Next-generation sequencing and data analysis:

Presented here are the steps for library preparation, in-lab sequencing and data analysis. Alternatively, samples may be sent for sequencing to a third-party service.

1. Sequencing Library Preparation

1.1. Quantify extracted DNA concentrations and quality.

NOTE: We use Qubit with the high sensitivity reagent for high accuracy (Table of Materials).

1.2. Amplify the V4 region of the 16S rRNA gene using the following tailed PCR primers:

515F:5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGCCAGCMGCCGCGGTAA-3'
806R:5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCTAAT-3'

NOTE: Tails provide compatibility with index and sequencing adapters.

Mixed according to the following scheme:

DNA sample	5µL
5X Phusion HF Buffer	4µL
2mM dNTPs	2µL
10µM 16S V4 515F Primer	1µL
10µM 16S V4 806R Primer	1µL
Phusion High-Fidelity DNA Polymerase	.2µL
Water	to 20µL

And with the following cycling parameters:

98 °C	30 s	
98 °C	**10 s**	
55-59 °C	**20s**	**35 Cycles**
72 °C	**30 s**	
72 °C	10 min	
4 °C	hold	

NOTE: Reactions using DNA extracted from compost samples can be efficiently amplified with 25 cycles. Worm samples have lower DNA concentrations and thus require additional amplification cycles. We therefore used 35 cycles for all reactions.

1.3. Clean PCR products from primers and nucleotides using a column or modified magnetic beads (See Table of Materials for our choice).

1.4. Add unique dual-index barcodes to cleaned amplicons from the different samples using the Nextera XT kit (Table of Materials) according to manufacturer instructions.

1.5. Purify Indexed amplicons as in step 1.3.

1.6. Quantify concentration of purified Indexed PCR products.

1.7. Pool indexed PCR products and load as per the sequencer manufacturer instructions (Table of Materials).

NOTE: Before loading, sequencing libraries are mixed with a library prepared from the PhiX viral genome (Table of Materials), to space clusters of templates of interest and prevent erroneous merging of adjacent distinct clusters. For many metagenomic applications, the starting percentage of the PhiX library in the final library mix is 1-2%. However, to overcome low sequence diversity in 16S sequences, the ratio of the PhiX library needs to be increased up to 30%.

2. Sequence Analysis with DADA2 [1]

2.1 Download demultiplexed FASTQ files from sequencer and upload files to computer.

2.2 Open RStudio (install first if necessary).

2.3 Load the necessary packages, starting with DADA2 (install first if necessary, available at https://benjjneb.github.io/dada2/dada-installation.html).

```
library(dada2)
library(phyloseq)
library(DECIPHER)
library(phangorn)
```

NOTE: DADA2 was chosen over alternative analysis pipelines for its high taxonomic resolution[2] and compatibility with *ggplot2*, an easy-to-use data visualization tool available in R. The following steps for DADA2 are based on a tutorial available at https://benjjneb.github.io/dada2/tutorial.html, showing only those that contain details directly relevant to the protocol. The full computational pipeline used to generate the representative results is available at https://github.com/kennytrang/CompostMicrocosms.

2.4 Inspect forward and reverse read quality for all samples using plotQualityProfile().

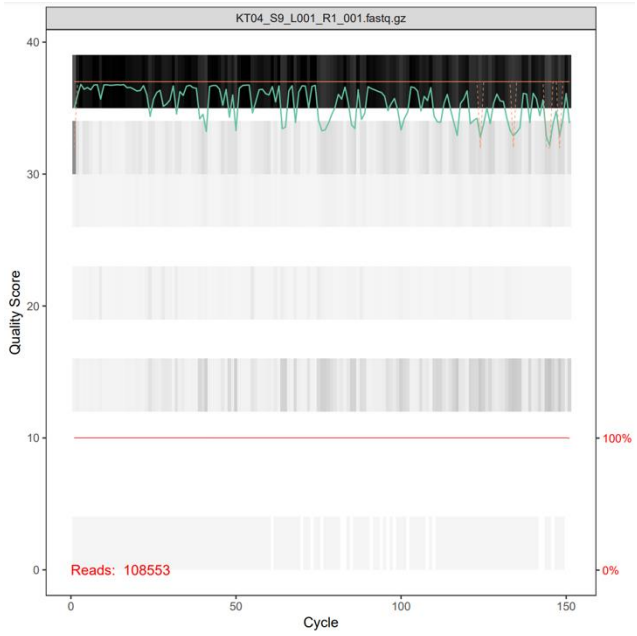

Figure 1. Shown is an example for quality control graph for the reverse reads from one sample. X-axis (cycle) shows nucleotide position along the sequence read. Left Y-axis shows the quality score. Grey-scale heatmap represents the frequency of the quality score at each nucleotide position; green line depicts the median quality score at each nucleotide position; top orange line depicts the quartiles of the quality score distribution; bottom red line depicts the percent of sequence reads that extended that nucleotide position (right y-axis, here 100%).

NOTE: The majority of forward and reverse reads should have quality scores between 30-40 across the entire length of the sequence.

2.5 Filter out low quality reads and remove PhiX sequences using filterAndTrim().

```
filterAndTrim(fnFs, filtFs, fnRs, filtRs,
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=FALSE)
```

NOTE: Quality scores are used to calculate expected errors in sample (EE) according to the formula EE = sum(10^(-Q/10)). Reads with higher EE than defined by the argument maxEE=c() will be removed by the filterAndTrim() function.

NOTE: Primers are not removed from the sequence reads, as those can be used by DADA2 to help build a sequence error model that later on is used to distinguish between sequence variance attributed to errors vs. that represented true sequence and taxonomic variation, providing better resolution among different species/strains.

2.6 Construct error models for each sample using learnErrors().

2.7 Inspect error rates for forward and reverse reads using plotErrors().

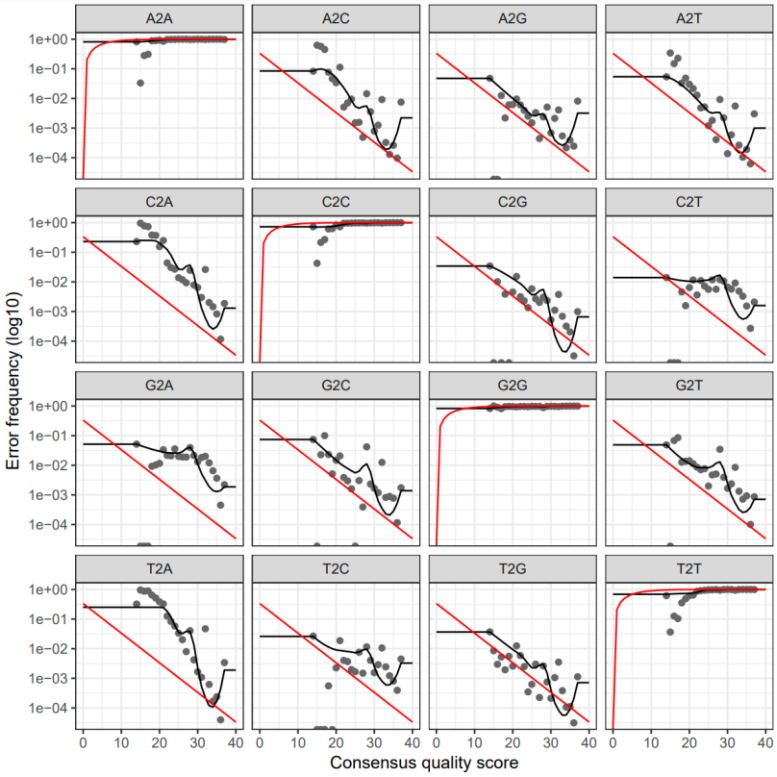

Figure 2. Error frequency in the different samples (black dots) should decrease with increasing quality score for each possible base pair substitution depicted, reflecting the expected trend line (red).

2.8 Bin sequence reads into unique sequences using dada().

NOTE: In the course of binning, dada() removes sequencing from reads (denoising).

2.9 Merge forward and reverse reads using mergePairs().

NOTE: For 2x150bp paired-end sequencing of the 291bp V4 amplicon, the default minimum overlap of 20bp will result in zero reads merged. Minimum overlap of is recommended to allow merging of forward and reverse reads.

2.10 Construct Amplicon Sequence Variant (ASV) table using makeSequenceTable().

2.11 Remove chimera sequences, which are the result of inappropriate merging, using removeBimeraDenovo().

Table 1. Sequential sequence filtering.

Sample	input[a]	filtered[b]	denoisedF[c]	denoisedR[c]	merged[d]	nonchim[d]
KT04	108553	106690	104472	104608	97007	76287
KT05	116008	114368	112429	112647	106384	82430
KT06	117301	115496	111894	112103	98151	69670
KT07	100635	99204	96928	97178	88588	68626
KT08	107581	106076	104044	104263	98327	74812
KT09	105131	103499	101242	101330	92945	64918
KT10	95858	94117	91810	92098	82533	60973
KT11	75303	74127	71951	72074	64099	45072
KT12	112863	111012	108591	109026	100827	80284
KT16	111852	109541	108169	108421	98576	79577
KT17	124179	121915	120259	120662	111581	83812
KT18	87561	85736	84298	84847	76951	63286
KT24	129980	128169	125988	126390	108812	56996
KT25	118163	116391	113723	114307	94927	51029
KT26	99333	97863	95845	96293	81984	46514
KT27	110536	108913	105509	106339	80689	51084
KT28	120512	118736	114625	115946	86059	52783
KT29	96912	95173	91917	93283	72473	42840
KT30	98846	97319	94394	95154	77463	44402
KT31	147246	144807	140698	141367	115911	68664
KT32	145840	143748	139493	140605	115917	69244
KT36	118194	116341	113690	114226	94948	58041
KT37	124923	122747	120118	120248	98416	61807
KT38	99289	97849	95445	96049	77361	41865

[a] Number of sequence reads before filtering.

[b-d] Each column represents the number of sequence reads remaining after a filtration step: filtering-out low quality reads (step 2.5), denoising algorithm performed by dada() (step 2.8), merging forward and reverse reads (step 2.9), and removing chimeras (step 2.11).

NOTE: Typically, around 80% of all reads remain following filtration in worm gut samples and about 60% in environmental compost samples.

2.12 Assign taxonomy to ASVs using assignTaxonomy() and the SILVA or GreenGenes databases.

NOTE: SILVA or GreenGenes are available online for download. SILVA v132 was used in the representative results section. Unlike GreenGenes, new versions are continuously released to update the SILVA database[3] .

2.13 Import metadata table with read.csv(), to associate samples with their identity.

Table 2. Metadata table.

Sample	Subject	SampleName	SampleCode	SampleType	CompostType
KT04	KT04	BP Worm #1	Worm #1	Worm	Bell Pepper
KT05	KT05	BP Worm #2	Worm #2	Worm	Bell Pepper
KT06	KT06	BP Worm #3	Worm #3	Worm	Bell Pepper
KT07	KT07	AP Worm #1	Worm #1	Worm	Apple
KT08	KT08	AP Worm #2	Worm #2	Worm	Apple
KT09	KT09	AP Worm #3	Worm #3	Worm	Apple
KT10	KT10	POT Worm #1	Worm #1	Worm	Potato
KT11	KT11	POT Worm #2	Worm #2	Worm	Potato
KT12	KT12	POT Worm #3	Worm #3	Worm	Potato
KT16	KT16	OR Worm #1	Worm #1	Worm	Orange
KT17	KT17	OR Worm #2	Worm #2	Worm	Orange
KT18	KT18	OR Worm #3	Worm #3	Worm	Orange
KT24	KT24	BP Soil #1	Soil #1	Soil	Bell Pepper
KT25	KT25	BP Soil #2	Soil #2	Soil	Bell Pepper
KT26	KT26	BP Soil #3	Soil #3	Soil	Bell Pepper
KT27	KT27	AP Soil #1	Soil #1	Soil	Apple
KT28	KT28	AP Soil #2	Soil #2	Soil	Apple
KT29	KT29	AP Soil #3	Soil #3	Soil	Apple
KT30	KT30	POT Soil #1	Soil #1	Soil	Potato
KT31	KT31	POT Soil #5	Soil #2	Soil	Potato
KT32	KT32	POT Soil #3	Soil #3	Soil	Potato
KT36	KT36	OR Soil #1	Soil #1	Soil	Orange
KT37	KT37	OR Soil #2	Soil #2	Soil	Orange
KT38	KT38	OR Soil #3	Soil #3	Soil	Orange

3. PCoA analysis

3.1. Select the most abundant ASVs (top 250) to construct a phylogenetic tree for the dataset, necessary to enable calculation of UniFrac distances between microbiomes,

3.2. Construct a phylogenetic tree with phangorn()[4].

NOTE: UniFrac distances take into account phylogenetic relationships between ASV/strains

4.1 Generate a *phyloseq* object with phyloseq(), combining the ASV table (step 2.10, the OTU table (step 2.12), metadata table (step 2.16), and phylogenetic tree (step 3.2).

4.2 Perform Principle Coordinate analysis (PCoA) based on UniFrac distances with ordinate().

NOTE: Ordination with weighted UniFrac distances is carried out with the optional parameter distance = “wunifrac”; with distance = “unifrac” for unweighted UniFrac, which ignore taxa relative abundances; or with distance = “bray” for ordination based on Bray-Curtis distances.

4.3 Plot PCoA results with plot_ordination().

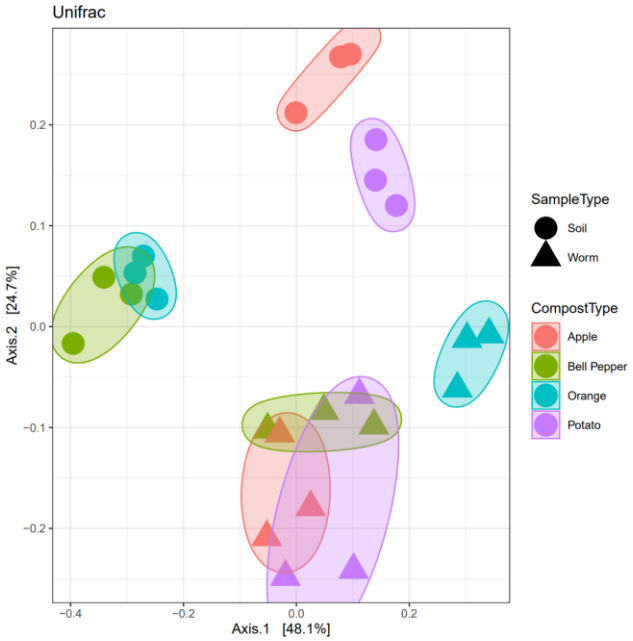

Figure 4. An example of PCoA based on weighted UniFrac distances. Group names shown in legend represent the produce used to enrich the compost used in the different microcosms.

References:

1. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. **13** (7), 581–583, doi: 10.1038/nmeth.3869 (2016).

2. Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A.H., Nieuwdorp, M., Levin, E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE*. **15** (1), doi: 10.1371/journal.pone.0227434 (2020).

3. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*. **41** (D1), doi: 10.1093/nar/gks1219 (2013).

4. Schliep, K.P. phangorn: Phylogenetic analysis in R. *Bioinformatics*. **27** (4), 592–593, doi: 10.1093/bioinformatics/btq706 (2011).