METHOD ARTICLE

REVISED The National Ecological Observatory Network's soil metagenomes: assembly and basic analysis [version 2; peer review: 1 approved, 2 approved with reservations]

Zoey R. Werbin[1], Briana Hackos[2], Jorge Lopez-Nava[3], Michael C. Dietze[4], Jennifer M. Bhatnagar[1]

[1]Department of Biology, Boston University, Boston, MA, 02215, USA
[2]Department of Mathematics, University of Colorado, Boulder, Boulder, CO, 80309, USA
[3]Department of Mathematics, Swarthmore College, Swarthmore, PA 19081, USA
[4]Department of Earth & Environment, Boston University, Boston, MA, 02215, USA

v2 **First published:** 19 Apr 2021, **10**:299
https://doi.org/10.12688/f1000research.51494.1
Latest published: 23 Mar 2022, **10**:299
https://doi.org/10.12688/f1000research.51494.2

Abstract

The largest dataset of soil metagenomes has recently been released by the National Ecological Observatory Network (NEON), which performs annual shotgun sequencing of soils at 47 sites across the United States. NEON serves as a valuable educational resource, thanks to its open data and programming tutorials, but there is currently no introductory tutorial for accessing and analyzing the soil shotgun metagenomic dataset. Here, we describe methods for processing raw soil metagenome sequencing reads using a bioinformatics pipeline tailored to the high complexity and diversity of the soil microbiome. We describe the rationale, necessary resources, and implementation of steps such as cleaning raw reads, taxonomic classification, assembly into contigs or genomes, annotation of predicted genes using custom protein databases, and exporting data for downstream analysis. The workflow presented here aims to increase the accessibility of NEON's shotgun metagenome data, which can provide important clues about soil microbial communities and their ecological roles.

Keywords

metagenomics, microbial ecology, soil microbiome, tutorial, workflow

Open Peer Review

Approval Status ? ? ✓

	1	2	3
version 2 (revision) 23 Mar 2022			✓ view
version 1 19 Apr 2021	? view	? view	

1. **Naupaka Zimmerman**, University of San Francisco, San Francisco, USA
2. **William Nelson**, Pacific Northwest National Laboratory, Richland, USA
3. **Olubukola Oluranti Babalola**, North-West University, Mmabatho, South Africa

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Bioinformatics gateway.

Corresponding author: Zoey R. Werbin (zrwerbin@bu.edu)

Author roles: Werbin ZR: Conceptualization, Data Curation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hackos B**: Data Curation, Methodology, Software, Writing – Original Draft Preparation; **Lopez-Nava J**: Software, Visualization; **Dietze MC**: Resources, Supervision, Writing – Review & Editing; **Bhatnagar JM**: Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: ZRW is funded by the National Science Foundation (NSF) Graduate Research Fellowship Program. ZRW, MCD, and JMB are funded by the NSF Macrosystems Biology Program (Award# 1638577). BH and JLN are funded by the BU Bioinformatics Research and Interdisciplinary Training Experience (BRITE) NSF-REU program (Award #1949968).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Werbin ZR, Hackos B, Lopez-Nava J *et al.* **The National Ecological Observatory Network's soil metagenomes: assembly and basic analysis [version 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2022, **10**:299 https://doi.org/10.12688/f1000research.51494.2

First published: 19 Apr 2021, **10**:299 https://doi.org/10.12688/f1000research.51494.1

REVISED **Amendments from Version 1**

We have reorganized each section to provide clearer background and rationale for each analysis step, with more detailed explanations of how each tool performs specifically on soil metagenomics data. We changed the main Snakemake workflow to a new pipeline, metaGEM (Zorilla *et al.*, 2021). This pipeline includes support for high-performance computing clusters, which is expected to help with scaling up to the increasing number of NEON samples. The raw sequence quality control step is now carried out by a single software, fastp (Chen *et al.*, 2018). The taxonomy assignment step no longer requires downloading large reference databases, and instead calls Kraken2 (Wood *et al.*, 2019) remotely through the Toolchest R package (Cai & Lebovic 2021). The genome-binning step now describes a multi-tool programmatic option as an alternative to the interactive approach. Figures 1, 2, and 3 have been revised to reflect the changes in software tool output.

Jorge Lopez-Nava helped implement and test the new bioinformatics pipeline and create figures and is therefore, added as a new author and credited with Software and Visualization.

We thank the two reviewers for thorough and helpful feedback, and we hope this revision improves the utility of this manuscript.

Any further responses from the reviewers can be found at the end of the article

Introduction

The soil microbiome is responsible for key ecological processes, such as decomposition and nitrogen cycling (Allison *et al.*, 2013). One powerful tool for studying the soil microbiome is shotgun metagenomic sequencing, in which all of the genetic material within the DNA extract of a soil sample is sequenced at once, without targeting specific organisms (Quince *et al.*, 2017; Pérez-Cobas *et al.*, 2020). The largest publicly available sequencing dataset of this type is updated annually by the National Ecological Observatory Network (NEON), which monitors ecological conditions at 47 terrestrial sites spanning 20 ecoclimatic domains across the US and its territories (Keller *et al.*, 2008). NEON is funded by the National Science Foundation (NSF), and collects soil samples and releases shotgun metagenomics data annually.

To date, the NEON soil metagenomics data can only be accessed in two formats: as completely raw reads released by NEON, or as processed files through the default protocols of the MG-RAST storage server. Neither format is suitable for most metagenomic analyses, which generally answer scientific questions using custom data processing pipelines that use specific algorithms and targeted reference databases (Ladoukakis *et al.*, 2014; Quince *et al.*, 2017). However, the hyperdiversity of soil ecosystems can pose a challenge for even the most cutting-edge genomic software: retrieving complete bacterial genomes is especially difficult from soil samples (Sieber *et al.*, 2018), and up to 95% of soil DNA reads cannot be identified to the genus level (Méric *et al.*, 2019). To facilitate future scientific analysis, we present a workflow for taking raw soil sequences and generating a processed dataset that can be linked to other NEON data products, which include soil biogeochemistry, root measurements, or aboveground plant communities.

NEON data is a valuable resource for ecology and bioinformatics, thanks to its open access software, robust documentation, and educational resources (Jones, 2020). The pipeline that we present here is designed to complement existing NEON educational resources, such that students and researchers with basic bioinformatics experience may use this dataset to learn about microbial communities within the soil. We present code and explanations for common analysis steps, including basic quality control (QC), assembling reads into larger genome fragments ("contig" assembly), predicting genes, quantifying gene counts for specific ecological or biogeochemical functions, genome assembly, and exporting to the KBase platform (Arkin *et al.*, 2018). We recommend the review by Pérez-Cobas *et al.*, (2020) for software alternatives for each step of this shotgun metagenomics analysis.

Methods

Dataset description

Soil samples are collected annually from 47 NEON sites during peak greenness. Soil samples are collected up to 30cm below the soil surface, the organic (O) and the mineral (M) horizons (when present) are separated, and subsamples from each horizon are homogenized into one composite sample per horizon, and frozen on dry ice until DNA extraction. Sample file names include the 4-letter site identifiers, soil horizons (O or M), sampling date, and replicate number. Three samples are collected within a NEON plot at a sampling time point. As of 2021, DNA extractions are performed using KAPA Hyper Plus kit (Kapa Biosystems). Samples from multiple sites are pooled into sets of 40 or 60 for 150 bp paired-end sequencing, which is conducted on an Illumina NextSeq at the Battelle Memorial Institute (NEON Metagenomics Standard Operating Procedure, v.3). While there is currently no versioned release of NEON's metagenomic data, the pipeline described here is designed to be robust to processing new short-read sequence data as they are released from NEON, approximately annually, though protocols may shift over NEON's 30-year time span (Stanish & Parnell, 2018).

Operation

We assume a Linux operating system and command-line interface. Storage and RAM requirements will depend on the specific analyses performed and the number of samples analyzed. To work with a large dataset (10+ samples), a significant amount of computational power will be necessary, ideally with 8 or more cores for parallel computation. For those without access to institutional high-performance clusters, the scientific computing platform CyVerse (Merchant *et al.*, 2016) offers free computational and storage resources.

The computing requirements for metagenomic analysis can sometimes overwhelm personal computers, or login nodes on shared computing clusters. Therefore, users may wish to test the pipeline in a local environment, then shift to a high-performance cluster for large numbers of samples. Due to the long duration of certain steps, users may benefit from Linux commands that prevent sessions from timing-out or dropping the connection, such as *tmux* or *screen*. Either method requires modifying the configuration file called "config.yaml." **Bolded text will be used to emphasize parameters that should be modified within the configuration file.**

Local analysis: Each metaGEM command can be run with a "--local" flag to run within your current environment. If you have access to multiple cores, then you will need to add the "--cores" flag to each metaGEM commands below, to take advantage of parallel computing. This command can check your available threads, though you may not want to use all of them if you share computing resources:

```
echo "CPU threads: $(grep -c processor/proc/cpuinfo)"
```

Cluster analysis: To run on a cluster, the pipeline will assume that jobs are submitted via a SLURM-based scheduling system, controlled using the file called "cluster_config.json." Clusters with SGE/OGE-based scheduling may require workarounds. **The "cores" section of the configuration file should be modified to reflect the number of computing cores for each step.** Contact your system administrator for information on appropriate scratch directories, or for guidance on scheduling and configuration files.

On shared computing clusters, some softwares must be loaded as "modules" before they are used. For instance, to use Miniconda (necessary for every step of this pipeline), this command will work if there is a shared installation:

```
module load miniconda # may need to specify version
```

If there is no existing Miniconda installation, follow the instructions from Conda for a new installation. Subsequent code will assume that analysis is running locally within a Miniconda environment.

Implementation

Once sequences are downloaded, we use the pipeline metaGEM (Zorrilla *et al.*, 2021), which links a variety of bioinformatics tools and users can develop customized extensions for specific purposes. metaGEM, and its underlying Snakemake framework (Köster & Rahmann, 2012), are designed to address common problems with software versioning and updating, as well as efficient data re-analysis (i.e. running the minimal tasks necessary to generate updated output files). We describe installation and use instructions for metaGEM below. In addition to metaGEM default steps for cleaning and assembling the raw reads, we describe taxonomic classification or protein annotation for predicted genes using custom databases.

To customize or expand on the workflow below, it is helpful to know the basic logic of Snakemake, which is the underlying framework for the metaGEM pipeline. Snakemake relies on a series of rules, which specify input files, output files, and any necessary commands. When a rule is called, Snakemake works backwards from the output files to decide if any input files are missing or outdated, and tries to re-run rules as needed (Köster & Rahmann, 2012).

Setup: installing metaGEM pipeline

Full details on installation can be found in the metaGEM wiki. In short, run the following commands to create and setup a new analysis directory called metaGEM:

```
git clone https://github.com/franciscozorrilla/metaGEM.git # Download metaGEM repo
cd metaGEM # enter directory
bash env_setup.sh # Run automated setup script
```

Confirm success of installation and environment setup:

```
bash metaGEM.sh -t check
```

If all went well, your screen will report messages about the installation. Otherwise, it will report any problems in specific package installations or environments. You can inspect at the new environments using:

```
conda env list
```

Activate the metaGEM conda environment. This will be used for most parts of the pipeline.

```
conda activate metaGEM
```

Open the configuration file called "config.yaml" and modify paths as needed. **Users must specify the location for the analysis environment, as well as a "scratch" directory for temporary files.**

1. Accessing raw sequence files

1.1 Download test dataset

We recommend an initial interactive test of the pipeline with two microbial samples. This will ensure that all necessary software is installed and that file paths are correct. From within the metaGEM directory, a sample set can be downloaded using the code block below:

```
cd dataset # enter data directory (within metaGEM directory)
wget  https://neon-microbial-raw-seq-files.s3.data.neonscience.org/2017/WOOD_002-M-20140925-comp_R1.fastq.gz
wget  https://neon-microbial-raw-seq-files.s3.data.neonscience.org/2017/WOOD_002-M-20140925-comp_R2.fastq.gz
wget  https://neon-microbial-raw-seq-files.s3.data.neonscience.org/2017/SCBI_012-M-20140915-comp_R1.fastq.gz
wget  https://neon-microbial-raw-seq-files.s3.data.neonscience.org/2017/SCBI_012-M-20140915-comp_R2.fastq.gz
cd ..# return to enclosing metaGEM directory
```

Next, we have metaGEM reorganize the raw sequence files into subfolders.

```
bash metaGEM.sh --task organizeData
```

1.2 Download custom dataset

Information about the metagenomic sequencing for each soil sample is contained in the NEON data product DP1.10107.001, which can be accessed using the interactive Data Portal.

Data from specific sites and dates can also be accessed via the neonUtilities R package (Lunch *et al.*, 2021). The R commands below will download the DP1.10107.001 metadata for all samples collected from the Harvard Forest site in the year 2018. This metadata can then be used to download raw sequences.

```
# install neonUtilities - can skip if already installed
install.packages("neonUtilities")
# load neonUtilities
library (neonUtilities)
metadata <- loadByProduct (dpID = 'DP1.10107.001', site="HARV", startdate = "2018-01", enddate = "2018-12", package = 'expanded')
```

Downloads will come with three tables of interest:

- mms_metagenomeDnaExtraction: reports the quantity of DNA extracted from the soil sample.
- mms_metagenomeSequencing: lists sequencing protocol for each sample, as well as the read counts. These read counts can be used to filter out low-quality samples.
- mms_rawDataFiles: lists the download URL for each sample. This table is included only with the "expanded" package setting, not the "basic" setting.

The sites and dates of interest should be determined by the goals of your analysis: a comparative study might require samples from Alaska as well as from Puerto Rico, or samples could be retrieved from sites that have accompanying multi-decadal data from the Long-Term Ecological Research (LTER) program. If samples have the extension.tar.gz, then they are bundled into a compressed folder with multiple samples and will need to be unbundled (see tutorial here). Samples must have forward and reverse reads and they should be compressed in.fastq.gz format for most downstream software. Even when compressed, each file may still require multiple GB of storage.

2. Quality control

2.1 Background and rationale

Raw sequences are shared online in FASTQ format, with only minimal quality control from NEON's sequencing facilities, since users may prefer to use specific protocols for quality control. Some aspects of quality control present a trade-off between data volume and data quality. Each base returned by a sequencing machine (e.g. "A", "C", "T", or "G") has an associated quality score, or *Q score* (Cock *et al.*, 2009). Q scores can be used to filter low-quality reads, which generally improves the reliability of genomic analysis (Illumina, 2014). Certain aspects of quality control are absolutely necessary for reliable analysis, such as removing adapter or primer sequences used in sequencing protocols. For these steps, Cutadapt (Martin, 2010) and Trimmomatic (Bolger *et al.*, 2014) are frequently-used tools and work well. Fastp (Chen *et al.*, 2018) is an all-in-one QC tool included in the metaGEM pipeline (Section 2.3) (Zorrilla *et al.*, 2021).

Optional steps of quality-control include removing low-complexity sequences and searching for contaminants. Low-complexity sequences are naturally occurring regions of DNA with highly biased distributions of bases, such as "AAAAAAAAAGCGCTTTTTTT." These regions can make matching to gene databases more difficult by causing spurious results (Clarke *et al.*, 2019). Users may wish to search for and remove contaminant sequences, such as those that match the PhiX genome, which is a common contaminant of Illumina metagenomic data due to its use as a control during sequencing (Mukherjee *et al.*, 2015).

2.2 Considerations for NEON data

Soil samples from NEON have a wide range of average quality scores, as well as a range of sequencing depths, which are affected by DNA amounts in soil, lab DNA extraction efficiency, and sequencer error. We recommend removing samples with lower sequencing depths, but the specific depth cutoff will vary based on your analysis goals (Brumfield *et al.*, 2020). Up to 100 Gbp may be required for characterizing full soil diversity (van der Walt *et al.*, 2017). None of NEON's metagenomes meet this ultra-high sequencing depth, but the majority are sequenced to at least 1.5 Gbp (Figure 1a).

In a subset of NEON metagenomes, we did not find PhiX contamination, so this step is not implemented in Section 2.3. However, tools for removing low-complexity sequences (Komplexity) and removing contaminant DNA are included in the Sunbeam pipeline (Clarke *et al.*, 2019), an alternative to the metaGEM pipeline used throughout.

2.3 Implementation via metaGEM pipeline

To run quality control on raw sample files (primer trimming, adapter trimming, read filtering, and base quality evaluation) run the following command:

```
bash metaGEM.sh --task fastp --local
```

Each sample will have detailed report files within the "qfiltered" directory. To summarize the results across all samples, run the following command:

```
bash metaGEM.sh --task qfilterVis --local
```

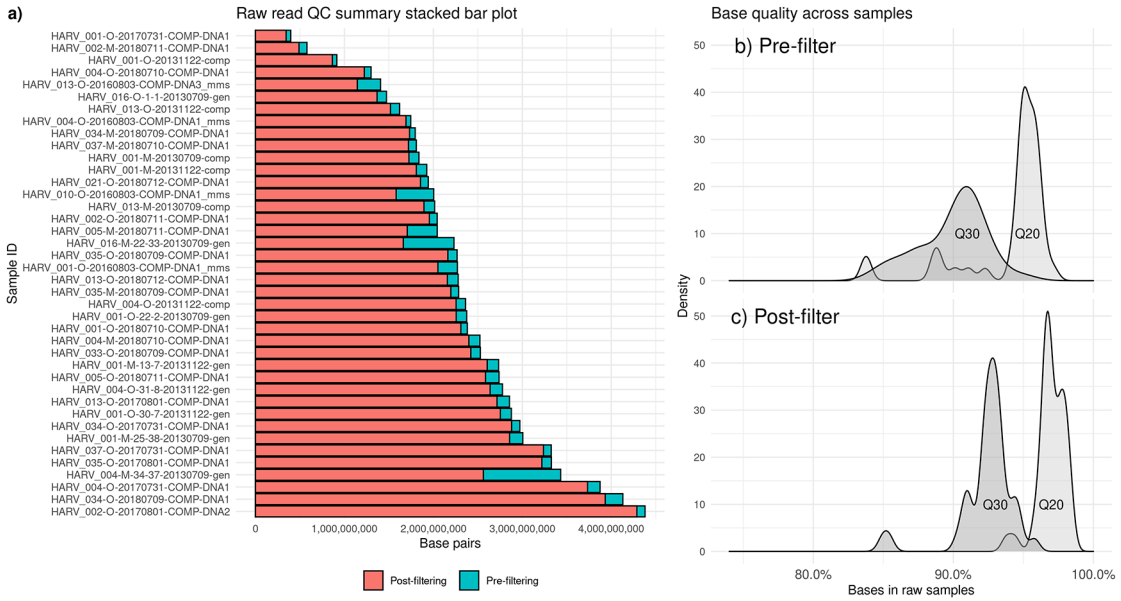

Figure 1. Quality control results for short reads using the Fastp software (Chen *et al.*, 2018). Short-read metagenomic samples are from the Harvard Forest site of the National Ecological Observatory Network (NEON). a) Counts of read pairs before (blue) and after (red) quality control steps. b) Base quality at Q30 (dark gray) and Q20 (light gray) before filtering. c) Base quality at Q30 (dark gray) and Q20 (light gray) after filtering.

Simple visualization of QC outputs will then be generated within the "stats" directory.

3. Assembly-free analysis

3.1 Background and rationale

Metagenomic analysis often involves assembling short reads into longer fragments, called contigs, which can be searched for genes. However, the assembly step is computationally intensive, and may be avoidable if the only desired output is a taxonomic profile, which can be generated by tools designed to work with unassembled short reads (Pearman *et al.*, 2020). These tools, such as Kraken2 (Wood *et al.*, 2019) or Kaiju (Menzel *et al.*, 2016), can assign taxonomic identities to reads by comparing sequences to reference databases. Compared to other classification tools, Kraken2 has been shown to perform favorably on soil datasets (Kalantar *et al.*, 2020; Lu & Salzberg 2020). However, the vast majority of soil reads remain unclassified with short-read classifiers. This may be due to the lack of complete genomes from soil organisms within reference databases (Quince *et al.*, 2017).

3.2 Considerations for NEON data

Taxonomic reference databases can include sequences from various biological domains, often using genomes from RefSeq (O'Leary *et al.*, 2016) or marker gene databases such as Silva (Quast *et al.*, 2013) and RDP (Cole *et al.*, 2014). The "Standard" pre-built database, shared by the Kraken2 developers, contains sequences from archaea, bacteria, viral, plasmid, human, and UniVec_Core. Due to the importance of fungi within soil ecosystems, we tested a larger database ("PlusPF") that also includes fungi and protozoa. Overall, approximately 17% of reads were identifiable to any kingdom, with fewer than 0.1% assigned to fungi. Given the increased memory costs of larger databases, and the low detection of fungi and protozoa, a smaller database (e.g. the Standard) is likely preferable for most microbial analyses. Other NEON microbial data products (such as amplicon sequences, qPCR, and PLFA) can provide domain-specific information on fungi, bacteria, and archaea.

3.3 Implementation

The Kraken2 reference databases that span multiple domains of life can reach 100 gigabytes, presenting a potential obstacle to running analyses on personal computers. The Toolchest R package (Cai & Lebovic, 2021) allows for remote Kraken2 analysis of samples from within the R environment. The example code below uses the "PlusPF" Kraken2 database, which includes sequences from archaea, bacteria, viral, plasmid, human, protozoa, fungi, and vector contaminants. Results for each sample are summarized in a "report" file, which sums the number of reads assigned to each taxon.

```
install.packages("toolchest")
library("toolchest")
toolchest::set_key("share.NjYyZDE2ZTUtNTU0Ny00OWQzLTlkNTktYjRmMTAzYmM4NWFh") # example
key with limited capacity - please download a new key from the Toolchest website
kraken2(read_one = "WOOD_002-M-20140925-comp_R1.fastq.gz",
        read_two = "WOOD_002-M-20140925-comp_R2.fastq.gz",
        output_path = "./kraken_output.txt")
```

Kraken2 report files can be visualized using the software Pavian (Breitwieser & Salzberg, 2020). Pavian can be run locally via R, or samples can be uploaded for analysis using the online application. Alternatively, output from Kraken2 can be converted to the BIOM file format for in-depth visualization using the metagenomics exploration software Phinch (Bik, 2014).

4. Contig assembly

4.1 Background and rationale

Assembling short reads into contigs can increase sensitivity and accuracy when predicting and annotating genes. Contig assembly generally requires more computational power and time than any other step within metagenomic analysis (Quince *et al.*, 2017).

Assembly of soil metagenomes is particularly difficult due to high amounts of biodiversity per sample and the absence of organisms in reference databases. Currently, the only assembly software designed for soils is Megahit (Li *et al.*, 2016), which is also one of the fastest tools for metagenome assembly. For some samples, this speed may come at the expense of sensitivity. metaSPAdes has been benchmarked with soil data and performs comparably, sometimes producing longer contigs, but requires additional memory and runtime (van der Walt *et al.*, 2017).

Co-assembly of reads, in which information is shared between samples, increases sensitivity to low-abundance reads (Sczyrba *et al.*, 2017), and can aid in recovering rare genomes (Albertsen *et al.*, 2013). However, co-assembly causes an exponential increase in assembly time and memory usage, possibly taking days or weeks to complete. Co-assembly can also increase the number of chimeric contigs for samples with high strain diversity (Ramos-Barbero *et al.*, 2019).

Other assembly decisions (such as minimum contig length) should depend on downstream analyses; for example, average prokaryotic genes are about 1000 bp (Xu *et al.*, 2006), so shorter contigs may not contain useful information on gene presence or absence. Some genome binning tools, such as metaBAT, will discard any contigs lower than 1500 bp. Very

Name	Number of raw reads	Classified reads	Bacterial reads	Viral reads	Fungal reads	Protozoan reads
HARV_002-O-20170801-COMP-DNA2	14,776,052	22.3%	21.8%	0.014%	0.0764%	0.0189%
HARV_004-O-20170731-COMP-DNA1	14,335,691	16.2%	15.7%	0.0147%	0.0645%	0.0194%
HARV_001-O-20160803-COMP-DNA1_mms	14,308,726	11.7%	10.7%	0.00465%	0.0549%	0.00641%
HARV_001-M-25-38-20130709-gen	14,146,431	13.6%	13.4%	0.00484%	0.0213%	0.004%
HARV_001-O-30-7-20131122-gen	13,635,178	13.5%	13.2%	0.00487%	0.0456%	0.00794%
HARV_034-O-20180709-COMP-DNA1	13,610,474	21.2%	20.7%	0.0144%	0.0682%	0.0178%
HARV_004-O-31-8-20131122-gen	13,064,107	13.7%	13.3%	0.00478%	0.0427%	0.0077%
HARV_001-M-13-7-20131122-gen	12,902,412	12.1%	11.9%	0.00402%	0.0392%	0.00516%
HARV_004-M-34-37-20130709-gen	12,689,315	11.7%	11.5%	0.00451%	0.0185%	0.00335%
HARV_035-O-20170801-COMP-DNA1	12,131,047	17.3%	16.9%	0.0139%	0.0663%	0.0168%
HARV_037-O-20170731-COMP-DNA1	11,918,825	18.3%	17.9%	0.0143%	0.0621%	0.0158%
HARV_010-O-20160803-COMP-DNA1_mms	11,318,403	13.1%	11.8%	0.00466%	0.027%	0.00461%
HARV_001-O-22-2-20130709-gen	11,176,185	15.1%	14.6%	0.005%	0.0955%	0.0112%
HARV_004-O-20131122-comp	11,174,535	17%	16.6%	0.00769%	0.0695%	0.0108%
HARV_013-O-20170801-COMP-DNA1	10,723,092	16.4%	15.9%	0.0171%	0.0745%	0.017%

Figure 2. Percentage of metagenomic short reads assigned to high-level taxonomic categories. Samples are from the Harvard Forest site of the National Ecological Observatory Network (NEON). Reads were assigned using the *PlusPF* database (release 5/17/21), which includes sequences from archaea, bacteria, viral, plasmid, human, UniVec_Core, protozoa & fungi. Image generated using the visualization software Pavian (Breitwieser & Salzberg, 2020).

low thresholds, such as 300 or 500 bp, will increase the percentage of raw reads that are represented in an assembly. Longer contigs generally represent higher confidence in longer regions of the genome, although misassemblies can occur and lead to long contigs (Sczyrba *et al.*, 2017). We recommend the tool metaQUAST to perform in-depth evaluation assembly, such as summaries of contig length distributions, detection of misassemblies and errors, or comparison with reference databases to estimate the abundance of unknown species (Mikheenko *et al.*, 2016). The review by Ayling *et al.* (2020) covers recent developments in short-read assembly approaches and reference-free assembly evaluation.

4.2 Considerations for NEON data

The variation in sequencing depth among NEON soil samples corresponds to high variation in assembly length (Figure 3A). Samples with deeper sequencing depths had, on average, longer contig lengths (Figure 3B). Most assemblies consisted of thousands of separate contigs (Figure 3C). Due to the effort required for assembly, it may be preferable to select a subset of high-quality samples for downstream analysis, rather than assembling all samples.

Co-assembly of samples may improve assemblies, but it is currently unclear how samples should be grouped for optimal results, since co-assembly can improve some aspects of an assembly while also introducing errors (Ramos-Barbero *et al.*, 2019). Some options include grouping samples by sampling plot, timepoint, soil horizons, or field site.

4.3 Implementation

For the contig assembly step, we recommend changing certain parameters in the configuration file. Under the "params" section, the assemblyPreset parameter is passed to the assembly software, Megahit. **The default value is "meta-sensitive", but the "meta-large" setting is optimized for complex soil datasets.**

To assemble contigs, run the following command, specifying the number of available cores:

```
bash metaGEM.sh --task megahit --local --cores 28
bash metaGEM.sh --task assemblyVis
```

Visualization of assembly outputs are also located within the "stats" subfolder.

5. Functional gene annotation

5.1 Background and rationale

To estimate the functional capabilities of a soil microbial community, gene annotation can be carried out using various gene reference databases. This annotation step can be performed on short reads (i.e. the output from the quality filtering steps), but this can lead to false positives due to short reads matching multiple ambiguous regions of reference genes (Quince *et al.*, 2017). More confident matches can often be obtained by searching for genes within assembled contigs. However, soils often have low assembly rates, in which only a small portion of reads end up as part of a contig (Vollmers *et al.*, 2017), which can skew functional profiles. The benefit of assembling before annotation can be diminished if fewer than 85% of reads map to contigs (Tamames *et al.*, 2019).

Functional gene annotation of unassembled reads is carried out for all NEON samples on MG-RAST at the time of their online publication, using a collection of functional gene databases such as eggNOG (Huerta-Cepas *et al.*, 2019), KEGG (Kanehisa *et al.*, 2017), and SwissProt (Boutet *et al.*, 2007). Gene annotation from multiple databases can dramatically increase the number of annotated genes, a trend that is especially pronounced for microbes (such as soil organisms) that are only distantly related to model organisms like *E. coli* (Griesemer *et al.*, 2018).

When annotating genes in assembled contigs, a preliminary step is to identify Open Reading Frames (ORFs) using software such as Prodigal (Hyatt *et al.*, 2010). Then, BLASTp (Altschul *et al.*, 1990) or DIAMOND2 (Buchfink *et al.*, 2021) can be used to search against protein gene databases. Gene presence does not necessarily mean that the genes are transcribed or active; however, due to the metabolically expensive nature of maintaining genomic pathways (Lynch, 2006), there is potentially meaningful correspondence between gene presence and functional potential (Pérez-Cobas *et al.*, 2020).

5.2 Considerations for NEON data

Soil metagenomes can be used to explore functions of biogeochemical, medical, or ecological interest. For example, the Comprehensive Antibiotic Resistance Database (CARD) (Alcock *et al.*, 2020) is a curated reference database of DNA sequences and proteins, designed to identify mutations and mechanisms of resistance to antibiotics, which can develop as a result of poor human stewardship (Brown & Wright 2016). However, antibiotic resistance can also be an ecological signifier of fungal-bacterial competition for nutrients (Bahram *et al.*, 2018). Another protein database with relevance

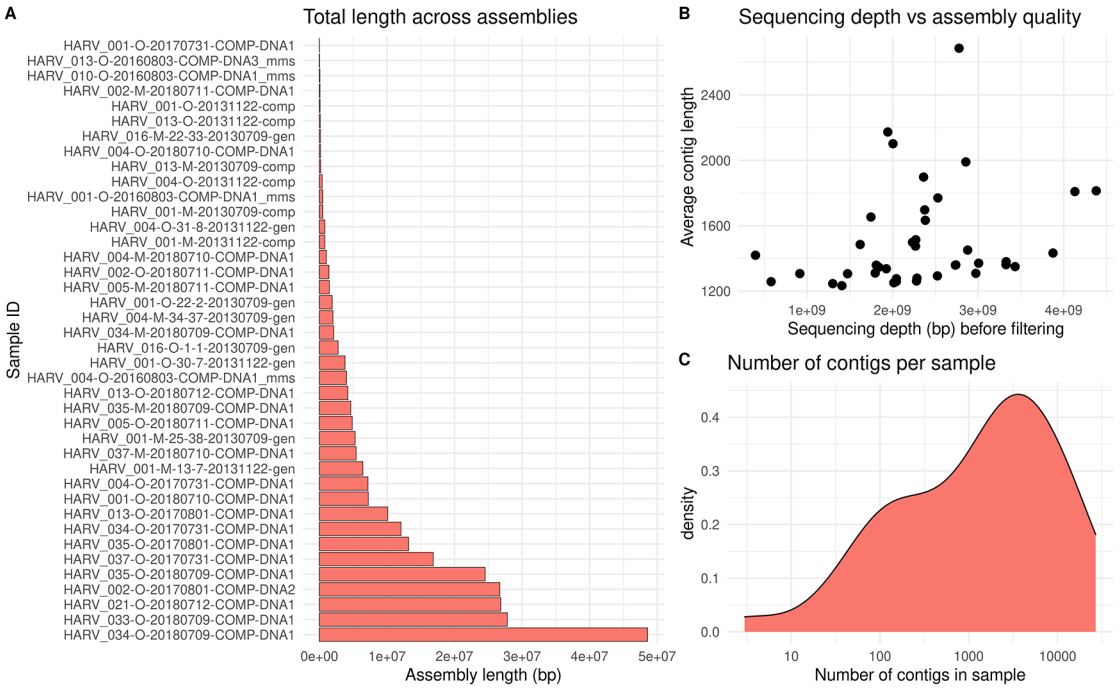

Figure 3. Results of contig assembly of short-read quality-filtered metagenomic samples. Contigs were assembled using the Megahit software, with samples from the Harvard Forest site of the National Ecological Observatory Network (NEON). The "meta-large" preset was used with a minimum contig length of 1000 base pairs (bp). a) Assembly length per sample, calculated as the sum of contig lengths within sample. b) Average contig length per sample, plotted against the sequencing depth before filtering. c) Density plot showing the number of contigs per sample.

to the soil microbiome is NCycDB, which categorizes genes into pathways that represent transformations such as nitrification, denitrification, and anammox. NCycDB was compiled from other sources, including COG, eggNOG, KEGG and the SEED (Tu *et al.*, 2019).

While functional gene profiling is more reliable with contigs rather than short reads (Anwar *et al.*, 2019), we note that only 5-10% of reads mapped to any contigs within select Harvard Forest samples (minimum contig length 1000, and pseudoalignment carried out using Kallisto with default settings (Bray *et al.*, 2016)). These low mapping rates may suggest that our assembled contigs represent only a small portion of the soil metagenome.

5.3 Implementation

For this example, we will search samples for genes from NCycDB. NCycDB has been shown to return fewer false positives when used with assembled contigs rather than unassembled short reads (Anwar *et al.*, 2019), so the following steps use the assembled contigs as input.

The NCycDB must be downloaded from Github and converted into a BLAST-compatible protein database. From the metaGEM directory, run the following commands to download the database:

```
svn export https://github.com/qichao1984/NCyc/trunk/data/NCyc_100_2019Jul.7z db/
NCyc_100_2019Jul.7z
```

This file must be decompressed from "7z" format into ".faa" format. Commands for this will vary based on your operating system.

Next, we use the program Diamond (Buchfink *et al.*, 2021) to convert to BLAST-compatible database for use within our pipeline:

```
diamond makedb --in db/NCyc_100_2019Jul.faa -d db/NCyc_DB
```

In your configuration file, the "blast_db" parameter should be modified to point to the database file name.

To predict the genes on the assembled contigs, run Prodigal via the following command:

```
bash metaGEM.sh --task run_prodigal
```

To compare the predicted genes with the NCycDB, run the following command:

```
bash metaGEM.sh --task run_blastp
```

To interpret the output files, each gene can be linked to its gene family using the "id2map" file associated with NCycDB:

```
svn export https://github.com/qichao1984/NCyc/trunk/data/id2gene.map.2019Jul db/
id2gene.map.2019Jul
```

To compare results across samples, gene counts must be normalized to account for variation in sequencing depths (Pereira *et al.*, 2018). One widely-used method is relative-log expression (RLE), which calculates scaling factors based on the geometric mean of gene abundances across all samples. RLE can be implemented using the DESeq R package (Love *et al.*, 2014), and can be used to identify genes that are differentially abundant between groups (such as field sites, or soil horizons).

6. Binning

6.1 Background and rationale

The vast majority of soil sequences match to no known organism (Figure 2). However, novel genomes can be assembled from metagenomes. These Metagenome-Assembled Genomes (MAGs) are more commonly assembled from human-associated samples, but they are quickly becoming a valuable resource for soil genomics: a recent collection of about 200 soil MAGs doubled the percentage of identifiable soil sequences, from 5% to 10% (Nayfach *et al.*, 2020). See Chen *et al.* (2020) for an overview of the strengths and pitfalls of MAG assembly and publication.

Because MAGs are assembled directly from contigs, rather than grown in an experimental setting, they often have no cultured relatives, representing a hidden source of genetic diversity in the microbiome (Nayfach *et al.*, 2020). For each putative genome, or "bin," summary statistics are produced that estimate the completeness and possible contamination of the genome, using a set of genes that are expected to be "single-copy" within a genome (Sieber *et al.*, 2018). Bins can be further refined manually, and genomes that are mostly complete with minimal contamination may be good candidates for submission to public databases (Bowers *et al.*, 2017). High-quality MAGs can uncover entirely new lineages in the microbial tree of life (Nayfach *et al.*, 2020).

Binning pipelines generally use a variety of separate binning tools, then refine and synthesize the best outputs from each tool. Bin refinement is essential for retrieving high-quality bins from soil than from other ecosystems, reflecting the challenges associated with soil bioinformatics (Sieber *et al.*, 2018; Uritskiy *et al.*, 2018).

6.2 Considerations for NEON data

Many of the genomes in reference databases such as RefSeq and Genbank are actually chimeric (consisting of multiple organisms). Chimeric genomes are especially prevalent in metagenome-assembled genomes, with chimerism identified in up to 30% of "high-quality" MAGs. Differential coverage data (obtained from multiple samples) can very quickly identify chimeric organisms. This makes the extensive NEON dataset particularly valuable for identifying novel soil genomes. Chimeric genomes can be identified by visualizing genomes in Anvi'o, or by running tools such as GUNC (Orakov *et al.*, 2021) that identify inconsistencies in the lineages of various genes.

6.3 Implementation

Genome binning is a well-supported feature of the KBase Predictive Biology platform, which was developed for microbiome analysis by the U.S. Department of Energy (Arkin *et al.*, 2018). KBase links hundreds of different software tools using an online interface, which allows users to create "Narratives" for specific data analysis projects. In an example Narrative (Figure 4), we combine the output from three tools, MaxBin2 (Wu *et al.*, 2016), MetaBAT2 (Kang *et al.*, 2019), and CONCOCT (Alneberg *et al.*, 2014). As inputs, we use the contigs assembled by MEGAHIT, as well as the quality-controlled sequencing reads. DAS Tool (Sieber *et al.*, 2018) and CheckM (Parks *et al.*, 2015) report on genome quality.

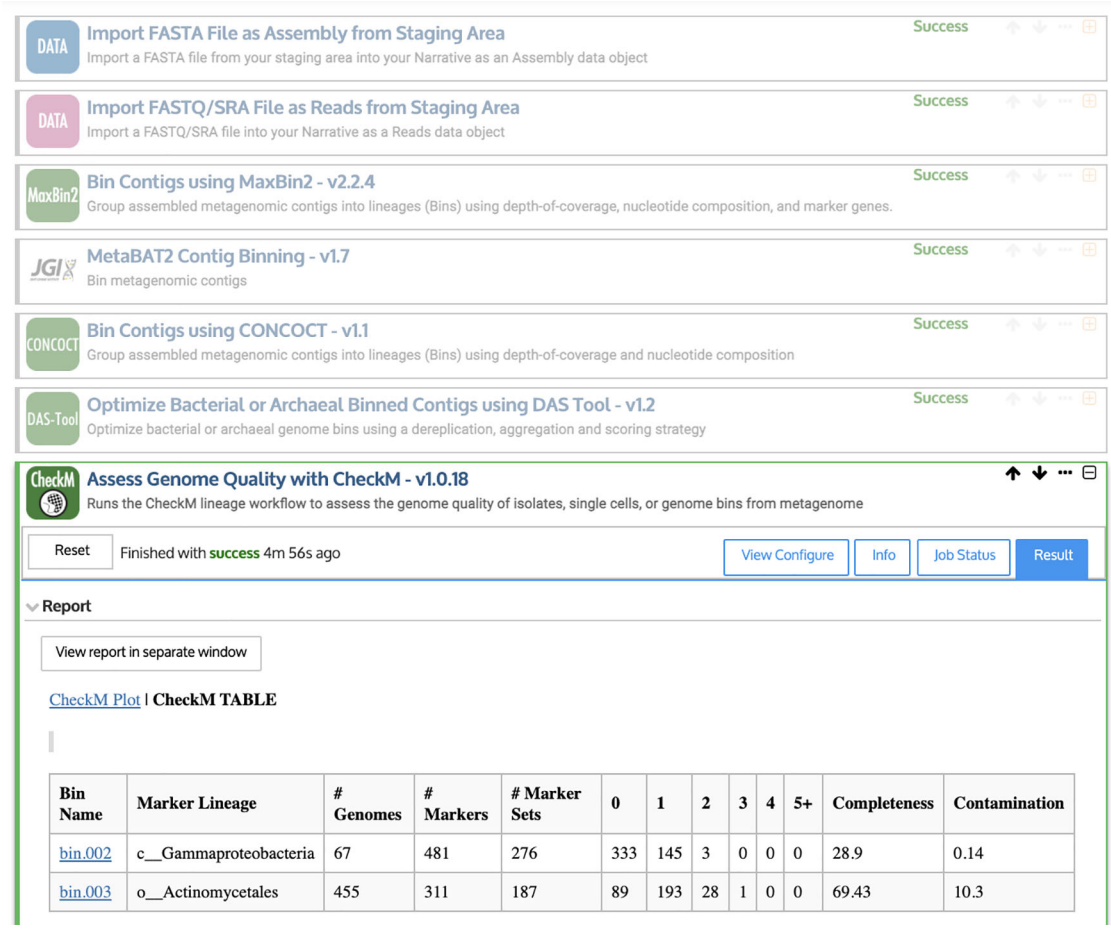

Figure 4. Example workflow for creating and evaluating Metagenome-Assembled Genomes (MAGs) using the KBase Narrative interface (Arkin *et al.*, 2018). First, quality-controlled sequencing reads and assembled contigs are imported using upload modules. Then, contigs are binned into putative genomes (or "bins") using MaxBin2 (Wu *et al.*, 2016), MetaBAT2 (Kang *et al.*, 2019), and CONCOCT (Alneberg *et al.*, 2014). DAS Tool (Sieber *et al.*, 2018) is used to identify the highest-quality bins. Finally, CheckM (Parks *et al.*, 2015) reports the completeness and contamination (among other statistics) for each putative genome.

However, there is currently a limited number of supported software tools within KBase, so the next section presents a Snakemake-based approach for carrying out similar tasks.

6.4 Genome binning

Assembled contigs can be grouped into bins using information such as read overlap and differential abundance across samples. The following metaGEM rule calculates differential abundance, and feeds this information into three binning tools: CONCOCT, metaBAT, and MaxBin:

```
bash metaGEM.sh --task binning --local --cores 28
```

6.5 Bin evaluation & refinement

To determine genome completeness, the metaGEM pipeline evaluates bins using a reference database called CheckM. The compressed database file can be downloaded as part of the env_setup.sh script (see Implementation section). Once the "checkM" folder is in your metaGEM directory, decompress it by running:

```
mkdir checkM
tar -xvzf checkm_data_2015_01_16.tar.gz -C checkM
checkm data setRoot checkM # may take a moment to complete
```

Next, the outputs from Concoct, metaBAT, and MaxBin are refined by metaWrap. **The default cutoffs for keeping a genome are 50% minimum completeness and 10% maximum contamination. These values can be modified within the configuration file.** To run the bin refinement step:

```
bash metaGEM.sh --task binEvaluation --local
```

To view the resulting bin quality for each sample, go to the sample name within the "reassembled_bins" directory and inspect the generated plots.

6.6 Genome taxonomy

The newly-assembled genomes can be evaluated against genome databases to determine taxonomy. First, users must set up the Genome Taxonomy Database (GTDB) (Parks *et al.*, 2020) and specify its location using the "GTDBTK_DATA_PATH" environment variable. For details on the download and installation of this database, see the GTDB-tk documentation (Chaumeil *et al.*, 2020).

Once the database is setup, run the following command for taxonomic assignment:

```
bash metaGEM.sh --task gtdbtk --local
```

6.7 Additional analysis

Additional analysis - such as metabolic modeling, and simulating interactions between MAGs - can be carried out with metaGEM, but has more complex software requirements. Details on implementation are in the metaGEM readme.

7. Applications

The NEON microbial sampling structure was designed to allow researchers to connect microbial community structure and functional potential (Stanish & Parnell, 2018). Complementary data streams can also be leveraged to link soil microbial data to ecosystem-level biogeochemical fluxes, plant growth, soil quality (Vestergaard *et al.*, 2017) and more. We recommend Qin *et al.* (2021) for a discussion of the high-level questions that may be tackled using NEON soil microbial data; below we highlight a few topics and recommended resources.

7.1 Microbial community structure

NEON microbial data is well-suited for elucidating basic patterns in soil microbial ecology, such as the variation between communities at different spatial and temporal scales (Qin *et al.*, 2021). The nested sampling, in which soil samples come from plots within each site, can be used to investigate spatial variability and autocorrelation among genes or taxa (Averill *et al.*, 2021). Longer-term change in microbial communities could be studied by integrating multi-decadal data from the Long-Term Ecological Research (LTER) program.

Shotgun metagenomes, which provide a snapshot of the entire genomic potential of a community, can be contrasted with amplicon sequencing, in which specific gene regions are amplified with the goal of distinguishing between taxa. NEON performs amplicon sequencing (NEON.DP1.10108.001) for soil fungi and bacteria, approximately 3 times per year at each site. These amplicon sequencing data can be accessed through the specialized neonMicrobe R package (Qin *et al.*, 2021). To link amplicon sequences with metagenome-assembled genomes (MAGs; Section 6), MAGs must include the gene regions used for amplicon sequencing. Tools such as phyloFlash (Gruber-Vodicka *et al.*, 2020) can be used to specifically assemble these gene regions and insert them into MAGs. This method provides an avenue for exploring the hidden diversity of the soil microbiome via genome assembly, while retaining the phylogenetic context of new genomes.

7.2 Biogeochemistry

The biogeochemical functions of soil microbes are poorly understood, despite their importance to global nutrient recycling. NEON measures many aspects of soil chemistry, which represents the nutrients available to microbial and plant communities. One-time characterizations of soil texture, bulk density, and detailed chemistry (including micro-nutrients such as zinc, iron, copper, etc.) are collected during the setup of each site (NEON.DP1.00096.001). Soil carbon and nitrogen are measured multiple times per year. (NEON.DP1.10086.001). Both datasets can be accessed using the neonUtilities R package or the NEON Data Portal. These can be used to investigate how microbial communities vary with chemical properties.

A subset of NEON metagenomes have an associated data stream on soil nitrogen transformations (NEON.DP1.10086.001), usually measured at each site once every five years. To calculate microbial rates of nitrogen mineralization and nitrification, soils are incubated for a month. Initial and final pools of ammonium, nitrites, and

nitrates can be converted into daily transformation rates using the neonNTrans R package (Weintraub, 2021). To link these nitrogen transformation rates to microbial data, users can estimate the abundances of pathway genes from NCycDB (Section 5.3), and match datasets with the dnaSampleID sample identifier. Genes that encode for enzymes like ammonia monooxygenase (AMO) are often used as proxies for nitrogen transformation activity, though the relationships between gene presence and functional activity are poorly characterized (Rocca *et al.*, 2015). NEON's soil nitrogen and microbial data can be used to clarify the strength of gene-function relationships across diverse biomes.

7.3 Plant communities

The soil microbiome is intimately linked with plant communities, which rely on (or compete with) soil microbes for nutrients (Bo *et al.*, 2022). NEON soil microbial data is collected alongside detailed inventories of plant species (DP1.10058.001), phenology (DP1.10055.001), tree biomass (DP1.10098.001), root biomass (DP1.10066.001), and root stable isotopes (DP1.10099.001). Summaries of plant diversity metrics at multiple spatial resolutions are available using the neonDiversity R package (Mahood, 2020). These data streams could be used to answer long-standing questions about spatio-temporal associations between plants and microbes (O'Brien *et al.*, 2021). For instance, soils form the "seed bank" from which plants recruit microbial symbionts (Bo *et al.*, 2022). The metabolic capacity of these symbionts can change the growth and stress tolerance of plants (Ravanbakhsh *et al.*, 2019). Soil metagenomes could be used to identify key microbial genes or symbionts affecting plant distributions across ecosystems (Cregger *et al.*, 2021).

7.4 Bioinformatics

Major challenges in soil bioinformatics include the lack of reference databases and specialized analysis tools, with different pipelines often leading to divergent conclusions (Pauvert *et al.*, 2019). NEON sequences can be used to develop bioinformatics pipelines that work well across biologically and physically heterogeneous soil biomes. Currently available pipelines that work well on some soils may perform poorly on other soils, because soil chemistry affects sequencing library preparation and can lead to downstream biases in sequence data. For instance, guanine-cytosine (GC) content of genomic regions can add bias to sample preparation steps, such as DNA lysing and sequencing (Benjamini & Speed, 2011). GC content is related, however, to temperature and nutrient conditions, and varies between species. While many bioinformatic tools attempt to correct for GC bias, these normalization steps may not be equally important for different soils. By freely providing sequences from a variety of biomes, researchers can calibrate tools against a reference dataset that reflects the full diversity of soils. More generally, NEON shotgun metagenomes can be used to investigate how variation in bioinformatic pipeline decisions affect ecological inferences. They may also act as a valuable resource for soil bioprospecting efforts, which use bioinformatic approaches to identify bioactive compounds with potential medical or industrial value (Vuong *et al.*, 2022).

Data availability

Raw metagenomics sequencing data is published in RELEASE-2021 as DP1.10107.001 from the National Ecological Observatory Network (https://data.neonscience.org/data-products/explore). All other data is previously published and cited throughout the paper.

Software availability

Bioconductor packages available at https://www.bioconductor.org/. CRAN packages available at https://cran.r-project.org/. metaGEM software is available at https://github.com/franciscozorrilla/metaGEM and the version used for this publication is archived at https://doi.org/10.5281/zenodo.4707723.

Author contributions

ZRW, BH, and JLN developed the software tools and tested the workflow. ZRW and BH wrote the initial manuscript draft; all authors contributed to revisions. JMB and MD provided project supervision and obtained funding.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation through the National Ecological Observatory Network, which is operated under cooperative agreement by Battelle Memorial Institute. We thank Michael Silverstein for assistance with parsing BLAST outputs in Python. We also thank the developers of metaGEM and Toolchest for assistance in troubleshooting code.

References

Albertsen M, Hugenholtz P, Skarshewski A, *et al.*: **Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.** *Nat Biotechnol.* 2013; **31**(6): 533–538.
PubMed Abstract | Publisher Full Text

Alcock BP, Raphenya AR, Lau TTY, *et al.*: **CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database.** *Nucleic Acids Res.* 2020.
Publisher Full Text

Allison SD, Lu Y, Weihe C, *et al.*: **Microbial abundance and composition influence litter decomposition response to environmental change.** *Ecology.* 2013; **94**(3): 714–725.
PubMed Abstract | Publisher Full Text

Alneberg J, Bjarnason BS, De Bruijn I, *et al.*: **Binning metagenomic contigs by coverage and composition.** *Nat Methods.* 2014; **11**(11): 1144–1146.
PubMed Abstract | Publisher Full Text

Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
Publisher Full Text

Anwar MZ, Lanzen A, Bang-Andreasen T, *et al.*: **To assemble or not to resemble-A validated Comparative Metatranscriptomics Workflow (CoMW).** *Gigascience.* 2019; **8**(8): 1–10.
Publisher Full Text

Arkin AP, Cottingham RW, Henry CS, *et al.*: **KBase: The United States department of energy systems biology knowledgebase.** *Nat Biotechnol.* 2018; **36**(7): 566–569.
PubMed Abstract | Publisher Full Text

Averill C, Werbin ZR, Atherton KF, *et al.*: **Soil microbiome predictability increases with spatial and taxonomic scale.** *Nat Ecol Evol [Internet].* 2021; **5**(6): 747–756.
PubMed Abstract | Publisher Full Text

Ayling M, Clark MD, Leggett RM: **New approaches for metagenome assembly with short reads.** *Brief Bioinform.* 2020; **21**(2): 584–594.
PubMed Abstract | Publisher Full Text

Bahram M, Hildebrand F, Forslund SK, *et al.*: **Structure and function of the global topsoil microbiome.** *Nature [Internet].* 2018; **560**(7717): 233–237.
PubMed Abstract | Publisher Full Text

Bai B, Liu W, Qiu X, *et al.*: **The root microbiome: Community assembly and its contributions to plant fitness.** *J Integr Plant Biol.* 2022.
PubMed Abstract | Publisher Full Text

Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic Acids Res.* 2011; **40**(10): 1–14.

Bik H: **Pitch Interactive Inc. Phinch: An interactive, exploratory data visualization framework for –Omic datasets.** *bioRxiv.* 2014:009944.

Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**: 2114–2120.
PubMed Abstract | Publisher Full Text

Boutet E, Lieberherr D, Tognolli M, *et al.*: **UniProtKB/Swiss-Prot: The manually annotated section of the UniProt KnowledgeBase.** *Methods Mol Biol.* 2007; **406**: 89–112.
PubMed Abstract

Bowers RM, Kyrpides NC, Stepanauskas R, *et al.*: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.** *Nat Biotechnol.* 2017; **35**(8): 725–731.
PubMed Abstract | Publisher Full Text

Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
Publisher Full Text

Breitwieser FP, Salzberg SL: **Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification.** *Bioinformatics.* 2020; **36**(4): 1303–1304.
PubMed Abstract | Publisher Full Text

Brown ED, Wright GD: **Antibacterial drug discovery in the resistance era.** *Nature.* 2016; **529**(7586): 336–343.
Publisher Full Text

Brumfield KD, Huq A, Colwell RR, *et al.*: **Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data.** *PLoS One.* 2020; **15**(2): 1–21.
Publisher Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat Methods [Internet].* 2021; **18**(4): 366–368.
Publisher Full Text

Cai B, Lebovic N. toolchest-client-r [Internet]. 2021.
Publisher Full Text

Chaumeil PA, Mussig AJ, Hugenholtz P, *et al.*: **GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database.** *Bioinformatics.* 2020; **36**(6): 1925–1927.

Chen LX, Anantharaman K, Shaiber A, *et al.*: **Accurate and complete genomes from metagenomes.** *Genome Res.* 2020; **30**(3): 315–333.
PubMed Abstract | Publisher Full Text

Chen S, Zhou Y, Chen Y, *et al.*: **Fastp: An ultra-fast all-in-one FASTQ preprocessor.** *Bioinformatics.* 2018; **34**(17): i884–i890.
PubMed Abstract | Publisher Full Text

Clarke EL, Taylor LJ, Zhao C, *et al.*: **Sunbeam: An extensible pipeline for analyzing metagenomic sequencing experiments.** *Microbiome.* 2019; **7**(1): 1–13.
Publisher Full Text

Cock PJA, Fields CJ, Goto N, *et al.*: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Res.* 2009; **38**(6): 1767–1771.
PubMed Abstract | Publisher Full Text

Cole JR, Wang Q, Fish JA, *et al.*: **Ribosomal Database Project: Data and tools for high throughput rRNA analysis.** *Nucleic Acids Res.* 2014; **42**: D633–D642.
PubMed Abstract | Publisher Full Text

Cregger MA, Carper DL, Christel S, *et al.*: **Plant-microbe interactions: From genes to ecosystems using populus as a model system.** *Phytobiomes J.* 2021; **5**(1): 29–38.
Publisher Full Text

Fierer N, Leff JW, Adams BJ, *et al.*: **Cross-biome metagenomic analyses of soil microbial communities and their functional attributes.** *Proc Natl Acad Sci [Internet].* 2012 Dec 26 [cited 2019 May 24]; **109**(52): 21390–21395.
Reference Source | PubMed Abstract | Publisher Full Text

Griesemer M, Kimbrel J, Zhou C, *et al.*: **Combining multiple functional annotation tools increases coverage of metabolic annotation.** *bioRxiv.* 2018: 1–11.

Gruber-Vodicka HR, Seah BKB, Pruesse E: **phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes.** *mSystems.* 2020; **5**(5).
PubMed Abstract | Publisher Full Text

Huerta-Cepas J, Szklarczyk D, Heller D, *et al.*: **EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.** *Nucleic Acids Res.* 2019; **47**(D1): D309–D314.
PubMed Abstract | Publisher Full Text

Hyatt D, Chen GL, LoCascio PF, *et al.*: **Prodigal: Prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics.* 2010; **11**.
PubMed Abstract | Publisher Full Text

Illumina.: **Understanding Illumina Quality Scores.** *Tech Note Informatics [Internet].* 2014 [cited 2020 Oct 14];1–2.
Reference Source

Jones M: **NEON Educational Resources for Online Teaching.** *NEON Obs Blog.* 2020.

Kalantar K, Carvalho T, de Bourcy C, Dimitrov B, Dingle G, Egger R, *et al.* **IDseq – An Open Source Cloud-based Pipeline and Analysis Service for Metagenomic Pathogen Detection and** *Monitoring.* 2020;(April): 1–14.

Kanehisa M, Furumichi M, Tanabe M, *et al.*: **KEGG: New perspectives on genomes, pathways, diseases and drugs.** *Nucleic Acids Res.* 2017; **45**(D1): D353–D361.
PubMed Abstract | Publisher Full Text

Kang DD, Li F, Kirton E, *et al.*: **MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.** *PeerJ.* 2019; **7**(7): e7359.
Publisher Full Text

Keller M, Schimel DS, Hargrove WW, *et al.*: **A continental strategy for the National Ecological Observatory Network.** *Front Ecol Environ.* 2008; **6**(5): 282–284.
Publisher Full Text

Köster J, Rahmann S: **Snakemake-a scalable bioinformatics workflow engine.** *Bioinformatics.* 2012; **28**(19): 2520–2522.
Publisher Full Text

Ladoukakis E, Kolisis FN, Chatziioannou AA. **Integrative workflows for metagenomic analysis. Front Cell** *Dev Biol.* 2014; **2**(NOV): 1–11.
Publisher Full Text, PMID: | PubMed Abstract

Li D, Luo R, Liu CM, *et al.*: **MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices.** *Methods.* 2016; **102**: 3–11.
PubMed Abstract | Publisher Full Text

Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 1–21.
Publisher Full Text

Lu J, Salzberg SL: **Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2.** *Microbiome.* 2020; **8**(1): 1–11.
Publisher Full Text

Lunch C, Laney C, Mietkiewicz N, *et al.*: **neonUtilities: Utilities for Working with NEON Data. R package version 2.1.1.** 2021.
Reference Source

Lynch M: **Streamlining and simplification of microbial genome architecture.** *Annu Rev Microbiol.* 2006; **60**: 327–349.
PubMed Abstract | Publisher Full Text

Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet.* 2010; **17**.
Publisher Full Text

Menzel P, Ng KL, Krogh A: **Fast and sensitive taxonomic classification for metagenomics with Kaiju.** *Nat Commun.* 2016; **7**.
PubMed Abstract | Publisher Full Text

Merchant N, Lyons E, Goff S, *et al.*: **The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences.** *PLoS Biol.* 2016; **14**(1): 1–9.
Publisher Full Text

Méric G, Wick RR, Watts SC, *et al.*: **Correcting index databases improves metagenomic studies.** *bioRxiv.* 2019.

Mikheenko A, Saveliev V, Gurevich A: **MetaQUAST: Evaluation of metagenome assemblies.** *Bioinformatics.* 2016; **32**(7): 1088–1090.
Publisher Full Text

Mukherjee S, Huntemann M, Ivanova N, *et al.*: **Large-scale contamination of microbial isolate genomes by illumina Phix control.** *Stand Genomic Sci.* 2015; **10**(APRIL2015): 1–4.
Publisher Full Text

Nasko DJ, Koren S, Phillippy AM, *et al.*: **RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification.** *Genome Biol [Internet].* 2018 Dec 30;19(1):165. ; **19**: 165.
PubMed Abstract | Publisher Full Text

Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, *et al.*: **A genomic catalog of Earth's microbiomes.** *Nat Biotechnol [Internet].* 2020 Nov 9.
Reference Source

O'Brien AM, Ginnan NA, Rebolleda-Gómez M, *et al.*: **Microbial effects on plant phenology and fitness.** *Am J Bot.* 2021; **108**(10): 1824–1837.
PubMed Abstract | Publisher Full Text

O'Leary NA, Wright MW, Brister JR, *et al.*: **Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res.* 2016; **44**: D733–D745.
PubMed Abstract | Publisher Full Text

Orakov A, Fullam A, Coelho LP, *et al.*: **GUNC: detection of chimerism and contamination in prokaryotic genomes.** *Genome Biol.* 2021; **22**(1): 1–19.
Publisher Full Text

Parks DH, Chuvochina M, Chaumeil PA, *et al.*: **A complete domain-to-species taxonomy for Bacteria and Archaea.** *Nat Biotechnol [Internet].* 2020; **38**(9): 1079–1086.
PubMed Abstract | Publisher Full Text

Parks DH, Imelfort M, Skennerton CT, *et al.*: **CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.** *Genome Res.* 2015; **25**(7): 1043–1055.
PubMed Abstract | Publisher Full Text

Pauvert C, Buée M, Laval V, *et al.*: **Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline.** *Fungal Ecol.* 2019; **41**: 23–33.
Publisher Full Text

Pearman WS, Freed NE, Silander OK: **Testing the advantages and disadvantages of short- And long-read eukaryotic metagenomics using simulated reads.** *BMC Bioinformatics.* 2020; **21**(1): 1–15.
Publisher Full Text

Pereira MB, Wallroth M, Jonsson V, *et al.*: **Comparison of normalization methods for the analysis of metagenomic gene abundance data.** *BMC Genomics.* 2018; **19**(1): 1–17.
Publisher Full Text

Pérez-Cobas AE, Gomez-Valero L, Buchrieser C: **Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses.** *Microb genomics.* 2020; **6**(8).
PubMed Abstract | Publisher Full Text

Qin C, Bartelme R, Chung YA, *et al.* **sequences to microbial ecology: Wrangling NEON soil microbe data with the neonMicrobe R package.** *Ecosphere [Internet].* 2021 Nov 24; **12**(11).
Publisher Full Text

Quast C, Pruesse E, Yilmaz P, *et al.*: **The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools.** *Nucleic Acids Res.* 2013; **41**(D1): D590–D596.
Publisher Full Text

Quince C, Walker AW, Simpson JT, *et al.*: **Shotgun metagenomics, from sampling to analysis.** *Nat Biotechnol.* 2017; **35**(9): 833–844.
Publisher Full Text

Ramos-Barbero MD, Martin-Cuadrado AB, Viver T, *et al.*: **Recovering microbial genomes from metagenomes in hypersaline environments: The Good, the Bad and the Ugly.** *Syst Appl Microbiol [Internet].* 2019; **42**(1): 30–40.
Publisher Full Text

Ravanbakhsh M, Kowalchuk GA, Jousset A: **Root-associated microorganisms reprogram plant life history along the growth–stress resistance tradeoff.** *ISME J [Internet].* 2019; **13**(12): 3093–3101.
Publisher Full Text

Rocca JD, Hall EK, Lennon JT, *et al.*: **Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed.** *ISME J.* 2015; **9**(8): 1693–1699.
PubMed Abstract | Publisher Full Text

Sczyrba A, Hofmann P, Belmann P, *et al.*: **Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software.** *Nat Methods.* 2017; **14**(11): 1063–1071.
PubMed Abstract | Publisher Full Text

Sieber CMK, Probst AJ, Sharrar A, *et al.*: **Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy.** *Nat Microbiol [Internet].* 2018; **3**(7): 836–843.
PubMed Abstract | Publisher Full Text

Stanish LF, Parnell J: **NEON.DOC.000908: TOS Science Design for Terrestrial Microbial Diversity.** *NEON Doc Libr [Internet].* 2018.
Reference Source

Tamames J, Cobo-Simón M, Puente-Sánchez F: **Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes.** *bioRxiv.* 2019: 1–16.

Tu Q, Lin L, Cheng L, *et al.*: **NCycDB: A curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes.** *Bioinformatics.* 2019; **35**(6): 1040–1048.
PubMed Abstract | Publisher Full Text

Uritskiy GV, DiRuggiero J, Taylor J: **MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis.** *Microbiome.* 2018; **6**(1): 158–113.
PubMed Abstract | Publisher Full Text

van der Walt AJ, van Goethem MW, Ramond JB, *et al.*: **Assembling metagenomes, one community at a time.** *BMC Genomics.* 2017; **18**(1): 521–513.
PubMed Abstract | Publisher Full Text

Vestergaard G, Schulz S, Schöler A, *et al.*: **Making big data smart—how to use metagenomics to understand soil quality.** *Biol Fertil Soils.* 2017; **53**(5): 479–484.
Publisher Full Text

Vollmers J, Wiegand S, Kaster AK: **Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters!** *PLoS ONE.* 2017; **12**: 1–31.
PubMed Abstract | Publisher Full Text

Vuong P, Wise MJ, Whiteley AS, *et al.*: **Small investments with big returns: environmental genomic bioprospecting of microbial life.** *Crit Rev Microbiol [Internet].* 2022; (Jan 31): 1–15.
PubMed Abstract | Publisher Full Text

Weintraub SR. neonNTrans R package: NEON Nitrogen Transformations. 2021.
Reference Source

Wood DE, Lu J, Langmead B: **Improved metagenomic analysis with Kraken 2.** *Genome Biol. [Internet].* 2019 Dec 28; **20**(1): 257.
PubMed Abstract | Publisher Full Text

Wu YW, Simmons BA, Singer SW: **MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets.** *Bioinformatics.* 2016; **32**(4): 605–607.
PubMed Abstract | Publisher Full Text

Xu L, Chen H, Hu X, *et al.*: **Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms.** *Mol Biol Evol [Internet].* 2006 Jun 1 [cited 2020 Oct 13]; **23**(6): 1107–1108.
Reference Source | PubMed Abstract | Publisher Full Text

Zorrilla F, Buric F, Patil KR, *et al.*: **metaGEM: reconstruction of genome scale metabolic models directly from metagenomes.** *Nucleic Acids Res.* 2021; 1–12.

Open Peer Review

Current Peer Review Status: ? ? ✓

Version 2

Reviewer Report 08 June 2022

https://doi.org/10.5256/f1000research.79544.r139218

✓ **Olubukola Oluranti Babalola**
Food Security and Safety Focus Area, Faculty of Natural and Agricultural Science, North-West University, Mmabatho, South Africa

The manuscript presents analysis and workflows of the NEON metagenomics data collected annually and how the datasets can be interrogated. The manuscript reported on the software that offers users, who are not yet confident enough, to build their pipeline from the start to use the software to analyze metagenomics, especially shotgun datasets. The software information has been improved to give room for the reproducibility of data. Each step involved in implementing the software was adequately described to ensure replication of the output.

A little addition would have been to present a user-friendly interface for beginners, who may not be familiar with or confident in using command lines, just as the part of KBase that was part of the workflow. Future studies can look into that.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant/Soil Microbe Interactions

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 19 July 2021

https://doi.org/10.5256/f1000research.54670.r84561

?

William Nelson
Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Rationale:
My main question is who is the audience for this pipeline? Is this intended to be used by students to learn some metagenomic analysis and how the NEON data set can be interrogated? Or is this intended to be used by researchers, in which case I think the downstream annotation and analysis components are somewhat thin. Is this officially recognized by NEON as a standard pipeline that will enable comparison between analyses? I don't wish to sound dismissive, but this reads like a Yet-Another-Metagenomics-Pipeline paper, which on one hand is fine - there's nothing technically or scientifically wrong with it - but this would be a more impactful report if the purpose behind it was more strongly presented.

Description:
There is nothing wrong with the description of the various steps, but the descriptions are superficial. There is little discussion of why the methods were chosen and what their strengths and weaknesses are.

Replication:
The code blocks are great, but the formatting rendered incorrectly in my browser (Firefox) - newlines were not present, making it hard to interpret what the actual commands are. Also, I tried to follow along with those commands on our institutional computing cluster and got stuck on the installation of sunbeam. I was able to install sunbeam on my desktop server, but the test of the install failed. I went ahead and tried to follow the analysis anyway, but ran into multiple problems. Just a caveat that providing the commands doesn't ensure replicability.

A few other comments:

End of Dataset description: "TOS Science Design for Terrestrial Microbial Diversity, NEON.DOC.000908" - What is this?

The comment about miniconda, "this command may work", is likely to be confusing. Might be best just to say that anaconda is required and to talk to local IT about its availability and how to use it.

The transition between section 1.2 and 2 should make it clearer that section 1.2 was describing constructing the configuration file and sections 2 through 5 are describing the individual steps that make up the sunbeam pipeline. As it reads now, it could be interpreted that the QC step is subsequent to the sunbeam run.

Is section 4.1b missing a code block?

I did not understand what you meant by "We use the homolog protein genes to construct our reference database." in section 5.

The Bowers 2017 reference appears to be missing from the bibliography.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I have 20 years experience performing microbial genomic and metagenomic analysis, including assembly, binning and annotation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Nov 2021

Zoey Werbin, Boston University, Boston, USA

- *My main question is who is the audience for this pipeline? Is this intended to be used by students to learn some metagenomic analysis and how the NEON data set can be interrogated? Or is this intended to be used by researchers, in which case I think the downstream annotation and analysis components are somewhat thin. Is this officially recognized by NEON as a standard pipeline that will enable comparison between analyses? I don't wish to sound dismissive, but this reads like a Yet-Another-Metagenomics-Pipeline paper, which on one hand is fine - there's nothing technically or scientifically wrong with it - but this would be a more impactful report if the purpose behind it was more strongly presented.*

Thank you for identifying these deficiencies within the manuscript. Our intended audience is both students and researchers working with NEON soil metagenomes. We have stated this explicitly in the last paragraph of the Introduction to the article, and strengthened each section of the paper to increase its value to these groups. Specifically, we have added subsections titled "Background and Rationale" and "Considerations for NEON data" to each analysis section. We plan to submit this revised manuscript for inclusion as a NEON community resource.

- *There is nothing wrong with the description of the various steps, but the descriptions are superficial. There is little discussion of why the methods were chosen and what their strengths and weaknesses are.*

Each step has now been supplemented with descriptions of our preferred methods as well as the strengths and weaknesses of alternative methods (in "Background and Rationale"). We describe which methods have or have not been benchmarked or optimized for soil metagenomes, specifically, as well as their usefulness for the NEON dataset, given the properties of the data (in "Considerations for NEON data").

- *The code blocks are great, but the formatting rendered incorrectly in my browser (Firefox) - newlines were not present, making it hard to interpret what the actual commands are. Also, I tried to follow along with those commands on our institutional computing cluster and got stuck on the installation of sunbeam. I was able to install sunbeam on my desktop server, but the test of the install failed. I went ahead and tried to follow the analysis anyway, but ran into multiple problems. Just a caveat that providing the commands doesn't ensure replicability.*

Great points. In response to this and to the comments of Reviewer #1, we have adjusted our specific bioinformatic methods to address Sunbeam installation issues. We now recommend the stable branch of the metaGEM pipeline, which has run successfully in multiple Linux environments. The code blocks have all been shortened to improve readability and cross-browser formatting.

- *End of Dataset description: "TOS Science Design for Terrestrial Microbial Diversity, NEON.DOC.000908" - What is this?*

The citation for this sampling protocol document has been changed to "Stanish & Parnell, 2018", with the full protocol version information within the Works Cited.

- *The comment about miniconda, "this command may work", is likely to be confusing. Might be best just to say that anaconda is required and to talk to local IT about its availability and how to use it.*

The sentence on miniconda requirements has been revised to point readers to their system administrators.

- *The transition between section 1.2 and 2 should make it clearer that section 1.2 was describing constructing the configuration file and sections 2 through 5 are describing the individual steps that make up the sunbeam pipeline. As it reads now, it could be interpreted that the QC step is subsequent to the sunbeam run.*

This recommendation is no longer relevant, given our shift in methods and manuscript organization.

- *Is section 4.1b missing a code block?*

This section is no longer present, given our shift in methods and manuscript organization.

- *I did not understand what you meant by "We use the homolog protein genes to construct our reference database." in section 5.*

This section is no longer present, given our shift in methods and manuscript organization.

- *The Bowers 2017 reference appears to be missing from the bibliography.*

This reference has been added to the bibliography.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 June 2021

https://doi.org/10.5256/f1000research.54670.r83581

?

Naupaka Zimmerman
Department of Biology, University of San Francisco, San Francisco, CA, USA

This is a timely and valuable contribution that has the potential to aid in the use of NEON data by a wider audience. The core approach (using Sunbeam, a snakemake pipeline, to analyze NEON metagenomics data) seems like a good one, and will offer advantages to users who are not yet comfortable enough to develop their own such pipeline from scratch.

While in general the approach is a good one and the need for the tool is real and well-articulated by the authors, there are a number of aspects that could be improved to maximize the value of this contribution. I will outline a few here, but I was unable to complete the full pipeline in my testing using the example data specified in the manuscript, and so I am not able to comment on all aspects of the pipeline at this time. I would be happy to do another view and assessment after hearing from the authors.

I outline some suggestions below:

In the last paragraph of the introduction, I would encourage the authors to revise this sentence: "The pipeline that we present here is designed to complement existing NEON educational resources, such that users without prior bioinformatics experience may use this dataset to learn about microbial communities within the soil." The background skills that are necessary to

successfully understand and implement the approach outlined here is not trivial and I don't think it's exactly best suited for someone "without prior bioinformatics experience". I think such a user would more likely need a graphical interface that did not presume comfort with the *nix command line etc. I think the approach outlined here is a valuable contribution because it targets users who may have some comfort with programmatic and command-line approaches, but does not yet have the skill to develop a flexible pipeline themselves.

In the methods section, first paragraph, I think I would revise to be more careful with tenses. In some cases the collection protocols will remain mostly unchanged (e.g. I don't think NEON is planning to add any core sites), but other things may change (the kits that they use, the sequencing depth or sequencer used, etc. Since NEON is a 30 year project, it might help the manuscript's longevity if this paragraph were worded to reflect possible future methodological changes.

I might encourage a mention or a suggestion that users use tmux or screen to run pipelines like this is they are connected to a remote server over something like ssh. If the connection drops during a many hours long pipeline, it can be quite frustrating.

In step 1.2, why do you suggest the use of the develop branch of Sunbeam? Isn't that more likely to include breaking changes that will be overly challenging for the target audience? Perhaps this could be adjusted to use a stable branch or version, and the text could highlight the develop branch alternative for those willing to trade troubleshooting time in exchange for quicker access to more advanced features.

For downloading the config file, it might be better to pull from an archival version of the file instead of the github version, or at the least include a version at a specific commit and not just the main branch, so that it remains stable. Otherwise either the code could break, or the authors would need to continually update the configuration to track with software changes.

In my testing of the approach in the manuscript, I am unable to get past the tests that occur after the installation of Sunbeam (`bash tests/run_tests.bash`). The tests repeatedly fail with segmentation faults during either the megahit or kraken steps. This is on an Ubuntu 20.04 machine with lots of RAM/disk space/cores. I am not sure where the issue is, and I would consider myself reasonably able to troubleshoot such problems, so I am concerned that similar problems might arise and be too challenging for the target audience/user. I would be happy to work with the authors in more detail to resolve this problem (share log files, etc). I shall share them via a comment when I am able to.

Overall, I think this is a valuable contribution that fills a need in the community and uses a good approach to do so. However, in its current form, I cannot successfully run the example code, even on the recommended sample files, and so I have concerns with the brittleness of the approach outlined. I'd encourage the authors to do some additional testing on other machines and settings, and/or build some more resilience into the installation walkthrough so that the average target user is able to make use of this contribution.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Environmental microbial ecology, including specific experience in bioinformatics and pipelines, and several years of experience working with large NEON sequencing datasets.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Nov 2021

Zoey Werbin, Boston University, Boston, USA

Original reviewer comments are italicized.

- *This is a timely and valuable contribution that has the potential to aid in the use of NEON data by a wider audience. The core approach (using Sunbeam, a snakemake pipeline, to analyze NEON metagenomics data) seems like a good one, and will offer advantages to users who are not yet comfortable enough to develop their own such pipeline from scratch. While in general the approach is a good one and the need for the tool is real and well-articulated by the authors, there are a number of aspects that could be improved to maximize the value of this contribution. I will outline a few here, but I was unable to complete the full pipeline in my testing using the example data specified in the manuscript, and so I am not able to comment on all aspects of the pipeline at this time. I would be happy to do another view and assessment after hearing from the authors.*

  Thank you for highlighting the issues with the reproducibility of the pipeline we outlined. Due to the referenced issues with installing software, we have switched to a similar Snakemake pipeline (metaGEM) that has been tested on various computing systems. We describe this new pipeline in the "Implementation" section of the revised manuscript.

- *In the last paragraph of the introduction, I would encourage the authors to revise this sentence: "The pipeline that we present here is designed to complement existing NEON educational resources, such that users without prior bioinformatics experience may use this dataset to learn about microbial communities within the soil." The background skills that are necessary to successfully understand and implement the approach outlined here is not trivial and I don't think it's exactly best suited for someone "without prior bioinformatics experience". I think such a user would more likely need a graphical interface that did not presume comfort with the *nix command line etc. I think the approach outlined here is a valuable contribution because it targets users who may have some comfort with programmatic and command-line approaches, but does not yet have the skill to develop a flexible pipeline themselves.*

  This sentence has been revised to reflect that our audience is those with basic bioinformatics experience. Further, each section of the manuscript has been expanded to include a thorough description of the rationale for various decisions in the subsections "Background and Rationale" and "Considerations for NEON data", so that this can be a more useful introductory guide to soil metagenomics.

- *In the methods section, first paragraph, I think I would revise to be more careful with tenses. In some cases the collection protocols will remain mostly unchanged (e.g. I don't think NEON is planning to add any core sites), but other things may change (the kits that they use, the sequencing depth or sequencer used, etc. Since NEON is a 30 year project, it might help the manuscript's longevity if this paragraph were worded to reflect possible future methodological changes.*

  Tenses in the "Dataset description" section have been modified to reflect that the reported sampling and sequencing protocols are accurate as of 2021. We state that this bioinformatics protocol is intended for short-read data specifically, and that NEON protocols may shift in the future.

- *I might encourage a mention or a suggestion that users use tmux or screen to run pipelines like this is they are connected to a remote server over something like ssh. If the connection drops during a many hours long pipeline, it can be quite frustrating.*

  We now reference tmux and screen in Implementation, within the sub-section "Local vs cluster analysis".

- *In step 1.2, why do you suggest the use of the develop branch of Sunbeam? Isn't that more likely to include breaking changes that will be overly challenging for the target audience? Perhaps this could be adjusted to use a stable branch or version, and the text could highlight the develop branch alternative for those willing to trade troubleshooting time in exchange for quicker access to more advanced features.*

  Due to our shift in methods, we no longer use either the develop or stable branch of Sunbeam. At the time of writing, however, the develop branch had implemented a potential fix for the segmentation fault errors, but it did not resolve errors on all operating systems. We hope the local and cluster options for running the metaGEM

pipeline will also help with reducing troubleshooting time.

- *For downloading the config file, it might be better to pull from an archival version of the file instead of the github version, or at the least include a version at a specific commit and not just the main branch, so that it remains stable. Otherwise either the code could break, or the authors would need to continually update the configuration to track with software changes.*

  With our shift from Sunbeam to metaGEM, we decided to remove the example configuration file. The configuration file that comes installed with metaGEM primarily needs file paths to be modified by the user, whereas most parameters can be left as-is. Throughout the text, we've bolded sentences that instruct the user to modify the configuration filepaths.
- *In my testing of the approach in the manuscript, I am unable to get past the tests that occur after the installation of Sunbeam (`bash tests/run_tests.bash`). The tests repeatedly fail with segmentation faults during either the megahit or kraken steps. This is on an Ubuntu 20.04 machine with lots of RAM/disk space/cores. I am not sure where the issue is, and I would consider myself reasonably able to troubleshoot such problems, so I am concerned that similar problems might arise and be too challenging for the target audience/user. I would be happy to work with the authors in more detail to resolve this problem (share log files, etc). I shall share them via a comment when I am able to.*

Overall, I think this is a valuable contribution that fills a need in the community and uses a good approach to do so. However, in its current form, I cannot successfully run the example code, even on the recommended sample files, and so I have concerns with the brittleness of the approach outlined. I'd encourage the authors to do some additional testing on other machines and settings, and/or build some more resilience into the installation walkthrough so that the average target user is able to make use of this contribution.

These are excellent points and led to a dramatic shift in the focus and implementation of this analysis pipeline. The main text of the manuscript now focuses on the various options available to users for each step of soil metagenomic analysis, and describes issues specific to soil ecology and the NEON dataset specifically. The code at the end of each section is now an example of how these decisions may be implemented via specific tools. For this revision, we have communicated with the developers of the tools mentioned (metaGEM and Toolchest) and are confident that these tools will maintain resilience in the coming years. We hope this sufficiently addresses problems of brittleness.

Competing Interests: No competing interests were disclosed.