Microbial Community Profiling Protocol with Full-length 16S rRNA Sequences and Emu

Kristen D. Curry,[1,4] Sirena Soriano,[2] Michael G. Nute,[1] Sonia Villapol,[2] Alexander Dilthey,[3] and Todd J. Treangen[1,4]

[1]Department of Computer Science, Rice University, Houston, Texas, USA
[2]Center for Neuroregeneration, Department of Neurosurgery, Houston Methodist Research Institute, Houston, Texas, USA
[3]Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[4]Corresponding authors: *Kristen.d.curry@rice.edu*; *treangen@rice.edu*

Published in the Bioinformatics section

16S rRNA targeted amplicon sequencing is an established standard for elucidating microbial community composition. While high-throughput short-read sequencing can elicit only a portion of the 16S rRNA gene due to their limited read length, third generation sequencing can read the 16S rRNA gene in its entirety and thus provide more precise taxonomic classification. Here, we present a protocol for generating full-length 16S rRNA sequences with Oxford Nanopore Technologies (ONT) and a microbial community profile with Emu. We select Emu for analyzing ONT sequences as it leverages information from the entire community to overcome errors due to incomplete reference databases and hardware limitations to ultimately obtain species-level resolution. This pipeline provides a low-cost solution for characterizing microbiome composition by exploiting real-time, long-read ONT sequencing and tailored software for accurate characterization of microbial communities. © 2024 Wiley Periodicals LLC.

Basic Protocol: Microbial community profiling with Emu
Support Protocol 1: Full-length 16S rRNA microbial sequences with Oxford Nanopore Technologies sequencing platform
Support Protocol 2: Building a custom reference database for Emu

Keywords: bioinformatics • community profile • long-read sequencing • microbiome • Oxford Nanopore Technologies • species • 16S rRNA

How to cite this article:
Curry, K. D., Soriano, S., Nute, M. G., Villapol, S., Dilthey, A., & Treangen, T. J. (2024). Microbial community profiling protocol with full-length 16S rRNA sequences and Emu. *Current Protocols*, *4,* e978. doi: 10.1002/cpz1.978

INTRODUCTION

Sequencing of the 16S subunit of the ribosomal RNA (rRNA) gene has been established as a reliable way to characterize diversity in a community of microbes without the cost and complexity of whole genome metagenome sequencing (Johnson et al., 2019). The 16S rRNA gene is approximately 1550 bp and thus targeted amplicon sequencing of this gene with high-throughput short-read sequencing is limited to only a portion of the gene. This constraint ultimately prevents taxonomic distinction between highly similar species and thus short-read 16S rRNA sequencing cannot reliably generate taxonomic profiles

CURRENT PROTOCOLS
A Wiley Brand
Current Protocols e978, Volume 4
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.978

with greater precision than the genus level in most cases (Martínez-Porchas et al., 2016). Recent developments in third-generation sequencing, from providers such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), permit amplification of sequences spanning the entire 16S rRNA gene and provide potential for species-level community profiles from full-length 16S rRNA sequences. ONT technology additionally provides the added benefits of real-time output, offline sequencing capabilities, and portability of its handheld MinION device. However, previous software tools developed for short reads are not equipped to handle the error profiles of ONT reads. Emu is a software tool developed to utilize information from the entire community to overcome this challenge (Curry et al., 2022). The expectation-maximization algorithm within Emu uses a probabilistic model for improved classification when the read assignment is ambiguous, which is likely to occur due to incomplete databases, sequence mutations and sequencing error. Due to new technologies from ONT allowing for portable, low-cost, real-time sequencing and the development of compatible software, we opt to use ONT full-length 16S sequencing and Emu for efficient microbial community profiling.

This protocol is a detailed explanation of how to install and run Emu for 16S rRNA sequences. Support Protocol 1 additionally walks through the steps required to extract DNA from fecal samples, perform 16S library preparation, and use an ONT MinION to generate 16S rRNA sequences. If a different type of sample is desired, the user can replace the DNA extraction from fecal samples steps with their desired protocol and begin this protocol at the 16S library preparation phase. We have also included Basic Protocol 2, which explains the curation process of a custom reference database for taxonomic profiling with Emu. Finally, detailed information on critical parameters and troubleshooting is also provided with the intent of making this pipeline as straightforward as possible.

BASIC PROTOCOL

MICROBIAL COMMUNITY PROFILING WITH EMU

This protocol begins with a fastq file of basecalled 16S rRNA microbiome sequencing reads and describes the computational steps taken to achieve a microbial community profile with the software tool Emu and its default bacterial reference database. Options of other previously curated Emu databases are described in Critical Parameters and a description of how to build a custom database is described in Support Protocol 2. This protocol was developed for full-length 16s rRNA sequences but can also be used for 16S rRNA short reads for any selected hypervariable region(s). If a different targeted amplicon is desired, this protocol can also be used by replacing the reference database with the corresponding region sequences. However, it is important to note that custom databases and selected amplicons aside from the 16S rRNA gene have not been validated with Emu.

Necessary Resources

The required resources for this protocol are 16S rRNA sequences and a computer. 16 GB of available RAM is recommended, although depending on the depth and complexity of the supplied fastq of sequences, this may need to be increased. The steps below use a command line interface (CLI), of which a Unix or Linux system is assumed.

1. Download the default Emu database from OSF (*https://osf.io/56uf7/*) under osfstorage/emu-prebuilt/emu.tar.
2. Open a command line interface (CLI) and set environment variable EMU_DATABASE_DIR to the location of the downloaded Emu database. <database_location> is the complete path to the directory containing the two necessary database files: species_taxid.fasta and taxonomy.tsv.

```
$export EMU_DATABASE=<database_location>
```

A

```
[M::mm_idx_gen::1.935*1.24] collected minimizers
[M::mm_idx_gen::2.396*1.57] sorted minimizers
[M::main::2.402*1.57] loaded/built the index for 49243 target sequence(s)
[M::mm_mapopt_update::2.417*1.57] mid_occ = 14113
[M::mm_idx_stat] kmer size: 15; skip: 10; is_hpc: 0; #seq: 49243
[M::mm_idx_stat::2.427*1.56] distinct minimizers: 462637 (45.14% are singletons);
        average occurrences: 28.757; average spacing: 5.572; total length: 74133434
[M::worker_pipeline::2.847*1.65] mapped 4 sequences
[M::main] Version: 2.24-r1122
[M::main] CMD: minimap2 -ax map-ont -t 3 -N 50 -p .9 -K 500000000 -o
        ./results/full_length_emu_alignments.sam
        /emu_database/species_taxid.fasta emu/example/full_length.fa
[M::main] Real time: 2.930 sec; CPU: 4.783 sec; Peak RSS: 0.473 GB
Unassigned read count: 0
Assigned read count: 4
Number of EM iterations: 3
```

B

tax_id	abundance	species	genus	family	order	class	phylum
2420510	0.5	Sphingobacterium puteale	Sphingobacterium	Sphingobacteriaceae	Sphingobacteriales	Sphingobacteriia	Bacteroidetes
220927	0.25	Mycobacterium saskatchewanense	Mycobacterium	Mycobacteriaceae	Corynebacteriales	Actinobacteria	Actinobacteria
1310	0.25	Streptococcus sobrinus	Streptococcus	Streptococcaceae	Lactobacillales	Bacilli	Firmicutes
unassigned	0						

C

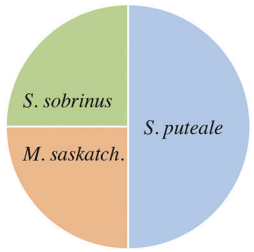

Figure 1 Example output. (**A**) Terminal output from the provided test example Emu run. Text may vary slightly between runs. (**B**) Expected content in the generated Emu test profile "full_length_rel-abundance.tsv" within the results directory. (**C**) Corresponding visualization for the predicted taxonomic community profile.

3. Install bioconda (*https://bioconda.github.io/*).
4. Install Emu with bioconda via CLI command.

```
$conda install -c bioconda emu
```

5. Test installation with provided sample data. To do so, first download sample data through the github repository.

```
$git clone https://github.com/treangenlab/emu
```

   Then, run Emu on the full_length.fa file in the examples directory.

```
$emu abundance emu/example/full_length.fa
```

6. Verify test run results are as expected (Fig. 1). A new folder titled "results" should be generated containing a single file titled "full_length_rel-abundance.tsv" with the estimated species relative abundance.
7. To prepare for Emu with your sequences, first establish if Emu default parameters are viable for your study or if altered parameters are required (see Critical Parameters).
8. Run Emu on a single barcode of sequences via CLI command, including any desired parameter settings.

```
$emu abundance <reads.fastq> (where <reads.fastq> is the file of sequences)
```

9. Repeat step 8 for all samples in the study.

FULL-LENGTH 16S RRNA MICROBIAL SEQUENCES WITH OXFORD NANOPORE TECHNOLOGIES SEQUENCING PLATFORM

SUPPORT PROTOCOL 1

This protocol includes the steps necessary for generating targeted amplicon sequences of the full-length 16S rRNA gene from microbes in the sampled community in preparation for the steps in the Basic Protocol (Fig. 2). The steps and materials below assume acquisition of a fecal sample, which is often used for the characterization of the gut microbiome. If a different sampling environment is required, please refer to Oxford Nanopore Technologies documentation to obtain the appropriate DNA extraction kit and begin this protocol at the 16S library preparation step. Indications have also been included for the sequencing of a mock microbial community for the purpose of benchmarking, if needed.

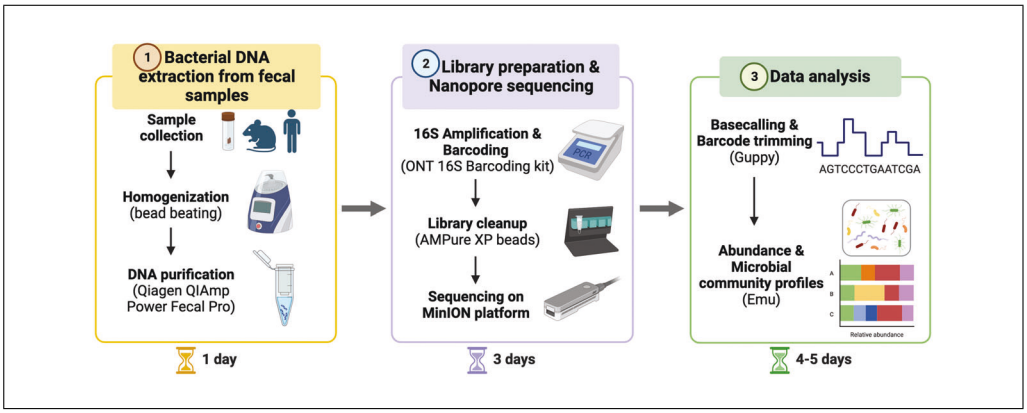

Figure 2 Experimental design. 1) Bacterial DNA extraction from fecal samples 2) Library preparation and Oxford Nanopore sequencing 3) Data analysis and Emu.

Materials

QIAamp PowerFecal Pro DNA kit (Qiagen, cat. no. 51804)
Microbial community standard II, optional (ZymoBIOMICS, cat. no. D6310)
16S Barcoding kit 24 v14 (Oxford Nanopore Technologies, cat. no. SQK-16S114.24)
LongAmp hot start taq 2X master mix (New England Biolabs, cat. no. M0533S or M0533L)
Nuclease-free water (Promega, cat. no. P1195 or equivalent)
70% ethanol, in nuclease-free water
10 mM Tris·HCl pH 8.0, with 50 mM NaCl
Qubit 1× dsDNA high sensitivity assay kit (Invitrogen, cat. no. Q33231)

Micropipettes (1000, 200, 20, and 10 µl)
Multichannel pipette, 10 µl, optional
Filter tips for micropipettes (1000 µl, 200 µl, 20 µl, 10 µl)
1.5-ml DNA LoBind tubes (Eppendorf, cat. no. 022431021)
0.2-ml PCR tubes (Fisherbrand, cat. no. 14-230-225 or equivalent)
Benchtop centrifuge
Vortex
FastPrep-24 homogenizer (MP Biomedicals, cat. no. 116004500 or equivalent)
Microvolume spectrophotometer (DeNovix, catn. no. DS-11 or equivalent)
Thermal cycler (Bio-Rad T100, cat. no. 1861096 or equivalent)
Magnetic separation rack, 12 tube (New England Biolabs, cat. no. S1509S)
Rotating mixer
Qubit fluorometer and assay tubes (Invitrogen, cat. no. Q32856)
Flow cell R10.4.1 (Oxford Nanopore Technologies, cat. no. FLO-MIN114)
MinION sequencing device (Oxford Nanopore Technologies, cat. no. MIN-101B)

Software: MinKNOW software and either Guppy or Dorado basecaller (Oxford Nanopore Technologies website)

DNA extraction from fecal samples

1. Collect fecal samples in sterile tubes. Samples may be stored at −80°C until DNA extraction is performed.
2. Follow the manufacturer's instructions for the Qiagen QIAamp PowerFecal Pro DNA kit.

Recommended starting material for the QIAamp PowerFecal Pro kit is 250 mg of fecal sample. If using the Zymobiomics Microbial Community Standard II, use 250 µl of the standard as starting material instead.

For the initial homogenization of the fecal samples, the FastPrep-24 bead-beater can be used for 1 min at 6.5 m/s, let cool for 1 min, repeating twice.

We recommend eluting the DNA in 100 µl of Solution C6, followed by measuring the DNA concentration on a microvolume spectrophotometer.

16S Library preparation, Nanopore Sequencing, and Basecalling

3. Follow the manufacturer's instructions for the ONT 16S Barcoding kit 24 v14 to prepare the 16S libraries.
4. Load the prepared library on the R10.4.1 flow cell, following ONT's directions.
5. On MinKNOW software, select start sequencing. Fill the experiment name and select the sequencing kit (SKQ-16S114.24). Select appropriate sequencing options and start.

   For sequencing of full length 16S with 24 barcodes on a new flow cell, we suggest selecting the following parameters: Fast basecalling, Barcoding: ON, Trim barcodes: ON, Min. barcoding score: 60 (default), Filtering: ON, $Q = 7$ (default).

   Alternatively, basecalling and barcode trimming can be performed after the sequencing experiment has been completed with Guppy or Dorado. On MinKNOW, turn the basecalling option to OFF. If using Guppy, use the Guppy basecalling tool with remove barcodes from demultiplexed sequences activate (–trim_barcodes).

BUILDING A CUSTOM REFERENCE DATABASE FOR EMU

SUPPORT PROTOCOL 2

Emu provides the functionality to build a custom reference database for relative abundance estimations. To construct a custom Emu database, a fasta file of reference sequences and their taxonomic lineages are required. Current acceptable formats for taxonomy include NCBI names.dmp and nodes.dmp files or a tab-separated file where each row is a unique taxonomic lineage, and the taxonomic id is the first column. In addition to sequences and taxonomy, a file mapping each sequence ID to the assigned taxonomic ID is required. The output will be an Emu database as a directory containing the two files required for Emu abundance estimation calls. To use this database for an Emu abundance call, either set the environment variable EMU_DATABASE_DIR to the directory containing the custom database or define this path using the –db parameter.

Necessary Resources

Computer used in Basic Protocol
Nucleotide reference sequences in a fasta file
A two-column tab-separated file defining the taxonomic ID for each sequence ID
Taxonomy, either in the form of NCBI names.dmp and nodes.dmp file, or a tab-separated file

1. Confirm all sequence IDs in sequence fasta file are included in the sequence ID to taxonomic ID mapping file.
2. Confirm all sequence classification taxonomic IDs (mapping file) are included in the taxonomy file(s).
3. If using a taxonomy list, ensure there are no repeats of identical entries.
4. Use Emu function to build database. Below, <db_name> is the desired folder name for the constructed database; <database.fasta> is the fasta file containing reference sequences; <seq2taxid.map> is the tab-separated file mapping each sequences id in the fasta to a taxonomic id; <dir-to-names/nodes.dmp> is the local directory containing NCBI names.dmp and nodes.dmp or <taxonomy.tsv> for the tab-separated taxonomy file.

5. If NCBI taxonomy:

```
$emu build-database <db_name> --sequences <database.fasta> --seq2tax <seq2taxid.map> --ncbi-taxonomy
    <dir-to-names/nodes.dmp>
```

   If taxonomy as a tab-separated file:

```
$emu build-database <db_name> --sequences <database.fasta> --seq2tax <seq2taxid.map> --taxonomy-list
    <taxonomy.tsv>
```

6. If the terminal output states "database creation successful" then a new folder at path <db_name> is generated with two files: taxonomy.tsv and species_taxid.fasta, and the custom database is built.

Understanding Results

In the <sample>_rel-abundance.tsv file, the first two columns contain a list of taxonomic IDs found in the sample and their corresponding relative abundances. The complete taxonomic lineage for each taxonomic ID is listed in the subsequent columns of each row. If the –keep-counts parameter is set to true, estimated counts are also included in results, which is simply the number of classified reads multiplied by the relative abundance. The last row in each results table is the unassigned row. Since relative abundance is calculated with unassigned reads discarded, the abundance of this column will always be zero. However, if the –keep-counts parameter is set to true, the estimated counts column will display the number of unassigned reads. A read is unassigned if it does not have any hits to the database based on the minimap2 settings. Figure 1 shows an example of a terminal output, relative abundance table, and the corresponding abundance pie chart of a successful Emu abundance call. There is an additional setting to –keep_read_assignments if the taxonomic classification probability distribution of each read is desired; further details on this process is shown in Critical Parameters. If you wish to benchmark the complete protocol, we recommend sequencing a mock community (ZymoBIOMICS Microbial Community Standards; *https://zymoresearch.eu/collections/zymobiomics-microbial-community-standards*) and verifying the reported community profile aligns with the known community.

COMMENTARY

Background Information

We opt for ONT sequencing due to its low cost, quick results, portability, and feasibility to perform sequencing in a laboratory without extensive equipment (Petersen et al., 2019). In addition, as basecalling algorithms and hardware technology continue to improve the accuracy of these devices, further bioinformatics platforms are being developed for specific scientific questions, which ultimately diversifies the utility of metagenome sequencing and ONT devices (Wick et al., 2019). Our mock community analyses found that previous taxonomic classification algorithms were unable to produce accurate species-level community profiles for 16S rRNA sequences – short read sequences did not have the length for species-level precision, while ONT long reads contained too high of error profiles. Thus, we developed Emu (Curry et al., 2022). Emu leverages the fundamental idea behind the error model in MetaMaps, which is to use read mapping to identify multiple candidates of taxonomic assignment for a given read then apply an expectation-maximization algorithm to adjust the relative confidence of the assignment (Dilthey et al., 2019). Since MetaMaps is designed for whole genome sequencing and thus includes approximate alignments and reference mapping locations, we found MetaMaps not suitable for 16S rRNA reads and pursued development of an entirely new algorithm. We developed Emu with the intent of analyzing ONT full-length 16S rRNA sequences for a reduced-cost pipeline of acquiring species-level microbial community composition, as described in this protocol.

This protocol is, however, a database-driven approach, which therefore limits the accuracy of classification to the extent by which the utilized database represents the species present in the sample. This protocol also inherits limitations that are implicit to 16S rRNA-based community profiles: community

profiles are only in relative abundance to each other rather than absolute counts, the profile is skewed by bias of differing quantities of 16S rRNA gene copies per genome, and downstream analysis is limited to since only the 16S rRNA gene is sequenced. Additionally, Emu does not give a single classification for each sequence. Rather, the method returns a community level profile, with the option of also obtaining a classification probability distribution for each read. Therefore, if the user requires classifying each read individually, further calculations are required. Yet when an accurate microbial community profile from 16S rRNA sequences is desired, recent studies are also favoring ONT sequencing paired with Emu (Petrone et al., 2023; Stevens et al., 2023).

Critical Parameters

With regards to sequencing (Support Protocol 1), the quality and quantity of the DNA is crucial in the full-length 16S sequencing pipeline. We recommend the concentration and purity of the samples following DNA extraction to be evaluated by microvolume spectrophotometry (NanoDrop, DeNovix). DNA purity is evaluated with the 260/280 absorbance ratio, which is in the range of 1.8 – 2 for good quality DNA. The 260/230 ratio is expected to be between 2 and 2.2 and lower ratios are indicative of contamination with organic compounds. Fluorescence based methods (Qubit, QuantiFluor) can also be used to assess the DNA concentration with more accuracy than absorbance. For sequencing of the full-length 16S rRNA gene, the integrity of the DNA strands is also critical, and care must be taken during the DNA extraction steps to avoid fragmentation. If concerns in this area are present, an additional step to assess fragment length can be taken prior to library prep with an instrument such as the Fragment Analyzer system.

As for Emu (Basic Protocol), the single parameter with the largest influence on the results is the selected database (–db). Sequence classifications are limited to only taxonomies that are in the database, thus if a species is in the sample but not in the database, it will likely be classified as the most similar species that is present in the database. However, increasing the size of the database can also have negative implications. With a larger database comes more opportunity for errors within the database and thus may lead to misclassifications and a heavier computational requirement. The Emu default database was curated to be a balance between these two extremes; however, when working with communities with known reference sequences specific to the sampled environment, it may be advantageous to construct a custom database accordingly. To construct a custom database, please follow Support Protocol 2 described in this article. In addition, two larger databases that have been previously curated for Emu (RDPv11.5 and SILVA v138.1) can also be downloaded from the same OSF location as the default Emu database (*https://osf.io/56uf7/*) (Cole et al., 2014; Quast et al., 2013). Note that these databases contain more sequences and more incomplete taxonomic lineages than the default database, which may require more computational resources and more attention to taxonomic gaps in downstream analyses.

A second influential parameter within Emu is the minimum abundance threshold (–min-abundance). Due to the nature of the EM algorithm, the estimated community composition often comprises a long tail of species with extremely low abundances. A default minimum abundance threshold parameter of 0.0001 has been set, such that relative abundance estimates below this value are deemed not present in the sample. If the input sequencing reads contains more than 100,000 reads, an additional composition estimate will be returned with a lower (more inclusive) minimum abundance threshold that is the equivalent of 10 reads. The user can then decide which profile to use based on the needs of the study. In the situation where the input sample has under 1000 reads, only one composition estimate is returned with the minimum abundance threshold is set to the equivalent of 1 read. A user may additionally alter this minimum abundance threshold parameter; a lower threshold keeps lower abundance species at the cost of increasing false positives, while a larger threshold may lead to false negatives of low abundant species.

Table 1 includes a list of parameter settings for Emu. The –type parameter is directly passed into the minimap2 alignment call within Emu to define the type of sequencer used to generate the reads. This impacts the alignments generated but does not alter the EM algorithm. There are also 4 parameters that can be used to include additional information in the generated output. The –keep-files keeps the generated minimap2 alignment in the output directory in the same file format. The –keep-counts parameter adds an additional column to the community profile tsv file to express the estimated number of counts

Table 1 Parameter Guide for Emu

Parameter	Function	Reason to apply
Type	Denote type of sequencer for minimap2 alignments	ONT: map-ont (default); PacBio: map-pb, Short-read: sr
Min-abundance	The abundance threshold where only estimated relative abundances above this value are marked as present in the sample	If study requires to detect species with relative abundance below 0.0001, decrease this value; if false positives are detrimental to your study, a more conservative approach would be to increase this value
db	Path to provided reference database	The default Emu database contains the full-length 16S rRNA gene from characterized bacteria and archaea; if specific species of interest are not in the default database, a custom database or addition of reference sequences to the default database is recommended; if targeted amplicon is not the 16S rRNA gene (i.e., 18S), an appropriate database is recommended
Keep-files	Keeps the output from minimap2 alignment in the sam file format in the output directory	Set this parameter to true if downstream analysis utilizing alignments is desired
Keep-counts	Includes an estimated read count (in addition to relative abundance) in the generated output relative abundance file, this is calculated by multiplying the relative abundance by the number of classified reads; a read is considered unclassified if no minimap2 alignments are generated for the read	Set this parameter to true if read counts for each species classification is desired
Keep-read-assignments	Creates an additional file containing the classification distribution for each read	Set this parameter to true if read-level classification is desired
Output-unclassified	Creates an additional fasta file of the sequences that did not generate any alignments during the minimap2 phase	Set this parameter to true if unclassified sequences are desired for downstream analysis
Threads	Number of threads used by minimap2	Increase this number if your computing system allows it to decrease the run time

for each taxonomy. These values are generated by multiplying the relative abundance by the number of classified reads. The "unclassified" read count is the number of reads that did not generate an alignment to the supplied database with minimap2 and its abundance will always be 0, as unclassified reads are not considered for relative abundance. The –keep-read-assignments flag generates an additional output file expressing the taxonomic classification likelihoods for each read. In this file, each row is a single read, and each column is a unique taxonomic id. The cells designate the likelihood that the read emanates from the corresponding taxonomic id such that each row sums to 1. Finally, the –output-unclassified flag generates an additional sequence file of all the reads that were left unclassified, which is defined here as reads without minimap2 alignments returned. The final parameter is –threads, which defines the number of threads used for the minimap2 alignment step.

Troubleshooting

Since nanopore sequencing can be challenging in the beginning, we have included a table of common issues that arise during 16S rRNA library prep (Table 2) and nanopore sequencing (Table 3). We also included a table of issues and solutions that may arise during the Emu installation or abundance estimation processes (Table 4).

Advanced Parameters

More advanced parameters for Emu include the –N and –K parameters, which are

Table 2 Troubleshooting Guide for 16S Library Preparation

Problem	Possible cause	Solution
No/ low PCR amplification	Not enough DNA	Check that at least 10 ng of DNA are added to each reaction
		Consider increasing the number of PCR cycles
	Low DNA purity (OD 260/280 < 1.8 and/or OD 260/230 < 2)	Add a DNA clean-up step prior to the PCR amplification
Low yield after AMPure beads clean-up	DNA lost due to low concentration of AMP beads	Mix thoroughly the AMPure beads and pipette out before they settle at the bottom of the stock tube
	The ethanol used for the washes was <70%	Ensure that the ethanol 70% is made fresh
	Pellet over dried resulting in decreased elution efficiency	Do not let the pellet dry to the point of cracking

Table 3 Troubleshooting Guide for Nanopore Sequencing

Problem	Possible cause	Solution
Flow cell check fails	The number of active pores <800	Check flow cell expiration date Ensure that the flow cells are stored at 4°C, inadequate storage temperature may result in the flow cell to be damaged irreversibly
The number of active pores is significantly lower when sequencing starts compared to the active number of flow cell check / the number of pores sequencing is low	Air was introduced in the flow cell	Remove any air bubble present by removing a small volume of buffer prior to adding the primer mix Avoid the introduction of new air bubbles when priming the flow cell
	The library had contaminant that damaged the pores of the flow cell	Consider adding extra library purification steps to remove any contaminant present
	Not enough library loaded	Make sure that at least 50 fmol of library was loaded into the flow cell

directly passed in to the minimap2 alignment step to alter the number of secondary alignments retained for each read and the mini-batch size, respectively. Each of these can be used to reduce the memory consumption. Additionally, there are two parameters in Emu to alter the output directory name and output file name(s) rather than the default settings, as shown in Table 5.

Further Analysis

Two supplementary functions included when Emu is installed are the collapse-taxonomy and combine-outputs scripts. Collapse-taxonomy is used to modify a taxonomic profile generated by Emu to a less specific taxonomic rank. For example, a default Emu community profile at the species level could be modified to the family level. This function works by summing all abundances for species under the same taxonomic family and removing the genus and species level information from the table. This function can alternatively be used at the genus, order, class, phylum, or superkingdom rank instead. Note this function will only work on Emu profiles that contain the desired rank in the header (first row). The collapse-taxonomy function can be called on an Emu profile tsv file at <file_path> and rank <rank> with this command.

```
$emu collapse-taxonomy
  <file_path> <rank>
```

The combine-outputs script is used to combine multiple Emu profiles into a single table, which is often desired for further statistical analyses or plot generation software tools. This function will take all the Emu profiles

Table 4 Troubleshooting Guide for Emu

Problem	Possible cause	Solution
Bioconda install unsuccessful	Dependency conflict within the environment	Create a new conda environment with python version 3.6+ or install directly from gitlab
Dependency package is not found	Required dependencies are not installed	Check that all dependencies in the environment.yaml file have been installed properly
	Required dependencies are installed at a different path	Emu search for dependencies in the path for "python3", ensure dependencies are installed at this path
Emu is taking too long	The sequence set is large, the database is large, and/or there are many strong alignments between the two sequence sets	Down sample the input sequences, decrease the complexity of the database, or increase the thread usage on the machine if available
Emu is taking up too much space	The sequence set is large, the database is large and/or there are many strong alignments between the two sequence sets	Down sample the input sequences, decrease the complexity of the database, or decrease the "N" parameter which restricts the number of alignments kept for each read
Expected species are not in results	Expected species are not in database	Add expected reference sequences to the database
	Primers were not able to amplify species	Primers can have limited degeneracy; update primer design

Table 5 Advanced Parameters for Emu

Parameter	Function	Reason to apply
N	Maximum number of minimap2 secondary alignments for each read	Default is 50, can reduce this to reduce time and/or memory consumption at the cost of retaining less information for the re-estimation step
K	Minibatch size for mapping in minimap2	Adjust the memory consumption without altering results
output-dir	Define directory for output results	An output directory other than "./results" at the current path is desired
Output-basename	Define basename for all output files	A different stem filename is desired for the output files than the input fastq file

from a provided directory (detected by containing "rel-abundance" in the file name) and generate a single table where each row is a different taxonomic lineage and each column is a different sample. The entries in the cell then describe the relative abundance of the described taxonomic lineage for the specified sample. The taxonomic rank used as the most specific rank is defined when calling the script. Again, the <rank> must be included in the headers of the relative abundance files.

```
$emu combine-outputs
  <directory_path> <rank>
```

A few metrics for comparing microbiome taxonomic profiles diversity are commonly used. One approach is to establish the diversity within each sample, often called alpha diversity, through metrics such as Simpson (Simpson, 1949), Shannon (Shannon, 1948), or Chao (Chao et al., 2016). If instead, a comparison between cohorts of samples is desired, a Principal Coordinate Analysis (PCoA) and analysis of variance (ANOVA) test can be conducted to determine if the separation between the taxonomic profiles of communities of differing cohorts is statistically significant. Further analysis to determine multivariable association between specific taxonomy and metadata can also be conducted.

Time Considerations

DNA extraction takes approximately 4 hr for a set of 24 samples. Preparation of the 16S library, including the amplification, barcoding, cleanup, and quantification steps can take from 8 to 10 hr. Depending on the desired depth of sequencing, this step may take up to 48 hr. The download and installation of Emu are expected to take a few minutes. Depending on the read depth, complexity of the sample, and number of computational cores used, determining the relative abundance with Emu can range from 1 to 24+ hr.

Acknowledgments

This work has been supported by Jürgen Manchot Foundation and Deutsche Forschungsgemeinschaft (DFG) award 428994620 (A.D.). Computational support and infrastructure were provided by the Centre for Information and Media Technology (ZIM) at the University of Düsseldorf (Germany). S.V. was supported in part by NIH grant R21NS106640 from the National Institute for Neurological Disorders and Stroke (NINDS). S.S. was funded by the Houston Methodist NeuralCODR Fellowship program. K.D.C. was supported in part by Ken Kennedy Institute Computational Science and Engineering Graduate Recruiting Fellowship, Rice University Wagoner Foreign Study Scholarship, and the Chateaubriand Fellowship. M.G.N., and T.J.T. were supported in part by NIH grant P01-AI152999 from the National Institute of Allergy and Infectious Diseases (NIAID). K.D.C., M.G.N., and T.J.T. were supported by the NSF MIM Universal Rules of Live (URoL) grant (EF-2126387, PI Treangen). T.J.T was also supported in part by the NSF CAREER award IIS-2239114 (PI Treangen). We would additionally like to acknowledge all the co-authors on the Emu publication for their contributions to the Emu method: Qi Wang, Michael G. Nute, Alona Tyshaieva, Elizabeth Reeves, Qinglong Wu, Enid Graeber, Patrick Fizner, Werner Mendling, and Tor Savidge, as well as technical support provided by Bryce Kille and Nicolae Sapoval.

Author Contributions

Kristen D. Curry: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; writing—original draft. **Sirena Soriano:** Conceptualization; data curation; formal analysis; investigation; methodology; validation; visualization; writing—original draft. **Michael G. Nute:** Methodology; software; writing—review and editing. **Sonia Villapol:** Conceptualization; funding acquisition; methodology; resources; supervision; validation; visualization; writing—review and editing. **Alexander Dilthey:** Conceptualization; funding acquisition; investigation; methodology; software; validation; writing—review and editing. **Todd J. Treangen:** Conceptualization; funding acquisition; investigation; methodology; project administration; resources; supervision; writing—review and editing.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Emu and all associate code are available on GitHub (*https://github.com/treangenlab/emu*). Emu can be installed via Bioconda (*https://anaconda.org/bioconda/emu*).

Literature Cited

Chao, A., Chiu, C.-H., & Jost, L. (2016). Phylogenetic diversity measures and their decomposition: A framework based on hill numbers. In R. Pellens & P. Grandcolas (Eds.), *Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis* (pp. 141–172). Springer International Publishing. https://doi.org/10.1007/978-3-319-22461-9_8

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633–D642. https://doi.org/10.1093/nar/gkt1244

Curry, K. D., Wang, Q., Nute, M. G., Tyshaieva, A., Reeves, E., Soriano, S., Wu, Q., Graeber, E., Finzer, P., Mendling, W., Savidge, T., Villapol, S., Dilthey, A., & Treangen, T. J. (2022). Emu: Species-level microbial community profiling of full- length 16S rRNA Oxford Nanopore sequencing data. *Nature Methods*, *19*(7), 845–853. https://doi.org/10.1038/s41592-022-01520-4

Dilthey, A. T., Jain, C., Koren, S., & Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature Communications*, *10*(1), 3066. https://doi.org/10.1038/s41467-019-10934-2

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 1. https://doi.org/10.1038/s41467-019-13036-1

Martínez-Porchas, M., Villalpando-Canchola, E., & Vargas-Albores, F. (2016). Significant loss of

sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon*, *2*(9), e00170. https://doi.org/10.1016/j.heliyon.2016.e00170

Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M., & Tsongalis, G. J. (2019). Third-generation sequencing in the clinical laboratory: Exploring the advantages and challenges of nanopore sequencing. *Journal of Clinical Microbiology*, *58*(1), e01315–e01319. https://doi.org/10.1128/jcm.01315-19

Petrone, J. R., Rios Glusberger, P., George, C. D., Milletich, P. L., Ahrens, A. P., Roesch, L. F. W., & Triplett, E. W. (2023). RESCUE: A validated Nanopore pipeline to classify bacteria through long-read, 16S-ITS-23S rRNA sequencing. *Frontiers in Microbiology*, *14*, 1201064. https://doi.org/10.3389/fmicb.2023.1201064

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

Stevens, B. M., Creed, T. B., Reardon, C. L., & Manter, D. K. (2023). Comparison of Oxford Nanopore Technologies and Illumina MiSeq sequencing with mock communities and agricultural soil. *Scientific Reports*, *13*(1), 9323. https://doi.org/10.1038/s41598-023-36101-8

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 4148. https://doi.org/10.1038/163688a0

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*(1), 129. https://doi.org/10.1186/s13059-019-1727-y