ARL-TR-10319 ● APR 2026

Building a Robust 16S Next-Generation Sequencing Analysis Workflow for Complex Microbial Samples

by Madelyn A. Szilagyi and Randi M. Pullen

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturers or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10319 ● APR 2026

Building a Robust 16S Next-Generation Sequencing Analysis Workflow for Complex Microbial Samples

Madelyn A. Szilagyi
Oak Ridge Associated Universities

Randi M. Pullen
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
April 2026	Technical Report	February 5, 2024	February 28, 2025

4. TITLE AND SUBTITLE
Building a Robust 16S Next-Generation Sequencing Analysis Workflow for Complex Microbial Samples

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Madelyn A. Szilagyi and Randi M. Pullen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-HD Austin, TX 28230	ARL-TR-10319

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID: Madelyn Szilagyi, 000-0001-9762-5194

14. ABSTRACT
Synthetic microbial communities present an opportunity to expand on current synthetic biology by harnessing bioproduction of Army-relevant materials critical to mission success. Population dynamics in a microbial community change over time, so the quickest way to determine the presence and proportion of any organism in the given community is via 16S sequencing. At ARL, we adapted and refined 16S sequencing workflows to study synthetic microbial communities in laboratory settings. Two model communities were used to validate the workflow: The Hitchhikers of the Rhizosphere (THOR), a simple community with three members, and the Caltech Synthetic Community (Caltech SynCom), a more complex community with 14 organisms. Nucleic acid extraction was identified as a major bottleneck in the process, and while the simpler community was successfully validated, genus-level classification of the complex community remained incomplete, likely due to reliance on pre-trained taxonomic classifiers in bioinformatics analysis. This workflow establishes a framework for future 16S sequencing efforts at ARL, with recommendations to train custom classifiers to ensure confident genus-level identification and support the Army's biotechnology initiatives.

15. SUBJECT TERMS
Biological and Biotechnology Sciences; next-generation sequencing; Illumina; bioinformatics; bacterial communities; synthetic biology

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	32

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Madelyn A. Szilagyi	(512) 720-0332

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Microbial communities are essential to the U.S. Army's mission, enabling advanced biomanufacturing and impacting military materiel in both natural and built environments. In biomanufacturing, a critical Department of War (DoW) technology, single microbial species are limited in their ability to synthesize complex bioproducts. Multi-species microbial communities, however, unlock greater potential for chemical synthesis and material production by distributing metabolic tasks across multiple organisms. This focus on in-house bioproduction not only enhances material output but also fortifies supply chain security, ensuring the Army's readiness for future challenges. Next-generation sequencing (NGS) is a critical tool for understanding how these communities are structured and function collaboratively. This technology is widely used for applications such as identifying organisms, studying genetic variations, and understanding complex biological systems. By leveraging DNA sequencing to characterize and understand these communities, the Army can harness their capabilities to drive innovation and enhance readiness across diverse applications.

Sequencing synthetic microbial communities poses unique challenges due to their diverse composition, sometimes including both bacteria and fungi. These organisms have distinct types of cell envelopes, requiring different lysis protocols and extensive optimization to process them effectively. To ensure our sequencing workflow is accurate and reliable, we validated it using model bacterial communities (Figure 1). In this study, we employed two model communities: The Hitchhikers of the Rhizosphere (THOR)[1] and the Caltech Synthetic Community (Caltech SynCom). This report provides a detailed account of the sequencing workflows tailored for laboratory characterization of synthetic microbial communities, providing a snapshot of their population dynamics.

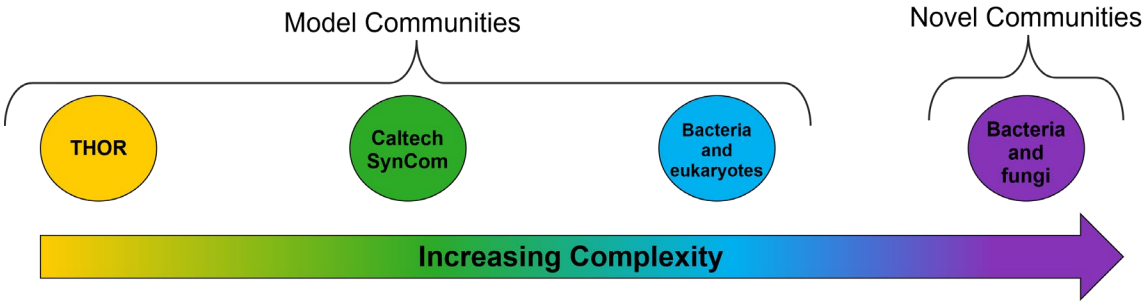

Figure 1. Overarching workflow for the microbial communities project.

2. Communities

2.1 Bacterial Communities

Bacterial communities are ubiquitous, found in diverse environments such as the human gut, kitchen counters, plant root soils, and deep-sea geothermal vents. Within these communities, bacteria interact in various ways, including promoting or inhibiting growth, exchanging metabolites, and engaging in quorum sensing.[2] For example, in the THOR community discussed below, *Pseudomonas koreensis* produces an antibiotic called koreenceine that inhibits the growth of *Flavobacterium johnsoniae*, but the production of koreenceine is suppressed by *Bacillus cereus*.[1] These dynamic interactions can drive shifts in community composition over time. By using 16S sequencing, researchers can capture a snapshot of these evolving microbial communities and better understand their structure and function.

2.2 Model Communities

The first synthetic community analyzed was the THOR community, a microbial community characterized by the Handelsman Lab at the University of Wisconsin–Madison. They chose three bacterial species representative of the dominant phyla in the rhizosphere to form a tractable model of community-level interactions: *Pseudomonas koreensis*, *Bacillus cereus*, and *Flavobacterium johnsoniae* (Table 1).[1] THOR was used as a model community to validate our bacterial community sequencing and data analysis workflows.

Further details about these organisms are listed in Table 1. Whether a microbe is Gram-positive or Gram-negative is critical information, especially for nucleic acid extraction. Gram-negative bacteria have an outer cell membrane, while Gram-positive bacteria do not. The outer membrane of Gram-negative bacteria makes cell lysis—and therefore nucleic acid extraction—more difficult. Another important metric to consider is guanine-cytosine (GC) content, because certain types of sequencers have difficulty reading sequences with high GC content. Knowing the genome size for each organism is also necessary, because genome size impacts steps in the sequencing workflow and is a critical number for calculating sequencing depth.

Table 1. THOR bacterial community composition.

Organism	Gram	NCBI reference genome ID	Genome size (Mb)	GC content (%)	Coding genes
Bacillus cereus	+	ASM167816v1	6.4	35	6,396
Flavobacterium johnsoniae	–	ASM1664v1	6.1	34	5,249
Pseudomonas koreensis	–	ASM200342v1	6.6	59	5,953

The second synthetic community selected to validate the sequencing and data analysis pipelines is called Caltech SynCom, which is more complex than THOR. The Newman Lab at the California Institute of Technology (Caltech) developed an experimental bacterial community consisting of 14 bacteria isolated from soil (Table 2).

Table 2. Caltech Synthetic bacterial community composition.

Organism	Gram	NCBI reference genome ID	Genome size (Mb)	GC content (%)	Coding genes
Paraburkholderia graminis	–	ASM333078v1	7.5	63	6,573
Microbacterium murale	+	ASM1463518v1	3.9	66.5	3,704
Pseudomonas synxantha 2-79	–	ASM385149v1	6.5	60	5,632
Arthrobacter agilis	+	ASM2873605v1	3.8	68	3,488
Phyllobacterium ifrigiyense	–	ASM3081788v1	4.8	53.5	4,443
Pedobacter kyongii	–	ASM431066v1	6.2	39	5,092
Pseudomonas synxantha (field isolate)	–	ASM385155v1 (recc ref)	6.6	59.5	5,855
Arthrobacter pascens DSM 20545	+	ASM1705246v1	4.8	65.5	4,277
Pantoea agglomerans DSM 3493	–	ASM3081543v1 *(strain DSM 3493_2)	4.8	55	4,336
Variovorax paradoxus	–	ASM3380785v1	7.1	67.5	6,571
Chryseobacterium shigense	–	ASM1420784v1	4.2	37	3,759
Chitinophaga ginsengisegetis	–	IMG-taxon 2595698247 annotated assembly	7.9	46.5	6,301
Sphingomonas faeni	–	ASM305374v1	4.4	65	3,995
Pseudomonas orientalis	–	ASM385204v1	5.9	60.5	5,148

3. Sequencing

3.1 16S Sequencing vs. Shotgun Metagenomic Sequencing

16S sequencing is a form of NGS[3] primarily used for taxonomic analysis of bacterial species. The 16S ribosomal RNA (rRNA) gene is the sequence of DNA coding for the 16S ribosomal subunit that is found in all bacterial genomes. It

contains nine hypervariable regions, labeled V1 to V9, which are flanked by evolutionarily conserved regions (Figure 2).[3] 16S sequencing requires a polymerase chain reaction (PCR) amplification step using primers targeting these hypervariable regions—in this case the V3–V4 region. The 16S rRNA gene is highly conserved within species, allowing it to be used for species identification and/or community composition analysis down to the genus level.[3] 16S sequencing allows researchers to perform taxonomic analysis by determining the composition of a community. 16S sequencing is particularly useful for longitudinal studies, where the same microbial community is observed over a period of time to track changes and trends.[4] 16S longitudinal studies can be used to observe changes in a variety of variables, such as relative abundance, response to changes in environmental factors, and intracommunity interactions.[4]

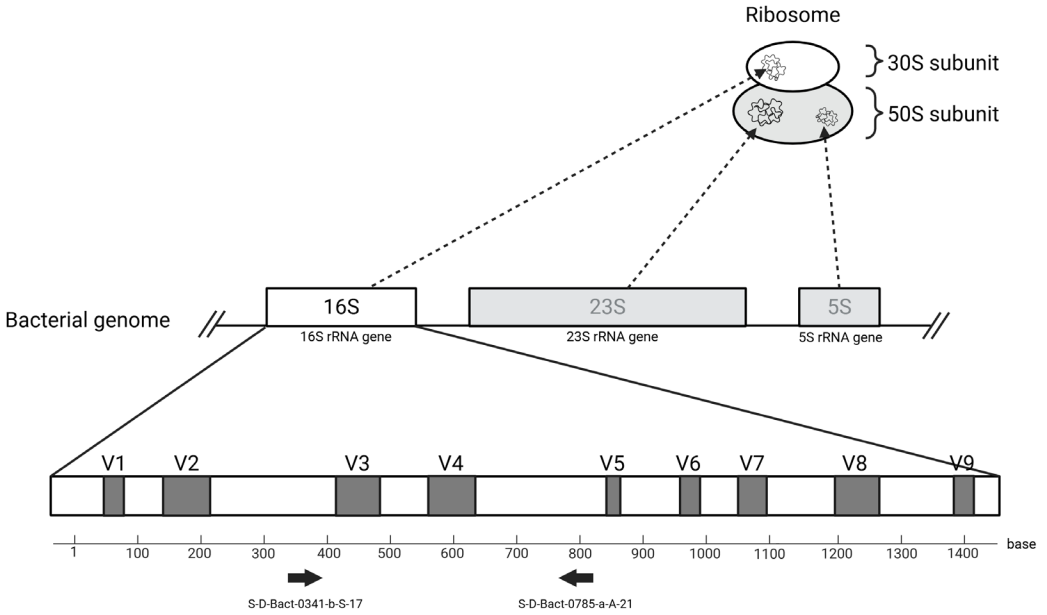

Figure 2. Diagram of the 16S rRNA gene with the nine hypervariable regions (V1–V9) marked. The two primers used for 16S amplification[5] are marked by the arrows at the bottom. (Diagram adapted from Fukuda et al.[6])

One alternative to 16S sequencing is shotgun metagenomic sequencing. This method utilizes whole-genome sequencing to capture and sequence all genomic DNA in a sample.[7]

Both 16S sequencing and shotgun metagenomic sequencing have their advantages and limitations. 16S sequencing is more cost-effective, simpler on the bioinformatics side, has lower rates of false positives, and requires less DNA input compared to shotgun metagenomic sequencing.[7,8] However, it is primarily limited to taxonomic composition analysis. 16S sequencing is particularly useful for bacterial identification, including bacteria without reference genomes, making it

valuable for studying unknown communities. For an unknown bacterium, 16S sequencing would likely be able to provide some sort of taxonomic identification, if only at a higher phylogenetic level. 16S reference databases are more comprehensive for bacteria than whole-genome databases, which tend to be biased toward organisms found in the human microbiome.

Shotgun metagenomic sequencing, on the other hand, offers broader capabilities. It is versatile, enabling analyses such as metabolic function profiling, metagenome assembly, and antibiotic resistance gene profiling.[7] It also provides higher resolution taxonomic identification, down to the species or strain level, and can analyze bacteria, archaea, viruses, and microbial eukaryotes.[7] One study found that shotgun metagenomic sequencing identified more genera, especially those that are less abundant in the sample.[8] However, it is more expensive, bioinformatically complex, and heavily reliant on the choice of reference database—unless a closely related reference genome is present in the chosen database, it is likely the organism will be missed.[7]

3.2 Short- vs. Long-Read Sequencing

There are two primary types of NGS methods: short-read sequencing using Illumina technology and long-read sequencing using Oxford Nanopore Technology (ONT); both of these technologies are considered state-of-the-art as of the time of this report. Illumina technology utilizes short-read sequencing called sequencing by synthesis. In this method, millions of DNA fragments are synthesized in parallel on a flow cell, using the input DNA library as template strands. As each deoxynucleotide triphosphate (dNTP) is added to the growing strand, the attached fluorescently labeled reversible terminator is imaged to identify the base before being cleaved to allow the next base to be incorporated.[9] This step-by-step imaging process enables the sequencing of each base as the growing strand is synthesized. The resulting short reads are then bioinformatically aligned to a reference genome or, in some cases, used to assemble a genome without a reference. Illumina is classified as short-read sequencing because its maximum read length is 2 × 300 base pairs (bp), requiring the DNA to be fragmented before sequencing.[9]

ONT is classified as long-read sequencing, which, unlike Illumina sequencing, does not require DNA fragmentation before sequencing. ONT uses a flow cell containing millions of specialized nanopores, each paired with an electrode and sensor chip to measure electric current. As a strand of target DNA passes through a nanopore, the current is disrupted, producing a characteristic electrochemical change corresponding to which base is present. The sequencing software then interprets these electrochemical changes and records the order of each base, a process called

basecalling.[10] Illumina's maximum read length of 2 × 300 bp is limiting, requiring longer DNA sequences (such as the 16S rRNA gene) to be fragmented before sequencing; however, it features very low error rates, with 99.9% accuracy.[11] ONT long-read sequencing, on the other hand, can sequence the entire 16S rRNA gene in a single read, providing increased taxonomic resolution and improving species identification accuracy.[12] Historically, ONT sequencing has been less accurate than Illumina sequencing, but recent advancements in library preparation chemistry and the ONT basecalling software have led to a level of accuracy comparable to Illumina.[13]

For this study, Illumina short-read 16S sequencing was chosen because the organisms in the synthetic communities were already known, making more in-depth sequencing unnecessary.

4. Methods

4.1 16S Sample Workflow

The workflow for analyzing 16S samples, illustrated in Figure 3, is summarized here and further detailed in the following sections. The process begins with receiving the samples as cell pellets, which are lysed to release their contents. Following lysis, DNA is isolated, and 16S amplification is performed using PCR. The amplified 16S PCR products are then quantified before moving to library preparation. After library preparation, the resulting libraries are normalized and pooled. The pooled libraries are diluted to the appropriate loading concentration and loaded onto the Illumina iSeq100 sequencer. Once sequencing is complete, the data is analyzed.

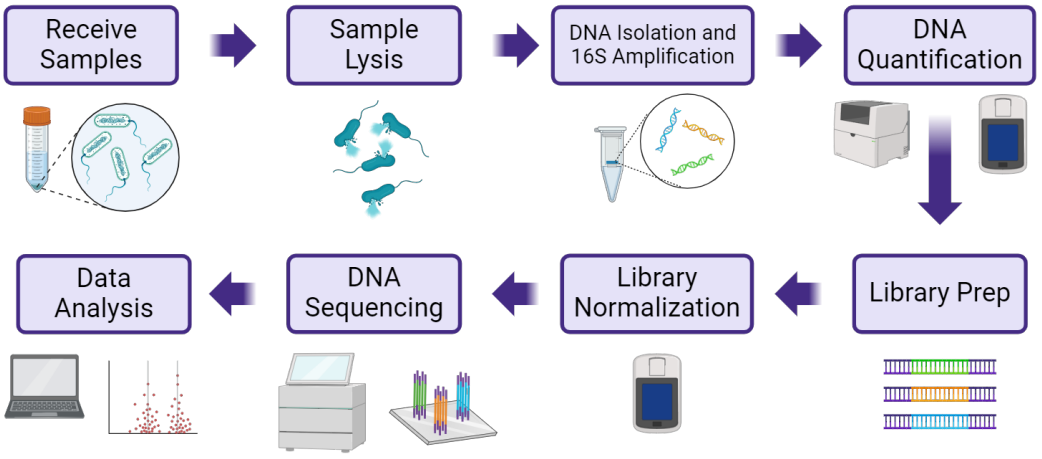

Figure 3. Workflow for analyzing 16S samples.

4.2 Sample Preparation

For both communities, samples were received from collaborators as frozen cell pellets. The THOR samples were lysed mechanically using QIAGEN PowerBead Tubes 0.1 mm (QIAGEN, cat. no. 13118-50). The tubes were placed on the QIAGEN TissueLyser LT (QIAGEN, cat. no. 85600) for 10 min, rested on ice for 5 min, then the procedure was repeated. The DNA and RNA were isolated using the ZymoBIOMICS DNA/RNA Miniprep Kit (Zymo Research Corp., cat. no. R2002) following manufacturer instructions. The DNA was quantified using the Thermo Qubit fluorometer and the Agilent TapeStation.

The Caltech SynCom samples were lysed using Metapolyzyme (MilliporeSigma, cat. no. MAC4L-5MG) and Proteinase K (New England Biolabs, cat. no. P8107S). We performed significant lysis optimization using isolates of each community member to ensure the final protocol would lyse all community members in the experimental samples. Total nucleic acid was isolated using the Lucigen MasterPure DNA/RNA Isolation Kit (LCG Biosearch Technologies, cat. no. MC85200). The resulting nucleic acid pellet was resuspended and split in half. One half was treated with DNase so only the RNA remained, while the other half was treated with RNase so only the DNA remained. The DNA was quantified using the Thermo Qubit fluorometer and the Agilent TapeStation.

4.3 16S Amplification

To sequence the 16S region, it was first amplified using PCR, which required primers with Illumina adapter "tails" specific to the Nextera XT kit. For this process, primers with Nextera-like tails targeting the 16S V3–V4 region were used. The PCR setup was then placed in a thermocycler, following the protocol outlined in Table 3.

Table 3. Thermocycler protocol used for 16S amplification.

No. of cycles	Temperature (°C)	Time
	95	2 min
25X	95	20 s
	55	10 s
	70	15 s

4.4 Library Preparation and Sequencing

For both communities, the amplified 16S DNA was converted into a library using the Illumina Nextera XT DNA Library Prep Kit (Illumina, cat. no. FC-131-1024)

with Illumina Nextera XT Index Kit Set A (Illumina, cat. no. TG-131-2001). The libraries were quantified using the Thermo Qubit fluorometer, normalized to 2 nM, then pooled. This pool was diluted to 150 pM, the loading concentration for the Nextera XT kit for the iSeq100 sequencer as provided by Illumina. A 10% PhiX spike-in was also added as a control. Both of the 16S libraries were sequenced on the Illumina iSeq100 sequencer. The Illumina iSeq100 sequencer has built-in software that demultiplexes the samples, converts the data into FASTQ format, and trims the adapters.

4.5 Sequencing Analysis Viewer (SAV) Metrics

After the sequencing run is completed, the sequencer generates a variety of run metrics. These metrics can be reviewed using Illumina's SAV, a program that provides detailed charts and data to assess the quality of the sequencing run (Table 4).

Table 4. Acceptable ranges for each of the SAV metrics that are used for judging the quality of a sequencing run.

SAV metric	Acceptable range	Experimental results – THOR	Experimental results – Caltech SynCom
% ≥ Q30	>80%[14]	89.27%	81.84%
Total yield	>1.2 Gbase[14]	1.94 Gbase	1.38 Gbase
% occupied	80%–90%[15]	91.53%	94.80%
% pass filter	>80%[16]	80.28%	57.16%

4.6 Data Analysis

Once the samples have been sequenced and the quality of the run in SAV has been confirmed, the sequencing data is transferred to a DoW high-performance computer for analysis (Figure 4). The software Quantitative Insights into Microbial Ecology 2, or QIIME2,[17] is used to analyze the 16S data. This software was developed specifically for 16S analysis of microbial communities.

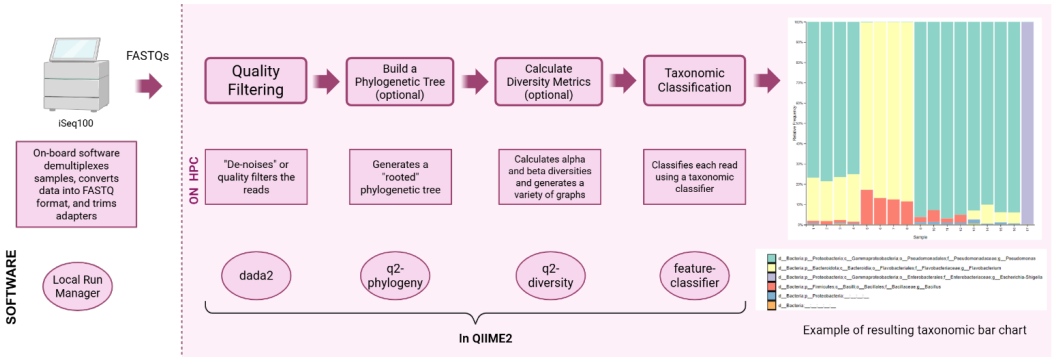

Figure 4. Workflow for analyzing 16S data using QIIME2.

First, a visualization is generated of the average quality of the data per base. A software package within QIIME2, Divisive Amplicon Denoising Algorithm 2 (DADA2), is then used for quality control and denoising Illumina sequencing data. If there is an average of low-quality reads at the beginning or the end of the reads, they can be removed with DADA2. Next, QIIME2 generates a phylogenetic tree for phylogenetic diversity analyses. This tree is then used to compute alpha and beta diversity metrics, including Principal Coordinate Analysis (PCoA) plots, using a variety of methods. The displayed plots allow the user to evaluate the similarity or dissimilarity between samples, providing valuable insights, particularly when analyzing replicates. Alpha diversity is explored as a function of sampling depth via alpha rarefaction plotting in QIIME2 and a taxonomic classifier is used to assess the taxonomic composition of the samples. Interactive taxonomic bar charts are generated to visualize the composition of each sample, providing an intuitive way to explore and compare microbial communities.

5. Results and Discussion

5.1 Comparison of SAV Metrics

The THOR sequencing run achieved better overall sequencing quality as compared to the Caltech SynCom. Figure 5 shows the % ≥Q30 for the two communities, or the percentage of sequencing reads with a quality score higher than Q30 per cycle. Q30 is a widely used quality benchmark, with 80%–100% of reads above Q30 indicating a high-quality sequencing run.[14] A lower-quality run can often be identified by the presence of "raindrops," which are cycles where the % ≥Q30 drops suddenly. In this analysis, the graph for THOR shows more raindrops compared to the graph for Caltech SynCom, but THOR has a higher overall % ≥Q30, indicating better overall sequencing quality.

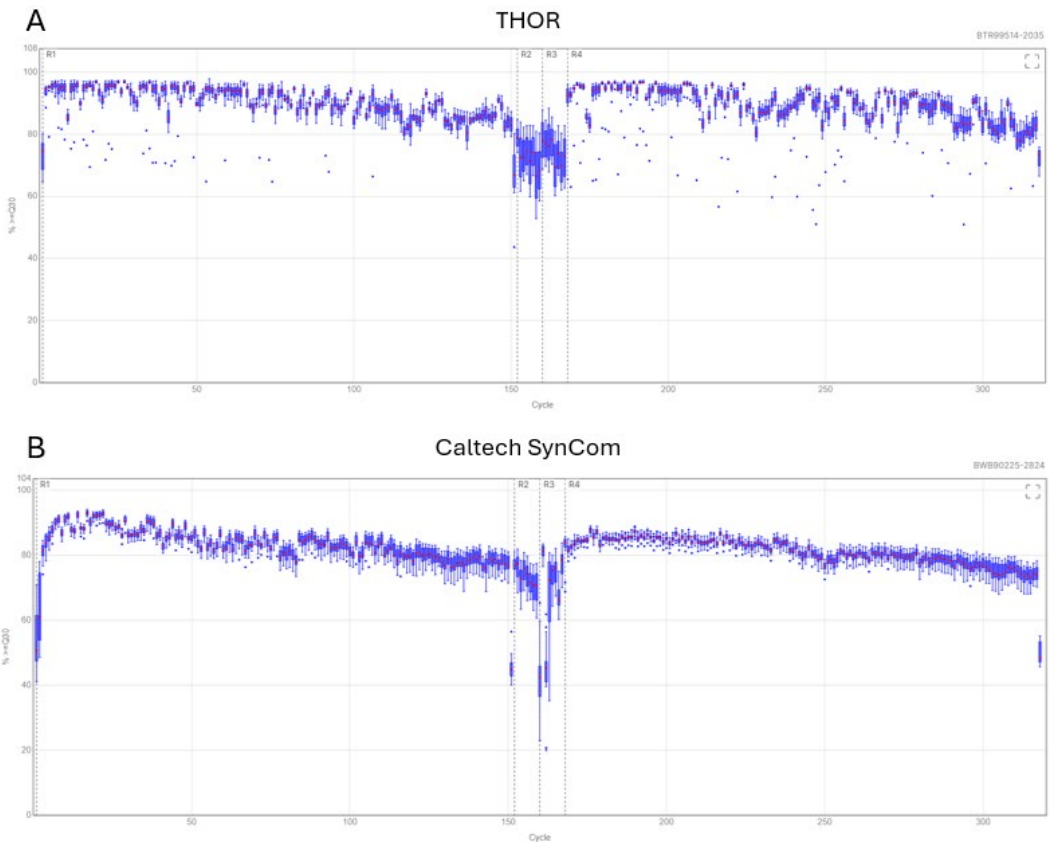

Figure 5. Graphs of % ≥Q30 per cycle from SAV for the THOR community (A) and the Caltech Synthetic Community (B).

For a read length of 2 × 150 bp, a good sequencing run typically generates a total yield of at least 1.2 Gbase.[14] In this study, the THOR run produced a total yield of 1.94 Gbase, while the Caltech SynCom run generated 1.38 Gbase (Figure 6). Additionally, the THOR run achieved a higher average % ≥Q30 compared to the Caltech SynCom run, indicating better overall sequencing quality (Figures 5 and 6). However, the THOR run had a slightly lower % Occupied (91.53%) than the Caltech SynCom run (94.8%). The % Occupied metric reflects how occupied the flow cell was and serves as an indicator of cluster density. Ideally, a good sequencing run should have a % Occupied between 80% and 90%.[15] Both runs exceeded this range, suggesting that the flow cell was overloaded. For future runs, the loading concentration should be reduced to prevent overloading and improve sequencing performance.

A

THOR

Level	Yield Total (GB)	Projected Yield (GB)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 1	% >= Q30	% Occupied
Read 1	0.93	0.93	20.58	0.76	95.06	73.53	90.44	91.53
Read 2 (I)	0.04	0.04	0	NaN	128	99.01	73.16	91.53
Read 3 (I)	0.04	0.04	0	NaN	141.13	109.16	74.78	91.53
Read 4	0.93	0.93	19.86	0.75	152.94	118.3	89.53	91.53

B

Caltech SynCom

Level	Yield Total (GB)	Projected Yield (GB)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 1	% >= Q30	% Occupied
Read 1	0.66	0.66	3.66	1.41	99.31	85.75	83.46	94.8
Read 2 (I)	0.03	0.03	0	NaN	129.5	111.82	74.01	94.8
Read 3 (I)	0.03	0.03	0	NaN	126.75	109.44	64.7	94.8
Read 4	0.66	0.66	3.31	1.23	107.69	92.98	81.39	94.8

Figure 6. Run summary tables from SAV for the THOR community (A) and the Caltech Synthetic Community (B).

% Occupied can be plotted against % Pass Filter to create a graph to look at cluster density (Figure 7). % Pass Filter represents the ratio of high-quality clusters to total clusters on a flow cell and indicates signal purity.[16] Cluster density is a critical factor in sequencing, as excessively high cluster density (overloading) can result in reduced data output and lower-quality sequencing data.[16] In this study, the THOR graph closely resembled the reference graph, whereas the Caltech SynCom graph deviated more significantly. This difference is attributed to the higher % Pass Filter observed in the THOR run compared to the Caltech SynCom run. Despite these differences, both runs produced sufficient high-quality data for downstream analysis.

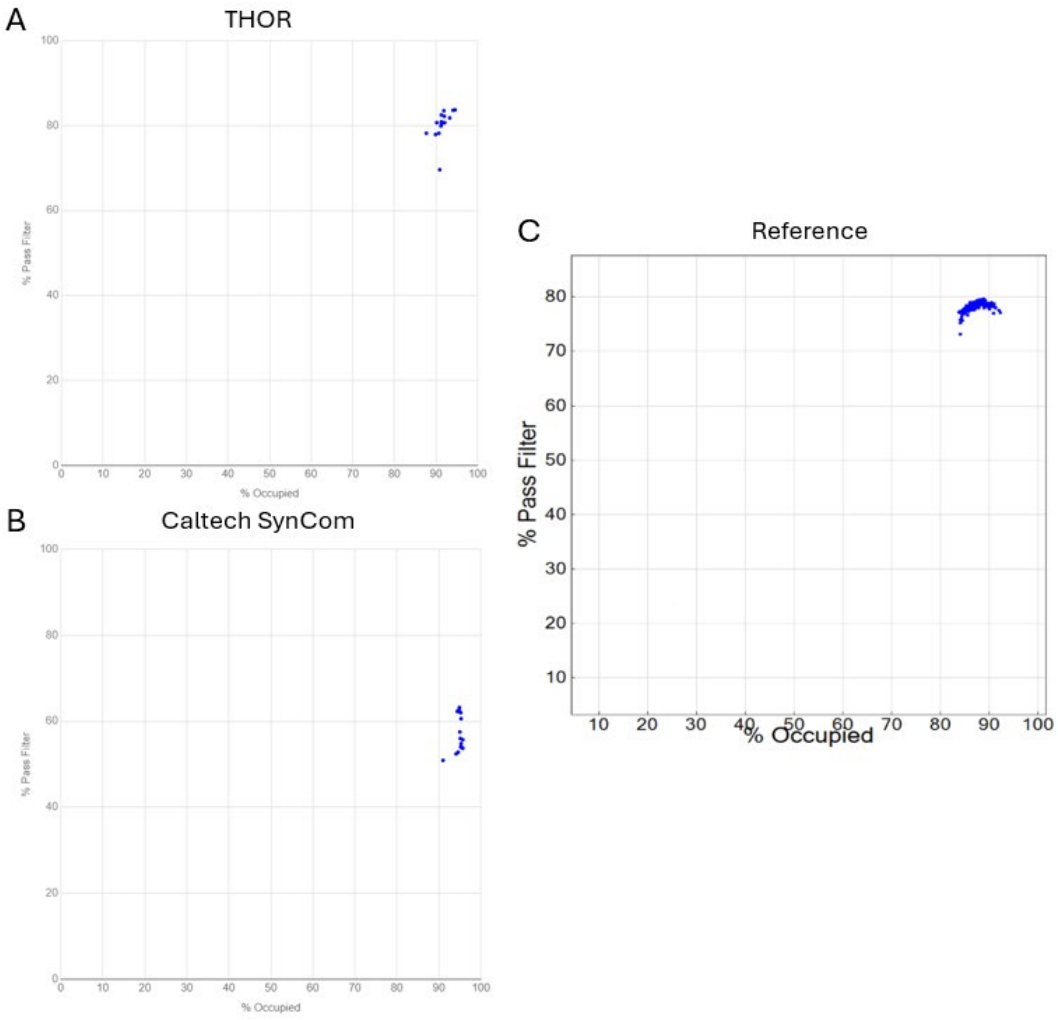

Figure 7. SAV graph % Occupied vs. % Pass Filter for the THOR community (A), the Caltech Synthetic Community (B), and the reference graph provided by Illumina[15] (C). Note that the scale for the reference graph does not match the scale of the experimental graphs.

Both % Base graphs show uneven proportions of the four nucleotides and high-intensity spikes at each cycle, which is expected for a 16S sequencing run due to the low-complexity nature of the amplicon-only library (Figure 8).[18] The low diversity of these libraries results in an uneven distribution of nucleotides across the flow cell that changes dramatically from one cycle to the next, leading to the intensity spikes observed in the graphs.[18] The % Base graph for the Caltech SynCom appears more compressed compared to THOR's % Base graph because the Caltech SynCom library has slightly higher complexity. In contrast, for fully diverse libraries, the lines on the % Base graph would be relatively flat and clustered around 25% for each nucleotide.[18]

Although differences in SAV metrics were observed between the two sequencing runs, they did not affect the quality of the final data. These metrics can serve as valuable references for optimizing future community sequencing runs.

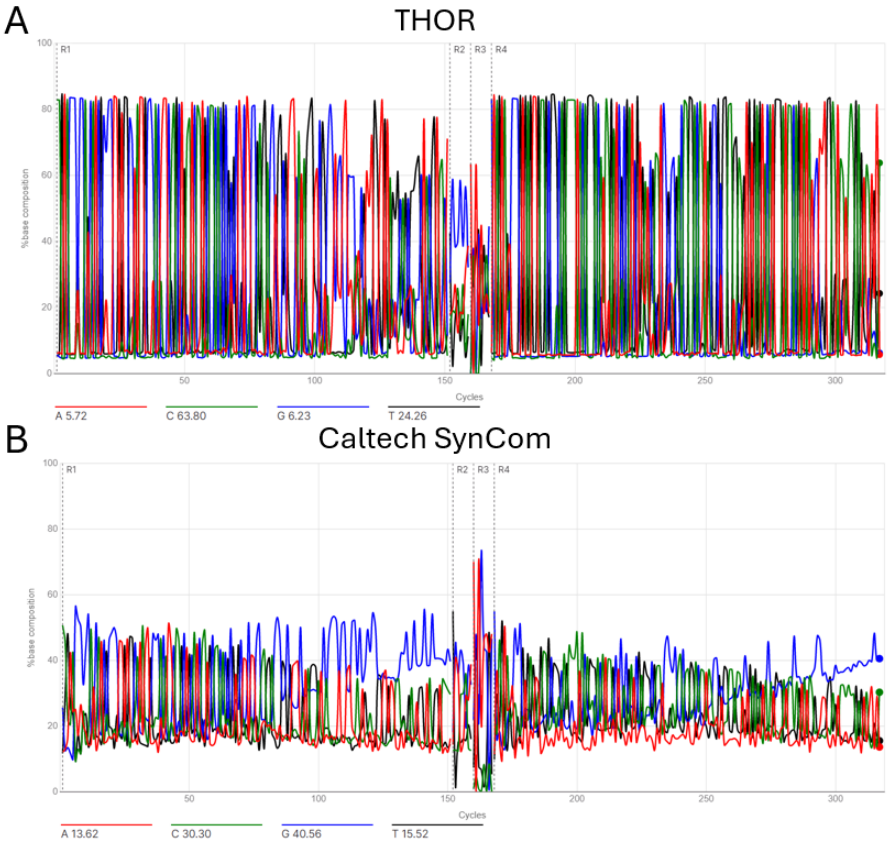

Figure 8. Graphs of % base per cycle from SAV for the THOR community (A) and the Caltech Synthetic Community (B).

5.2 16S Results for THOR

The bar charts generated by the taxonomic classifiers show the proportion of each organism in each sample. The index on the right shows the identification corresponding to each color; the identifications are listed in size order (largest to smallest).

Samples 1–4 are a mixture of all three organisms in the THOR community: *Bacillus cereus*, *Flavobacterium johnsoniae*, and *Psuedomonas koreensis*. Samples 5–8 are composed of *B. cereus* and *F. johnsoniae*, Samples 9–12 are composed of *B. cereus* and *P. koreensis*, and Samples 13–15 are composed of *F. johnsoniae* and *P. koreensis*. The inoculations of cocultures were normalized so that the starting number of cells for each member was about the same. The cultures were incubated at 20 °C for 19 h before being processed for 16S sequencing. The 16S sequencing reads were analyzed using the standard taxonomic bacterial classifier in QIIME2

and a taxonomic bar chart was generated (Figure 9). Although the taxonomic classifier was only able to identify organisms down to the genus level, the genii identified in the taxonomic bar chart correspond with the three member organisms. The community dynamics of the THOR community are reflected in the relative proportions: *P. koreensis* inhibits the growth of *F. johnsoniae* by producing an antibiotic called koreenceine, but production of koreenceine is inhibited by *B. cereus*. Samples 5–8, where *F. johnsoniae* is present and *P. koreensis* is not, show a majority of *F. johnsoniae*. The results of the 16S sequencing portray the trend in population dynamics we expect based on the known small molecule-based cell–cell communication that has been characterized for THOR.

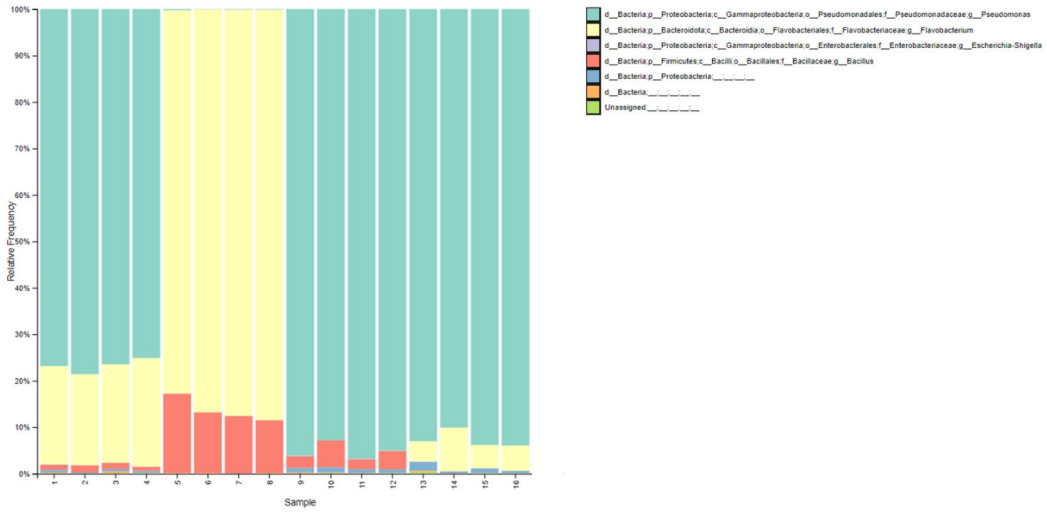

Figure 9. Taxonomic bar plot generated at the end of the QIIME2 workflow using the standard bacterial taxonomic classifier for the THOR community.

These results are considered to be reliable because the SAV run metrics indicated high-quality data, quality filtering was performed using QIIME2, and nearly all of the identifications correspond to the three representative species.

5.3 16S Results for Caltech SynCom

The taxonomic analysis for Caltech SynCom was more complicated than for THOR due to the higher complexity of the community. To address this, we employed two different bacterial taxonomic classifiers: a "standard" classifier (used for the THOR analysis) and a "diverse" classifier, which was trained on bacteria isolated from a wide range of environments. Both classifiers revealed that Groups 1 and 3 are similar, as are Groups 2 and 4. This observation is further supported by PCoA plots of the data (not shown). Groups 1 and 3 were grown at pH 5, while Groups 2 and 4 were grown at pH 7, indicating that the pH of the environment significantly influences the composition of the microbial community.

The results from the standard bacterial taxonomic classifier are shown in Figure 10. The second-largest category identified with this classifier is "unassigned," meaning the classifier could not identify the reads as bacterial. The largest category is only identified to the order level, Enterobacterales, which could correspond to more than one organism in the community. There is also a significant percentage identified as "Eukaryota," which suggests there may be fungal contamination in the samples. Table 5 lists the organisms expected to be found in the Caltech SynCom sample and the average percentage identified within each experimental group using the standard bacterial classifier.

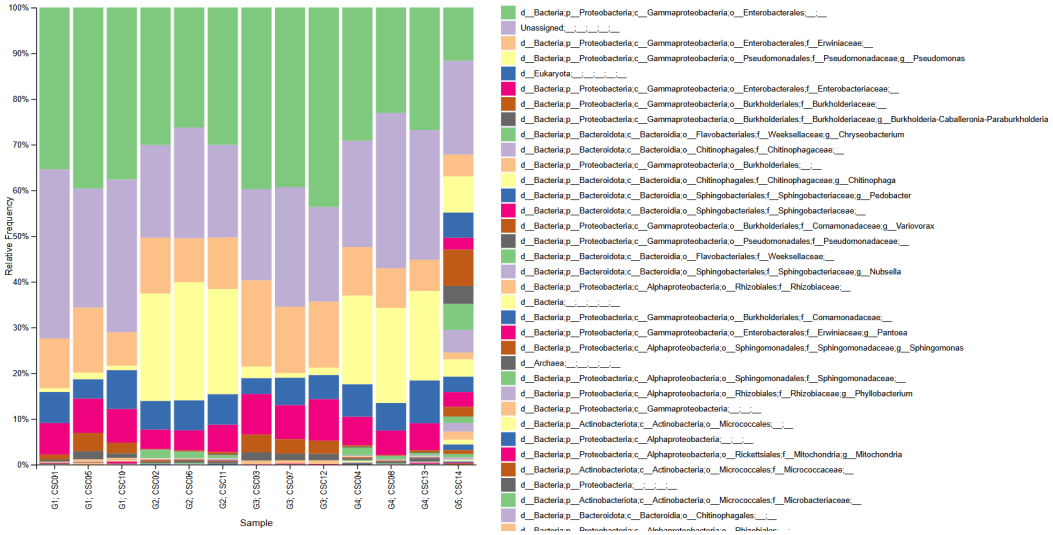

Figure 10. Taxonomic bar plot generated at the end of the QIIME2 workflow using the standard bacterial taxonomic classifier for the Caltech Synthetic Community

Table 5. Percentage identified per group of each expected organism in the Caltech Synthetic Community using the standard bacterial taxonomic classifier.

Organism	Family or genus level	% Identified in Group 1 (avg.)	% Identified in Group 2 (avg.)	% Identified in Group 3 (avg.)	% Identified in Group 4 (avg.)
Paraburkholderia graminis	F	2.49	0.29	3.32	0.34
Microbacterium murale	F	…	…	…	…
Pseudomonas synxantha 2-79	G[a]	1.24	24.58	1.73	20.50
Arthrobacter agilis	F[a]	…	…	…	…
Phyllobacterium ifrigiyense	G	…	…	…	…
Pedobacter kyongii	G	…	0.11	…	0.07
Pseudomonas synxantha (field isolate)	G[a]	1.24	24.58	1.73	20.50

[a] Because the classifier can only classify down to the genus level, these organisms have the same percentages.

Table 5. Percentage identified per group of each expected organism in the Caltech Synthetic Community using the standard bacterial taxonomic classifier (continued).

Organism	Family or genus level	% Identified in Group 1 (avg.)	% Identified in Group 2 (avg.)	% Identified in Group 3 (avg.)	% Identified in Group 4 (avg.)
Arthrobacter pascens	F[a]	…	…	…	…
Pantoea agglomerans	F	11.10	11.05	16.06	8.92
Variovorax paradoxus	G	…	0.25	…	0.20
Chryseobacterium shigense	G	…	1.42	…	1.13
Chitinophaga ginsengisegetis	G	…	0.13	…	0.19
Sphingomonas faeni	G	…	…	…	…

[a] Because the classifier can only classify down to the genus level, these organisms have the same percentages.

The results from the Caltech SynCom analysis are considered less trustworthy compared to those from the THOR analysis because the second largest group is labeled as "unassigned," indicating that the standard bacterial taxonomic classifier was unable to identify those reads. This suggests that the organisms present in the Caltech SynCom were likely not included in the training set for the standard classifier. To address this limitation, taxonomic analysis was performed using a diverse bacterial classifier, which was trained on bacteria from disparate environments, offering broader coverage for identification.

The results from using the diverse bacterial taxonomic classifier (Figure 11) are overall more specific, or able to identify to a greater classification level, than the results from the standard bacterial taxonomic classifier. The most identified group was of the genus *Pantoea*, which most likely corresponds with the organism *Pantoea agglomerans* in our community. The unassigned category is now the fifth largest group, but the generic "bacteria" group is the second-largest identified group, so this classifier is still not as successful as we hoped. Table 6 lists the organisms expected to be found in the Caltech SynCom sample and the average percentage identified within each experimental group using the diverse bacterial taxonomic classifier.

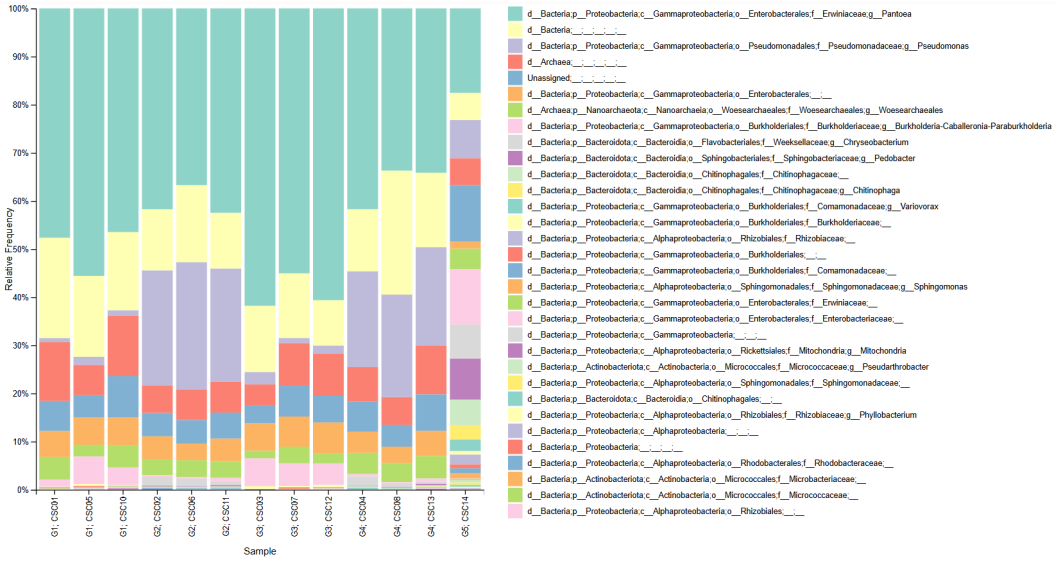

Figure 11. Taxonomic bar plot generated at the end of the QIIME2 workflow using the diverse bacterial taxonomic classifier for the Caltech Synthetic Community.

Table 6. Percentage identified per group of each expected organism in the Caltech Synthetic Community using the diverse bacterial taxonomic classifier.

Organism	Family or genus level	% Identified in Group 1 (avg.)	% Identified in Group 2 (avg.)	% Identified in Group 3 (avg.)	% Identified in Group 4 (avg.)
Paraburkholderia graminis	F	3.63	0.50	4.99	0.45
Microbacterium murale	F	…	…	…	…
Psuedomonas synxantha 2-79	G [a]	1.19	24.58	1.73	20.53
Arthrobacter agilis	F [a]	…	…	…	…
Phyllobacterium ifrigiyense	G	…	…	…	…
Pedobacter kyongii	G	…	0.16	…	0.19
Pseudomonas synxantha (field isolate)	G [a]	1.19	24.58	1.73	20.53
Arthrobacter pascens	F [a]	…	…	…	…
Pantoea agglomerans	F	50.11	40.44	59.17	28.48
Variovorax paradoxus	G	…	0.11	…	0.35
Chryseobacterium shigense	G	…	1.38	…	1.07
Chitinophaga ginsengisegetis	G	…	0.14	…	0.14
Sphingomonas faeni	G	…	…	…	…
Pseudomonas orientalis	G [a]	1.19	24.58	1.73	20.53

[a] Because the classifier can only classify down to the genus level, these organisms have the same percentages.

These results are considered slightly more trustworthy than those obtained using the standard bacterial taxonomic classifier because the diverse classifier was trained on a broader set of bacteria from diverse environments, increasing the likelihood of accurately identifying uncommon organisms. This classifier also identified more organisms in the community at a higher taxonomic resolution (e.g., genus level) compared to the standard classifier.

Some of the organisms in this community are relatively uncommon, so it is not surprising that the pretrained taxonomic classifiers were unable to successfully identify all the organisms expected to be present. For more accurate classification, it would likely be necessary to train a custom taxonomic classifier. Another possible explanation is that competition between organisms within the community may have allowed keystone organisms to dominate, outcompeting other organisms in the community. This highlights the importance of studying bacterial community dynamics over time, as the community composition determined by 16S sequencing represents only a snapshot at the specific time the sample was taken.

5.4 Optimal Timeline for 16S Analysis

The optimal timeline for completing sample preparation, sequencing and data analysis is approximately 10–13 business days from the reception of cell pellets (Figure 12). The workflow is as follows:

Days 1–5: Sample lysis, DNA isolation, and DNA quantification. These steps require multiple days due to the need for lysis optimization, as the complex communities may contain members with varying and potentially unknown cell envelope compositions.

Day 6: Samples are diluted, 16S PCR amplification is carried out and the resulting PCR products are analyzed using an agarose gel or TapeStation tape to confirm successful amplification. The PCR products are subsequently cleaned and quantified.

Day 7: The cleaned 16S PCR products are normalized to the input amount for library preparation. Samples are manually prepared into libraries using the Illumina Nextera XT Library Prep kit.

Day 8: 16S libraries are quantified, normalized, and pooled. They are then loaded onto the Illumina iSeq100 sequencer, which runs overnight until Day 9.

Days 9–13: The remaining time is dedicated to data analysis, primarily taxonomic analysis.

This timeline ensures a structured approach to processing and analyzing complex microbial communities efficiently.

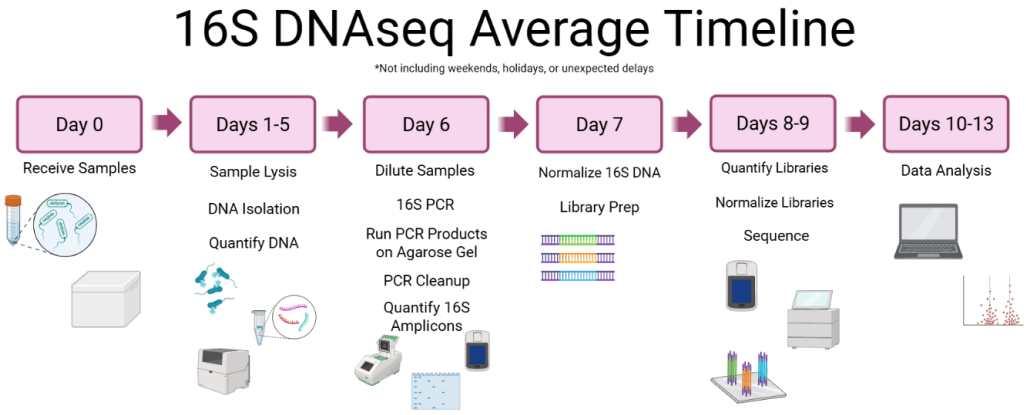

Figure 12. Average timeline from raw samples to analyzed data for 16S analysis.

6. Conclusion

Synthetic microbial communities offer transformative potential for advancing synthetic biology projects critical to the Army's mission, particularly for biomanufacturing. By genetically engineering microbial communities instead of individual organisms, the metabolic burden is distributed across multiple members, enabling the production of more complex mission-relevant materials. To fully harness this capability, 16S sequencing plays a critical role in understanding the composition and relative abundance of community members. However, because 16S sequencing provides only a snapshot in time, regular sampling and sequencing are necessary to maintain up-to-date insights into community dynamics. This iterative approach ensures the Army can effectively leverage synthetic microbial communities for innovative bioproduction solutions.

To fully realize the potential of synthetic microbial communities, it is essential to establish a robust experimental pipeline capable of handling the complexities of diverse samples. While 16S sequencing provides valuable snapshots of community composition and dynamics, the success of the workflow hinges on overcoming challenges such as nucleic acid extraction, which is often the most difficult step in the process. This project aimed to develop a framework to navigate these challenges, though success for any given sample cannot be guaranteed due to the numerous variables that influence extraction quality and yield. Optimization of nucleic acid extraction is expected when working with new communities, and manufacturers are continually improving extraction kits to better accommodate specific organisms, sample types, and communities. For complex bacterial

communities, screening multiple extraction kits is essential to identify the most effective option.

Despite the challenges posed by nucleic acid extraction, the bioinformatics pipeline developed in this project offers a versatile and modular solution that can be applied across workflows, independent of the nucleic acid extraction kit used. This adaptability ensures the pipeline remains relevant as new technologies and specialized kits emerge. Currently, the bioinformatics pipeline operates through the command line, requiring users to have basic proficiency in Unix. However, future development of a front-end user interface could simplify the workflow, making it more accessible to researchers and broadening its usability. These enhancements would establish the pipeline as a valuable tool for characterizing synthetic microbial communities.

7. Future Outlook

Competency in genetic sequencing is vital to the Army's mission, underpinning biodefense, force health protection, operational readiness, and advancements in biotechnology. Within the realm of biotechnology, sequencing supports synthetic biology efforts, aiding in the biomanufacturing of mission-critical materials. To further enhance these capabilities, the integration of automation and artificial intelligence (AI) into sequencing pipelines could streamline workflows, improve data quality, and bolster efficiency. These advancements would ensure the Army remains at the forefront of genetic sequencing innovation.

Automation provides a clear pathway for improving the efficiency and reliability of the genetic sequencing pipeline. Nucleic acid extraction can be automated using liquid handling robots, which are compatible with both magnetic bead–based and column-based extraction methods. Additionally, the increasing availability of library preparation kits designed for liquid handling platforms further streamlines the workflow. The integration of new Army resources, such as Army AI, has the potential to further increase sequencing throughput by automating and accelerating key steps of the data analysis workflow.

8. References

1. Lozano GL et al. Introducing THOR, a model microbiome for genetic dissection of community behavior. mBio. 2019;10(2):e02846-18. https://doi.org/10.1128/mBio.02846-18

2. Pascual-García A, Rivett DW, Jones ML, Bell T. Replicating community dynamics reveals how initial composition shapes the functional outcomes of bacterial communities. Nat Commun. 2025;16:3002. https://doi.org/10.1038/s41467-025-57591-2

3. Illumina. 16S and ITS rRNA sequencing; n.d. https://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/16s-rrna-sequencing.html

4. Lyu R, Qu Y, Divaris K, Wu D. Methodological considerations in longitudinal analyses of microbiome data: a comprehensive review. Genes 2024; 15(1):51. https://doi.org/10.3390/genes15010051

5. Klindworth A et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2012;41(1):e1. https://doi.org/10.1093/nar/gks808

6. Fukuda K, Ogawa M, Taniguchi H, Saito M. Molecular approaches to studying microbial communities: targeting the 16S ribosomal RNA gene. J UOEH 2016;38(3):223–232. https://doi.org/10.7888/juoeh.38.223

7. Zymo Research Corporation. 16S sequencing vs shotgun metagenomic sequencing; 2019 Nov 13. https://www.zymoresearch.com/blogs/blog/16s-sequencing-vs-shotgun-metagenomic-sequencing

8. Durazzi F et al. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Sci Rep. 2021;11:3030. https://doi.org/10.1038/s41598-021-82726-y

9. Illumina. Introduction to SBS technology; n.d. https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

10. Oxford Nanopore Technologies. How nanopore sequencing works; n.d. https://nanoporetech.com/platform/technology

11. Szoboszlay M et al. Nanopore is preferable over Illumina for 16S amplicon sequencing of the gut microbiota when species-level taxonomic classification, accurate estimation of richness, or focus on rare taxa is

required. Microorganisms. 2023;11(3):804. https://doi.org/10.3390/microorganisms11030804

12. Performing accurate species-level bacterial identification with nanopore sequencing. Oxford Nanopore Technologies. 2025. https://nanoporetech.com/resource-centre/workflow-16s-sequencing

13. Prior K et al. Accurate and reproducible whole-genome genotyping for bacterial genomic surveillance with Nanopore sequencing data. J Clin Microbiol. 2025;63(7):e00369-25. https://doi.org/10.1128/jcm.00369-25

14. Illumina. iSeq 100 sequencing system specifications; n.d. https://www.illumina.com/systems/sequencing-platforms/iseq/specifications.html

15. Illumina. Plotting %Occupied vs %PF to optimize loading for the NovaSeq 6000 and X, MiSeq i100, and iSeq 100; n.d. https://knowledge.illumina.com/instrumentation/general/instrumentation-general-troubleshooting-list/000002308

16. Optimizing cluster density on Illumina sequencing systems. Illumina; c2016. https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf

17. Bolyen E et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–857. https://doi.org/10.1038/s41587-019-0209-9

18. Illumina. Nucleotide diversity; n.d. https://support-docs.illumina.com/SHARE/ClusterOptimize/Content/SHARE/ClusterOptimize/NucleotideDiversity.htm

List of Symbols, Abbreviations, and Acronyms

AI	Artificial Intelligence
ARL	Army Research Laboratory
bp	Base Pair
Caltech SynCom	California Institute of Technology Synthetic Community
DADA2	Divisive Amplicon Denoising Algorithm 2
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
DoW	Department of War
GC	Guanine-Cytosine
ID	Identification
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
ONT	Oxford Nanopore Technologies
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
QIIME2	Quantitative Insights into Microbial Ecology 2
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
SAV	Sequencing Analysis Viewer
THOR	The Hitchhikers of the Rhizosphere

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CB TECH LIB