Cap the sediment conical tube and then swirl to mix it.

3.1.1. Clean the work area with 10% Bleach and 70% Ethanol before beginning sample transfer.

3.1.2. For an example DNA extraction protocol, see[12].

3.4.1. For an example Nanodrop quantification protocol, see[16].

4.1.1. Perform this lab procedure under laminar flow conditions to reduce the chance of contamination.

4.1.2. As with DNA extractions, clean the work area with 10% Bleach and 70% Ethanol before beginning sample transfer.

5.5. A commercial kit can then be used for excising the DNA from the gel or for an example gel purification protocol, see[20].

Note: This step should be performed in a different area than DNA or RNA extractions to prevent future contamination, as cutting the gel will spread PCR amplicons onto both the experimenter and the surrounding area.

6.1.1. Regardless of the kit chosen, work should be performed in a sterile environment in a hood under laminar flow.

6.2. Check the libraries for detectable concentrations of DNA.

6.2.1. Once more, a bioanalyzer, fluorometer, or spectrophotometer could be used.

6.3. Pool the metatranscriptomic libraries in a roughly equimolar ratio.

6.4. Purify the library following the same protocol for the 16S library purification, except fragments between 250 and 450 bp should be excised instead.

Note: Whereas the 16S library had a distinct band representing the amplified region, the result here will be more of a smear.

6.5. Ship the purified library with dry ice to a sequencing facility.

Part 8 (formerly Part 7) has been greatly expanded to include an example metatranscriptomics analysis pipeline per the editor and Reviewer 1's comments.

Unfortunately, the expanded Microbial Community Analysis section puts the Protocol over the 10 page maximum. Would it be acceptable to include the example metatranscriptomics analysis pipeline as a Supplemental File if our paper is accepted for publication, instead of in the main text?

9. Please include a one line spacer between each protocol step/substep and then highlight up to 3 pages of protocol text for inclusion in the video. This is a hard production limit to ensure that videography can occur in a single day.

Thank you for this comment. Lines have been added between each step and substep and three pages have been highlighted.

10. As we are a methods journal, please revise the Discussion to explicitly cover the following in detail in 3-6 paragraphs with citations:
a) The significance with respect to existing methods
b) Any future applications of the technique

Thank you for these comments. We modified the first paragraph in the Discussion to address them. See below.

> The methods described in this paper have been developed and refined over the course of several studies published by our group between 2014 and 2018[7,8,10], and have been employed successfully in a collaborative project to investigate the impacts of hydraulic fracturing on aquatic communities in a three year project that will soon submit a paper for publication. These methods will continue to be utilized over the course of the remainder of the project. Additionally, other current literature investigating the impact of hydraulic fracturing on streams and ecosystems describe similar methods for sample collection, processing, and analysis[7,8,10,11]. However, the methods presented here for sample collection are more detailed, and specific notes are included throughout this paper to avoid contamination. Furthermore, none of those papers utilized metatranscriptomic analysis, making this paper the first to describe how that analysis can be used to elucidate hydraulic fracturing's impact on nearby streams.

11. JoVE cannot publish manuscripts containing commercial language. This includes trademark symbols (™), registered symbols (®), and company names before an instrument or reagent. Please remove all commercial language from your manuscript and use generic terms instead. All commercial products should be sufficiently referenced in the Table of Materials and Reagents.
For example: The company name “Invitrogen” in figure 1.

> Thank you for this comment.
>
> “Invitrogen” has been removed from Figure 1, and all uses of “Qubit” have been changed to just “fluorometer”. All uses of NanoDrop have likewise been changed to “spectrophotometer”.

12. Please include a table of the essential supplies, reagents, and equipment. The table should include the name, company, and catalog number of all relevant materials in separate columns in an xls/xlsx file. Please sort the Materials Table alphabetically by the name of the material. -

Thank you for this comment.

We created a table to include all required information. Please see Chen_See_et_al_jove-materials-list.xlsx

---

Reviewers' comments:

Reviewer #1:

Manuscript Summary:
This manuscript describes sampling techniques and laboratory protocols to collect sediment and water samples in streams and prepare them for next generation 16S rRNA gene sequencing and metatranscriptomics analyses. The goal of these analyses is to assess the impact of hydraulic fracturing on water quality by examining changes in microbial community composition.

Major Concerns:

While this is a very well written and easy to follow methods paper, I wonder how the described methods are specific for fracking studies. The described protocol summarized common sampling techniques in the field as well as routine sample extraction and processing for 16S rRNA gene sequencing and metatranscriptomics in the lab. How any of this is specific to fracking eludes me.

Thank you for this comment.

Many of these methods are applicable to a variety of different studies. However, when used in combination with each other, as described in this paper, they allow researchers to see whether the total bacterial community, active microbial community, and functional profiles of potentially impacted streams have been altered in association with hydraulic fracturing. In other words, while these methods may be applicable to other projects, they can be applied to this field to see if hydraulic fracturing has had an impact on nearby streams. Moreover, by detailing these methods to this extent, it improves comparability and reproducibility of fracking microbial studies, which is vital for comparing results across studies.

The authors mention that contamination of low biomass samples can be a specific concern for these samples, however, they never define what low biomass means. What are the average cell counts in sediments and water samples of these systems? While I expect cell numbers to be lower in the water, I doubt that the cell counts in the sediments are particularly low.

Thank you for this comment.

To clarify, fewer than 10^4 cells per mL is what we had in mind when we said “low biomass”. However, upon consultation with one of our collaborators, we think it would be better to say “low extraction yield” since they felt our cutoff was arbitrary. Moreover, the average cell counts for our current project for the water samples (115,000 cells/mL based on AODC) exceed what was our cutoff, but are still less than the average for the sediment samples (408,000 cells/mL). The relatively low biomass is often the culprit for low nucleic acid yields when extracting from filters (water samples). However, sediment samples can also have low yields if they are mainly rocks and gravel, which are hard to avoid in some streams, instead of true sediment. If we would keep the original wording, “sediment” should be omitted, but with the change, its inclusion is appropriate. See below for the altered text in the first paragraph of page 22.

> Similar to field collection, avoiding or minimizing contamination is also important during nucleic acid extraction and sample preparation, especially when working with low nucleic acid yield samples, such as suboptimal sediment samples (samples containing a large amount of gravel or rocks) or water samples

As the authors state in the discussion it is important to avoid contamination both during sampling and later on during sample work-up in the lab. Therefore, I suggest to include an extra section in the protocol on the special care that needs to be taken to avoid contamination. For instance, that gloves should be worn at all times, how tools are best sterilized and during which steps contamination can arise and how to best avoid it.

> Thank you for this comment.
>
> We added an additional section (7) with additional recommendations for avoiding contamination. See below.

7. Contamination Minimization Recommendations

7.1. Gloves should be worn at all steps (except for 8.) to minimize the chance of introducing microbes or human cells into the sample.

7.2. For step 2.1, the 1 L bottle can be sterilized by rinsing with 10% bleach for 2 minutes, followed by rinsing three times with deionized water and then once with 70% ethanol, and finally autoclaving with settings: 30 minutes exposure time at 121.1oC and 15 minute drying time.

Note: During autoclaving, the cap on the bottle should be very loose to avoid the bottle being compressed in the process.

7.3. All work surfaces used during lab procedures should be sterilized beforehand by wiping with a 10% bleach solution, followed by a 70% ethanol solution.

7.4. For pipetting steps (3-6), filter tips should be used to avoid contamination due to the pipette itself. All tools used for lab work, including pipettes, should be wiped down before and after with the bleach and ethanol solutions. The filter tips are an extra precaution. Be sure to change tips every time they touch a non-sterile surface.

7.5. For step 3.3, a sterile, DNA and RNA free-surface can be created by folding aluminum foil so that the inner part of the fold is not exposed to the outside environment and autoclaving the folded piece with the settings: 121.1oC and 5 minute drying time.

Figure 3 and the steps included to make these plots are hardly touched upon in the text (they are only mentioned in the figure caption). I suggest to expand on this in the main text as an example how to process metatranscriptomics data sets.

Thank you for this comment.

Section 8 (Microbial Community Analysis) has been expanded to include a link to an official Qiime2 16S tutorial, as well as an example metatranscriptomics pipeline, which can be used to generate the analyses shown in Figure 3.

However, the expansion to Section 8 makes the protocol exceed the maximum page limit. Therefore, as an alternative to including it in the main text, we may have to include it as a Supplemental File.

Minor Concerns:
For step 1.1 I suggest to add a note that gloves should be worn during sampling.

Thank you for this comment.

We added the note below immediately after step 1.1. See below.

Note: Wear gloves during sample collection to avoid introducing unwanted human contamination.

For step. 2.2 define the filter type/material.

Thank you for the comment.

We changed the filter step (now 2.3) to read:

> Once on a stable surface, use a sterile Luer lock syringe and draw up a full volume. Then connect the syringe to a sterile and DNA/RNA-free 1.7 cm diameter polyethersulfone filter with a pore size of 0.22 µm and push the entire volume through the filter by pressing the plunger all the way down. Repeat this process until the total volume collected in the bottle (1 L) is pushed through the filter.

I wonder how do the authors define low biomass and "enough biomass" (can they provide cell numbers that define this?) as they refer to sediment and water samples as low biomass.

> Thank you for this comment.
>
> By “low biomass”, we initially meant less than 10^4 cells per mL. Referring to sediment samples as “low biomass” was a mistake.
>
> Considering the results of our current project, 61 out of 63 sediment samples and 19 out of 21 water samples yielded enough sequences to be analyzed. The lower quartile for sediment was 217,000 cell per mL and for water 20,000 cells per mL. Therefore, based on the results of this project, at least 217,000 cells per mL for sediment and 20,000 cells per mL for water (assuming at least 200 mL is pushed through the filter) should be “enough biomass” to reliably extract enough nucleic acids for genetic analysis.
>
> The second note under step 2.3 has been revised to:
>
> > Note: Streams with a lot of sediment after rainfall or a disturbance can make it difficult to push the entire volume through the filter. While 1 L is ideal, anecdotally, a volume of at least 200 mL would likely still collect enough biomass (assuming ~20,000 cells per mL) for the extraction of DNA and RNA. Some sediment may stick to the filter paper, but there should be no sediment inside of the filter otherwise.

How do the authors suggest to deal with (lab) contaminants in the metatranscriptome data sets? Should they all be kicked out or should additional processing pipelines be included to make sure they are indeed lab contaminants and are not indigenous to the sample.

> Thank you for this comment.

Some contaminants are obviously contaminants, such as sequences belonging to humans. Those can be identified and removed using several programs, including Kneaddata as detailed in steps 8.4.3 and 8.4.4 (see below).

> 8.4.3. Create a database of human sequences. Download the human genome with "wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.28_GRCh38.p13/GCA_000001405.28_GRCh38.p13_genomic.fna.gz" and decompress using gunzip. Then run "bowtie2-build GCA_000001405.28_GRCh38.p13_genomic.fna.gz human"
>
> Note: If the gunzip command is not found, run "conda install -c ostrokach gzip" to install it.
>
> Note: Bowtie2 and Kneaddata2 can be installed using: "conda install -c bioconda kneaddata"
>
> 8.4.4. Remove human contamination. Run Kneaddata2 using "kneaddata --bypass-trim --input $FILTERED_R1 --input $FILTERED_R2 -o $OUTPUT -db $PATH/human" where $FILTERED_R1 is the filtered R1 file from fastp, $FILTERED_R2 is the filtered R2 file from fastp, $OUTPUT is the name of the output folder, and $PATH is the absolute path to the database created in step 8.4.3.

Microbial contaminants are more challenging to deal with. The only way to know for sure that a microbe was introduced as a contaminant, instead of being indigenous to the sample, is to run and sequence at least one negative control. However, metatranscriptomics projects often do not include negative controls to save on cost. If no negative control is sequenced, it is hard to defend throwing out sequences just because they might be contaminants.

Therefore, to properly account for microbial contamination, at least one negative control must be processed in parallel with the samples and sequenced.

Our recommendation is to process the negative(s) with the samples and to use decontam afterwards only if at least one of the samples seems to have been definitively impacted by contamination based on it having a similar composition to the negative.

This recommendation is included in the paper as a note under step 8.4.4, reading:

> Note: Microbial contamination is more challenging to deal with due to the difficulty in determining which sequences are from the samples as opposed to contamination. Therefore, at least one negative control must be sequenced to account for this type of contamination. The control should be processed the same as the actual samples. If any of the samples are nearly identical in composition to the negative(s), then the R package decontam should be used to remove the contaminants from the samples. However, this should only be done if there is clear evidence of contamination, as decontam can potentially remove features that are not actually the result of contamination.

Reviewer #2:

Manuscript Summary:
The authors of the manuscript entitled "Molecular Microbial Methods for Evaluating the Impact of Hydraulic Fracturing on Streams" present a DNA/RNA extraction method for the analysis of the bacterial communities in streams near fracturing sites.

Major Concerns:
The method used is general and can be applied to the analysis of water and sediment samples in several environments. In fact, is not different from several others that have been published. The authors state at the end of the manuscript that "The methods in this paper describe field collection and sample processing methods for both 16S rRNA and metatranscriptomic analyses specific for fracking studies." However, it is not clear why is specific for fracking studies.

Thank you for this comment.

The Reviewer #1 raised this concern as well.

Many of these methods are applicable to a variety of different studies. However, when used in combination with each other, as described in this paper, they allow researchers to see whether the total bacterial community, active microbial community, and functional profiles of potentially impacted streams have been altered in association with hydraulic fracturing. In other words, while these methods may be applicable to other projects, they can be applied to this field to see if hydraulic fracturing has had an impact on nearby streams. Moreover, by detailing these methods to this extent, it improves

comparability and reproducibility of fracking microbial studies, which is vital for comparing results across studies.

Since no data is provided regarding Figure 3, how can someone who does an analysis of a bacterial community in a given site compare the community to those presented by the authors to assess if the site is contaminated or not?

Thank you for the comment. The representative results section has been expanded to include an interpretation of the analyses shown in Figure 3. See below.

> Those differences could take the form of consistent compositional shifts based on fracking status. If that were the case, HF+ and HF- samples would be expected to cluster apart from each other in a PCoA plot, as is the case in Figure 3A and 3B. To confirm that those apparent shifts are not just an artifact of the ordination method, further statistical analysis is needed. For example, a PERMANOVA test on the distance matrix that Figure 3A and 3B are based on revealed significant clustering based on fracking status, meaning that the separation observed in the plot is consistent with the underlying data. A significant PERMANOVA or ANOSIM result is a strong indication of consistent differences between HF+ and HF- samples, which would indicate that the HF+ samples were impacted by fracking, while a high p-value would indicate that the samples were not impacted. Metatranscriptomic data can likewise be visualized and evaluated using the same methods.
>
> Examining differential features (microbes or functions) can reveal evidence that samples have been impacted too. One method of determining differential features is to create a random forest model. The random forest model can be used to see how well the samples' fracking status can be correctly classified. If the model performs better than expected by chance, that would be additional evidence of differences dependent on fracking status. Moreover, the most important predictors would reveal which features were most important for correctly differentiating samples (Figure 3C). Those features also then would have had consistently different values based on fracking status. Once those differential features are determined, the literature can be reviewed to see if they have been previously associated with fracking. However, it may be challenging to find studies that determined differential functions, as most have only used 16S rRNA compositional data. Therefore, for evaluating the implications of

> differential functions, one possible method would be to see if they have been previously associated with potential resistance to biocides commonly used in fracking fluid or if they could aid in tolerating highly saline conditions. Furthermore, examining the functional profile of a taxon of interest could reveal evidence of fracking's impact (Figure 3D). For example, if a taxon is identified as differential by the random forest model, its antimicrobial resistance profile in HF+ samples could be compared to its profile in HF- samples and if they differ greatly, that could suggest that fracking fluid containing biocides entered the stream.

The outcome of this paper, according to title should allow an evaluation of the impact of fracturing and not a protocol to extract DNA/RNA and run 16S RNA or metatranscriptomics analyses.

> Thank you for this comment. The additions to the Representative Results section (see above) should help the reader interpret the analyses presented in Figure 3 and hopefully their own analyses to understand if hydraulic fracturing did impact their samples or not.

English grammar needs to be revised and the text to be carefully read as some sentences do not make sense or are incomplete. Specific comments are listed below.

Minor Concerns:
Abstract
"The bacteria within those streams can be used as indicators of stream health, as the types of bacteria present and their abundance in a disturbed stream would be expected to differ from those in an otherwise comparable but undisturbed stream."
The authors should give information about which bacteria are indicators of undisturbed and of disturbed streams.

> Thank you for this comment. We added specific examples of previously identified differential bacteria to the end of the third paragraph in the Introduction. See below.
>
> > For example, Beijerinckia, Burkholderia, and Methanobacterium were identified as enriched in streams near fracking while Pseudonocardia, Nitrospria, and Rhodobacter were enriched in the streams not near fracking[8].

"A common practice within the field of molecular ecology is to use the highly variable V3-V4 region of the 16S rRNA gene for sequencing resolution, often down to the genus level with a wide scope of identification, as is ideal for unpredictable environmental samples."
How does the technique depend on the primers used? The paper cited by the authors (ref 9) shows for example that is very difficult to classify Escherichia/Shigella due to close sequences. Shotgun metagenomics allows the determination of functional content of samples directly, however 16S RNA sequencing does not. How can he authors infer about the quantity of cells from each genus and about the viable cells of the community?

Thank you for this comment.

With respect to lab work, the target size of the sequences selected for sequencing differs based on the amplicon size, which depends on the primers. Considering computational analysis, the biggest consideration would be the database used to classify sequences. For example, on Qiime2's website, there is a link to download a pre-trained Silva database for sequences amplified with the 515F/806R primers (the V4 region of 16S rRNA). However, that database should not be used if the sequences were amplified using different primers, such as the ones used in the paper (ref 9) to amplify the V1-V2 region. Essentially, it is vital that the database used actually contains sequences that are comparable to the data generated.

We agree that some genera are challenging to differentiate, especially when using amplicon based approaches. However, as shown in Figure 1 in ref 9, analysis based on 16S rRNA sequences can yield genus-level identifications and abundances that are close to reality.

We also agree that functions cannot be conclusively determined through 16S analysis alone, but predictive tools, such as PICRUSt2 do exist; however, that tool is not covered in this paper though PICRUSt is used in the paper that we cited (ref 9).

Quantity is difficult to assess due to the variable copy number of the 16S rRNA across various clades.Though several tools exist for adjusting abundances based on copy number, the accuracy of their adjustments is questionable, with at least one review even recommending against their use (ref 10) to keep methods between studies comparable.

Viability cannot be inferred through 16S rRNA since it is based on all DNA present in the samples. Therefore, DNA from unviable cells may still be amplified and sequenced. However, metatranscriptomics can be used so that only data from viable cells are analyzed (as RNA degrades rapidly in the environment).

We added a sentence to the end of the fourth paragraph of the Introduction to note the variability in copy number. It reads:

> However, it is worth noting that bacteria have varying numbers of the 16S rRNA gene, which affects their detected abundances[10]. There are a few tools to account for this, but their efficacy is questionable[10].

Protocol:
The authors refer to the tubes used for sampling only as "conical". This is wrong and must be corrected to "conical tube" or "conical centrifuge tube" throughout the manuscript. Example:
- "Submerge a sterile 50 mL conical into the stream water." should be "Submerge a sterile 50 mL conical tube into the stream water." or "Submerge a sterile 50 mL conical centrifuge tube into the stream water."

Thank you for this comment.

We changed all instances of “conical” to “conical tube”.

- "3. Remove the conical from the water and dump out the water covering the collected sediment."
State if all water should be removed or if the sediment should be left with some water covering it.

Thank you for your comment.

A thin layer of water should be left over the sediment sample (approximately 1mL) should be left. We added this information to step 1.3. See below.

> 1.3. Remove the conical tube from the water and dump out all water, except for a thin layer covering the sediment sample (approximately 1 mL).

- "4. Using a 1000 microliter pipette and appropriate pipette tips, add 3 mL of DNA/RNA preservative to the collected sample."
The authors must state which DNA/RNA preservation agents should be added.

Thank you for your comment.

The specific DNA/RNA preservative is listed in the newly created Materials Table. It will not be included in the text though to avoid violating Jove's policy on trademarks.

- "5. If using a preservative, swirl to mix the conical (after capping it)."

English grammar has to be revised and corrected. This sentence is an example of sentences that must be corrected. In this case, the sentence may be corrected to "5. If using a preservation agent, swirl to mix it with the sample inside the conical tube (after capping it).

Thank you for your comment. We have revised the sentence to this:

1.4.5. Swirl the capped conical tube for 5 seconds to ensure the preservative and sample are thoroughly mixed.

- "6. (…) 16S analysis later (DNA)" - RNA?? Confirm the whole sentence.

Thank you for your comment. Step 1.5 has been revised to

Place the samples on ice for the rest of sample collection. Upon returning from the field, store in a freezer. The freezer should be kept at -20°Cif the samples are to be used for 16S analysis (DNA), or -70°Cif they are to be used for metatranscriptomics analysis (RNA).

- "Once on a stable surface, use a sterile Luer lock syringe and draw up a full volume." - State the preferable volume of the syringe to be used and the diameter of the filter. Should a filter be used in a filter holder or a syringe filter is better?

Thank you for the comment. We revised step 2.3 to include the diameter and added a note after it about the preferred volume. The filter detailed in the Table of Materials is a syringe filter that can be simply twisted onto the syringe.

2.3. Once on a stable surface, use a sterile Luer lock syringe and draw up a full volume. Then connect the syringe to a sterile and DNA/RNA-free 1.7 cm diameter polyethersulfone filter with a pore size of 0.22 µm and push the entire volume through the filter by pressing the plunger all the way down. Repeat this process until the total volume collected in the bottle (1 L) is pushed through the filter.

Note: The volume of the syringe can be variable, as long as the total amount of water pushed through the filter is tracked. However, generally, 60 mL is preferred.

- "4. Using a P1000 micropipette, add 2 mL of a DNA/RNA preservative by discharging it through the filter's larger opening while holding the filter horizontally."
This is not clear. Should the same side of the filter, where the syringe used to be, be used? If yes, how can 2 mL of a preservation agent be added with a P1000, as the filter should have a significant amount of sediment and P1000 does not make enough pressure for the 2 mL to pass?

Thank you for your comment. We revised 2.4 to include additional details for adding the preservative and added a sentence to the note prior about sediment, as well as another note after 2.4 about the total volume of fluid the filter can process.

Note: Streams with a lot of sediment after rainfall or a disturbance can make it difficult to push the entire volume through the filter. While 1 L is ideal, anecdotally, a volume of at least 200 mL would likely still collect enough biomass (assuming ~20,000 cells per mL) for the extraction of DNA and RNA. Some sediment may stick to the filter paper, but there should be no sediment inside of the filter otherwise.

2.4 Using a P1000 micropipette, add 2 mL of a DNA/RNA preservative by discharging it through the filter’s larger opening (where it was attached to the syringe) while holding the filter horizontally. The tip of the pipette tip should be within the barrel of the filter when it is depressed to ensure the preservative enters the filter.

> Note: As with sediment collection, this step is not absolutely necessary, but is strongly recommended for increased nucleic acid yield later, especially for RNA.
>
> Note: The maximum volume for the filter is 1 L, but the maximum process volume is 2 L. Therefore, some of the preservative may flow out of the filter. Pushing around 10 mL of air through the filter after the water is processed will force out excess water and help reduce the amount of preservative leakage.

- "Note: Ensure that the side used to wrap the filter is the one that was covered in plastic, as that is the side that is sterile."
Most brands wrap the filters and there is no specific side covered in plastic. The sentence should be generic.

Thank you for your comment. We've revised this in the note following step 2.5 to read

> Note: Ensure that the side used to wrap the filter is sterile, i.e. not previously exposed to the environment.

- "Make sure that the filter paper does not come into contact with any non-sterile surfaces during this process, as that would lead to unwanted contamination."
Not only sterilized should be used but especially DNA-free material. Something that has been sterilized may still contain RNA or DNA.

Thank you for your comment. We revised the last sentence in step 3.3 to read:

> Make sure that the filter paper does not come into contact with any surfaces which are not sterilized, or could have nucleic acid present, as this would lead to unwanted contamination of the sample.

- "For environmental samples, inhibition is often the culprit."
Inhibition of what? What is the critical step?

Thank you for your comment. We've revised that sentence in the note following step 4.3 to read

> ...For environmental samples, inhibition of the PCR reaction is often the culprit, which can be due to a variety of substances interfering with Taq polymerase[19]...

- "Diluting the DNA extracts for the failed samples before PCR"
State what should be used to dilute the DNA extracts (which buffer).

> Thank you for your comment. PCR grade water (see Table of Materials) can be used to dilute the DNA extracts. Accordingly, we added a sentence to the note after step 4.3 that reads
>
> ...If inhibition is suspected, PCR grade water (see Materials List) can be used to dilute the DNA extracts.

- "Moreover, a bright band in the negative would also indicate a failure (…)" - in the negative what? Please read the text carefully as there are other sentences similar to this where it is unclear to what the authors are referring to. In the same sentence, the authors have "that that".

> Thank you for your comment. We revised the sentence to read
>
> Moreover, a bright band in the gel lane that contains a negative PCR control would also indicate a failure since it would be risky to assume that the contamination impacting the negative control(s) did not affect the samples.

- "16S samples should have a minimum of 1,000 sequences" - To what are the authors referring to? Is it to sequences or reads? The same is valid for the caption of Fig. 2.

> Thank you for the comment. We are referring to sequences in both cases and have removed all instances of “reads” from the text to avoid confusion.

- The quality of Figures 2 and 3 provided is insufficient for their evaluation in this review.

> Thank you for this comment. Both figures have been remade so that they will be legible when embedded in the paper.

Supplemental Example MT pipeline

1. Microbial Community Analysis Pipeline

1.1. Install the necessary programs.

1.1.1. Install Conda by following the instructions here https://docs.conda.io/projects/continuumio-conda/en/latest/user-guide/install/index.html

1.1.2. Install FastQC with: conda install -c bioconda fastqc

1.1.3. Install fastp with: conda install -c bioconda fastp

1.1.4. Install Kneaddata with: conda install -c bioconda kneaddata

1.1.5. Install HUMAnN2 with: conda install -c bioconda humann2

1.1.6. Install PEAR with: conda install -c bioconda pear

1.1.7. Install BLAST with: conda install -c bioconda blast

1.1.8. Install QIIME2 with: conda install -c qiime2 qiime2

1.1.9. Install R with: conda install -c r r

1.2. Evaluate the data's raw quality by using FastQC's "fastqc" command in the form of "fastqc $data -o $output". Once FastQC is complete, the html in the output folder should be inspected to see sequence quality as defined by q scores.

Note: The variables in the command are as follows: $data is the name of an individual raw data file (with the extension of either ".fastq" or ".fq") and $output is the name of the output folder.

1.3. Run fastp's "fastp" command to discard poor quality data in the form of "fastp -r -i $FORWARD -I $REVERSE -t $FORWARD_#_BASES_DROPPED -T $REVERSE_#_BASES_DROPPED --out1 $FILTERED_FORWARD --out2 $FILTERED_REVERSE".

Note: The variables in the command are as follows: $FORWARD is the file containing the forward sequences, $REVERSE is the file containing the reverse sequences, $FORWARD_#_BASES_DROPPED is the number of bases to cut from the end of the forward sequences, $REVERSE_#_BASES_DROPPED is the number of bases to cut from the end of the reverse sequences, $FILTERED_FORWARD is the name of the file containing the filtered forward sequences, and $FILTERED_REVERSE is the name of the file containing the filtered reverse sequences.

1.4. Create a database of human sequences for decontamination. Download the human genome with "wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.28_GRCh38.p13/GCA_000001405.28_GRCh38.p13_genomic.fna.gz" and decompress using gunzip. Then run "bowtie2-build GCA_000001405.28_GRCh38.p13_genomic.fna.gz human".

Note: If the wget command is not found, run "conda install -c anaconda wget" to install it. If the gunzip command is not found, run "conda install -c ostrokach gzip" to install it.

1.5. To remove human contamination, run Kneaddata2 using "kneaddata --bypass-trim --input $FILTERED_FORWARD --input $FILTERED_REVERSE -o $OUTPUT -db $PATH/human"

Note: The variables in the command are as follows: $FILTERED_FORWARD is the filtered forward file from fastp, $FILTERED_REVERSE is the filtered reverse file from fastp, $OUTPUT is the name of the output folder, and $PATH is the absolute path to the database created in step 1.4.

1.6. Prepare the data for general functional and compositional analysis with HUMAnN2 by combining the filtered Kneaddata2 forward and reverse sequences for each sample using "cat $K_FORWARD $K_REVERSE > $COMBINED".

Note: The variables in the command are as follows: $K_FORWARD is the file containing the fastp and Kneaddata2 filtered forward sequences, $K_REVERSE is the file containing the Kneaddata2 fastp filtered reverse sequences, and $COMBINED is the name of the resulting file.

1.7. Run HUMAnN2 on all combined files with "humann2 --input $COMBINED --output $OUTPUT".

Note: The variables in the command are as follows: $COMBINED is the combined file from the previous step and $OUTPUT is the base name of the folder that will contain all of the temporary output files generated as well as the final output tables for the genes and pathways.

1.8. Format the HUMAnN2 data. For each sample, there are typically three files of interest: the "genefamilies.tsv" file (contain UniRef90 genes), the "pathabundance.tsv" file (containing Metacyc pathways), and the "metaphlan_bugs_list.tsv" file (containing the relative abundances of microbes present at the kingdom through species levels and located within the outputted folder). Create three folders, one for each type of file, and move all files for each type into their respective folder. Then run "humann2_join_tables --input $FOLDER --output $OUTPUT" for each folder.

Note: The variables in the command are as follows: $FOLDER is the folder containing the files of interest and $OUTPUT is the resulting table.

1.9. Prepare the data for antimicrobial resistance analysis. If data are paired-end, pair the files by using PEAR's "pear" command in the form of "pear -f $K_FORWARD -r $K_REVERSE -o $PAIRED".

Note: The variables in the command are as follows: $K_FORWARD is the name of the file containing the sample's fastp and Kneaddata2 filtered forward sequences, $K_REVERSE is the name of the file containing its fastp and Kneaddata2 filtered reverse sequences, and $PAIRED is the name of the resulting paired file

1.10. Prepare the antimicrobial resistance database. Download the MEGARes 2.0 database (.fasta file) from https://megares.meglab.org/download/index.php. Then format it as a BLAST database using "makeblastdb -in $FASTA -out $DATABASE -dbtype nucl". If the ".txt" extension is appended to the MEGARes2.0 file, simply delete ".txt" before running the database command.

Note: The variables in the command are as follows: with $FASTA is the name of the downloaded database file and $DATABASE is the name of the newly created database

1.11. Use BLAST with the MEGARes 2.0 database to determine which antibiotic, biocide, and metal resistance genes are expressed in the samples with a command in the form of 'blastn -task megablast -evalue 0.001 -max_target_seqs 1 -query $PAIRED -db $DATABASE -out $OUTPUT -outfmt "6 sseqid slen qseqid bitscore evalue qlen pident mismatch length staxids sscinames scomnames""'.

Note: The variables in the command are as follows: $PAIRED is the filtered, paired file from running PEAR, the $DATABASE is the database created in the previous step, and $OUTPUT is the desired name of the output file

1.12. Reformat the output file.

1.12.1. Open it with a spreadsheet editor.

1.12.2. Create a pivot table such that sseqid values are the rows and the number of times they appear in the table (count) are in the first column and their slen value (how long the sequence is in the database) is in the second.

Note: The initial BLAST output file will not have headers. However, the columns will be in the order specified, meaning the first column will be sseqid and the second slen and so on.

1.12.3. Create a third column that contains sequences per kilobase normalized (rpk) values (the result of dividing the first column by the second and multiplying that dividend by 1000).

1.12.4. Make the header for third column the name of the sample and delete the first (count) and second (slen) columns.

1.12.5. Check that the resulting table has a list of hits (sseqids) as the row names with the header for that column being (sseqid) and another column with the rpk normalized values.

1.12.6. Save the pivoted table as a tab-delimited (.txt) file and put the tables for all samples in the same folder. HUMAnN2 can then be used to merge the tables with the command "humann2_join_tables --input $FOLDER --output $OUTPUT" as before.

1.13. Import the three combined functional tables (AMR genes, UniRef90 genes, and Metacyc pathways) into Qiime2.

1.13.1. First convert them to hdf5 biom format with the command 'biom convert -i $INPUT -o $BIOM --to-hdf5 --table-type="OTU table"'

Note: The variables in the command are as follows: $INPUT is the combined table and $BIOM is the name of the resulting .biom file.

1.13.2. Import the resulting biom files into QIIME2 using "qiime tools import --type 'FeatureTable[Frequency]' --input-format BIOMV210Format --input-path $BIOM --output-path $QZA"

Note: The variables in the command are as follows: $BIOM is the output table from the "biom convert" command and $QZA is the name of the newly created Qiime2 artifact.

1.14. Create a metadata file with all the sample names and their fracking classification status. Format it as described here https://docs.qiime2.org/2020.8/tutorials/metadata/.

1.15. Run diversity analysis through Qiime2 with "qiime diversity core-metrics --i-table $QZA --m-metadata-file $METADATA --p-sampling-depth $DEPTH --output-dir $OUTPUT". Pick a $DEPTH that is no greater than the smallest sample (based on the sum of all of its features) that is to be retained

Note: The variables in the command are as follows: $QZA is the imported table from the previous step, $METADATA is the .txt file containing the sample groupings (HF+/-), $DEPTH is the depth to rarefy (subsample) to account for differences in the number of sequences, and $OUTPUT is the folder containing all of the output files. From this point on, the steps are written with the assumption that the metadata file is "metadata.txt" with a column named "Fracking_Status".

1.16. View the resulting qzv files with view.qiime2.org (Figure 3A).

Note: If more control over the PCoA plot's appearance is desired, the bray_curtis_distance_matrix.qza file can be exported with "qiime tools export --input-path bray_curtis_distance_matrix.qza --output-path $OUTPUT". The distance_matrix.tsv file in that output folder ($OUTPUT) can be used to remake the PCoA in R (Figure 3B).

1.17. Run beta diversity statistics on the resulting Bray-Curtis distance matrix using “qiime diversity beta-group-significance --i-distance-matrix bray_curtis_distance_matrix.qza --p-method permanova --m-metadata-file metadata.txt --m-metadata-column Fracking_Status --o-visualization $OUTPUT”.

Note: The variable in the command is as follows: $OUTPUT being the Qiime2 visualization (.qzv) file that contains the results of the PERMANOVA test.

1.18. Run alpha diversity statistics with the command “qiime diversity alpha-group-significance --i-alpha-diversity observed_features.qza --m-metadata-file metadata.txt --o-visualization $OUTPUT” Repeat this step twice using the evenness_vector.qza and shannon_vector.qza files instead of the observed_features.qza file.

Notes: The variables in the command are as follows: $OUTPUT is the Qiime2 visualization file that contains the results of the statistical test (Kruskal-Wallis).

1.19. Run random forest analysis to determine which features can be used to differentiate samples based on fracking status and how effective the dataset overall is at that task. This can be done through the randomForest package in R or Qiime2. The Qiime2 method is easier (1.19.1), but the R method (1.19.2-1.19.7) gives more control over how the analysis is conducted.

1.19.1. Using Qiime, run “qiime sample-classifier classify-samples-ncv --i-table rarefied_table.qza --m-metadata-file metadata.txt --m-metadata-column Fracking_Status --output-dir $OUTPUT” where $OUTPUT is the resulting folder. See Qiime2’s classification tutorial https://docs.qiime2.org/2020.8/tutorials/sample-classifier/ for more details about this command.

Note: The output folder will contain three files: feature_importance.qza, predictions.qza, and probabilities.qza. See https://docs.qiime2.org/2020.8/tutorials/sample-classifier/ for a description of those files and https://scikit-learn.org/stable/modules/ensemble.html#feature-importance for a description of how accuracy is measured

1.19.2. Get R ready. Activate R by entering “R” into the console. Install the Qiime2R package with “install.packages("remotes”)” followed by remotes::install_github("jbisanz/qiime2R"). Install the randomForest package with install.packages(“randomForest”).

1.19.3. Load the libraries with “library(qiime2R)” and “library(randomForest)”.

1.19.4. Load the data with “asv_table = read_qza(“rarefied_table.qza”)” and “metadata = read.table(metadata.txt, header=T, check.names = F, sep = "\t", quote = "")” The same metadata file that was used with other analyses should work here as well, as long as the header for the first column does not contain a “#”.

Note: Both of those lines store the loaded data in objects, asv_table and metadata, respectively so that they can be easily referenced later.

1.19.5. Format the feature table. Extract the feature frequencies by running "table = asv_table$data". Then transpose it using "t_table = t(table)".

1.19.6. Run random forest with the command "rf = randomForest(t_table, meta$Fracking_Status, ntree=100, votes = T, norm.votes = T, importance = T)".

1.19.7. Extract the results with the following commands: "confusion = rf[["confusion"]]", "importance = rf[["importance"]]", and "votes = rf[["votes"]]". The information in the importance object is equivalent to the feature_importance.qza (albeit with different accuracy metrics), and the information in the votes object is equivalent to the probabilities.qza. The confusion object contains the confusion matrix and the proportion of times samples within a group were classified incorrectly.

Note: See https://cran.r-project.org/web/packages/randomForest/randomForest.pdf for the randomForest's package documentation. The results of the random forest done through R can be visualized with a variety of functions and packages. For example, the varImpPlot function can be used like so "varImpPlot(rf)" to create a variable importance plot (Figure 3C). Such a plot could also be generated using the importance object and ggplot2.

1.20. Repeat steps 1.15 through 1.19 for all functional datasets.

2. Important Considerations

This pipeline can be difficult to run, especially since it assumes familiarity with running programs via command line. Many online tutorials for learning how to use command line programs are available. For example, the one at https://www.learnenough.com/command-line-tutorial/basics goes through some basic commands with pictures. Furthermore, there are a wide variety of possible errors even when following this example pipeline. However, the two most common are essentially "File not found" and "Command not found" errors. The former can be dealt with by checking that the desired input file is present in the folder that the command is being executed in by running the "ls" command if using Mac's Terminal or a Linux operating system or "dir" if using Windows. The latter error is usually the result of the program not being installed properly. This tutorial recommends installing everything through the "conda" command. However, some programs can have conflicting dependencies. Therefore, even if the "conda install" command is run, the program might not install properly, leading to that error. To avoid that, a new environment can be created with "conda create -n $NAME" with $NAME being the name of the new environment and activated with "conda activate $NAME" before the install command is run so that each of the programs described below is in a different environment. To switch between environments, run "conda deactivate" and then activate the desired one as before. Moreover, some of the steps are simply impractical to run on a laptop. Namely, 1.2, 1.3, 1.5, 1.7, and 1.11

will take an inordinately large amount of time if run locally on a laptop. Therefore, if the user has access to a computer cluster, they should be run on that. Alternatively, computing power can be rented through a private company for those steps.

As with lab protocols, contamination is once more an important consideration. Kneaddata2 is used to remove human sequences, and additional genomes can be downloaded and used from NCBI to remove potential contamination from other eukaryotes. Microbial contamination is more challenging to address. After analysis, if any of the samples are revealed to be nearly identical in composition to the negative(s), then the R package decontam should be used to remove the contaminants from the samples. However, this should only be done if there is clear evidence of contamination, as decontam can potentially remove features that are not actually the result of contamination.

Importantly, this is just one example of a potential metatranscriptomics pipeline. The analyses and methods presented here are by no means comprehensive. Furthermore, for the sake of simplicity, it uses Qiime2's "qiime diversity core-metrics" command, which always rarefies the data. However, many bioinformaticists dislike rarefaction as a normalization strategy due to the large amount of data discarded. The same analyses can be run through Qiime2 using a combination of several commands to avoid rarefaction, which can be found on the program's website https://docs.qiime2.org/2020.8/ The goal of normalization is to make samples more comparable to each other. Notably, all of the functional datasets in this pipeline are normalized by the reads per kilobase method to account for the fact that longer genes would be expected to contribute more sequences.

Contrarily, HUMAnN2's taxonomic assignments are normalized by converting their counts to relative abundances. The combined bugs_list file can be used to created cladograms and heatmaps as described here https://github.com/biobakery/biobakery/wiki/metaphlan2#visualize-results, being equivalent to that tutorial's "merged_abundance_table.txt" file. Because HUMAnN2 outputs microbial compositional as relative abundances, this table should not be used as input for QIIME2's "qiime diversity core-metrics" command.

As indicated above, this pipeline does not provide an exhaustive list of all possible analyses that can be performed. Notably, though it is not described in the pipeline proper, it is possible to determine contributors to the antimicrobial resistance genes. To determine contributors for a certain antimicrobial resistance gene, NCBI's seqkit (https://bioinf.shenwei.me/seqkit/usage/) can be used to extract the sequences that mapped to it, with that subset of sequences then being used as input for BLAST with a database containing microbe genomes, e.g. RefSeq, to identify them. See here for information about using BLAST's "update_blastdb.pl" command to acquire databases https://www.ncbi.nlm.nih.gov/books/NBK279680/ If the resistance profile of a specific taxon of interest is desired, all sequences that hit against the MEGARes 2.0 database could be extracted and taxonomically identified. Once that is done, the sequences belonging to that taxon can be searched for in the initial BLAST results (under the qseqid column). After those AMRs are found, the taxon's resistance profile can be visualized using a simple pie chart (Figure 3D). Still, though it does not by any means detail all possible analyses, this example pipeline can be used to investigate hydraulic fracturing's impact on nearby streams through analyzing metatranscriptomics (RNA) data.