Identifying and Quantifying Novel Bacteria

Identifying and Quantifying Novel Bacteria

A thesis presented by

Nkaziewoh N. Nchinda-Pungong

to

the Faculty of the

Harvard John A. Paulson School of Engineering and Applied Sciences

in partial fulfillment of the requirements for

the Bachelor of Arts degree with honors in

Biomedical Engineering

Faculty Adviser: Prof. Curtis Huttenhower

Harvard University

Cambridge, MA

March 26, 2021

Honor Code

In submitting this thesis to the Harvard John A. Paulson School of Engineering and Applied Sciences in partial fulfillment of the requirements for the degree with honors of Bachelor of Arts, I affirm my awareness of the standards of the Harvard College Honor Code.

Name: Nkaziewoh Nchinda-Pungong

Signature: 

The Harvard College Honor Code Members of the Harvard College community commit themselves to producing academic work of integrity – that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgment of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.

Acknowledgments

First and foremost, I have to thank my mentor and principal investigator, Dr. Nate Cira. This thesis spans nearly three years of work in his lab, through which I learned almost everything that I know about microbes and research. I entered his lab with little prior research experience but leave planning to pursue a research career for the rest of my life. Without his dedicated interest and involvement, both in my personal and academic development, this thesis never would have been possible.

I would also like to thank Dr. Huttenhower for his willingness to formally advise my thesis. His lab's prior experience developing tools for microbial bioinformatics proved invaluable as I began developing a bioinformatics pipeline of my own.

I would like to extend sincere thanks to the members of the Cira Lab who watched me take my first steps into research, were unafraid to push me to keep clarifying ideas, and were beams of light in the darkest hours of research: Adarsh Singh, Arnold Chen, Dieter Baumgartner, Jonathan Albo, Mohsin Qazi, Samira Shiri, Shayandev Sinha, and Shenghao Tan.

A special thanks goes to Adarsh Singh, who frequently stepped in with suggestions when my code failed or I was unsure where in the wide world of bioinformatics to identify a package that performed the singular task I needed. I would also like to thank him for performing the whole genome assembly and analysis that made the final Gracilibacteria discovery of my thesis possible.

I would like to thank the faculty and staff at the Rowland Institute at Harvard for supporting me through my first presentation and introducing me to fields of research that I never knew existed. I am especially grateful to Mike Burns for allowing me to sample his car and enthusiastically following my project every step of the way. If I every find a space bacterium or silicon-based life, you will be the first to know.

I would like to express endless gratitude to Dr. Jessica "Jessie" Wilks and Tony D. Jones who first introduced me to the worlds of research, bacteria, and DNA sequencing. Four years after MITES, a plush Staphylococcus aureus still sits in my bedroom.

I would like to thank the staff at the Molecular Biology Core Facilities at Dana-Farber Cancer Institute for their sequencing services.

I would like to thank the sources that funded my work across school years and summers: the Harvard College Research Program (Summer 2018), the Program for Research in Science and Engineering (Summer 2020), and the Rowland Institute at Harvard (Summer 2018; Fall 2018 - Spring 2021 Terms).

To Linsey Moyer, my BME advisor, thank you for introducing me to engineering, supporting me on every venture across four years, and welcoming my teaching aspirations with open arms.

Lastly, I am deeply grateful to my family and friends for their continuous encouragement through research highs, lows, summers, and school years. Their unwavering faith made this work possible.

List of Contributions

This research project was jointly conceived and designed by Nate Cira and Nkazi Nchinda. Hawaii soil samples were collected by Nate Cira, while other samples were jointly collected by Nate Cira and Nkazi Nchinda. Whole-genome Gracilibacteria assembly and V4 extraction from SILVA using mothur and V-Xtractor were completed by Adarsh Singh. All other portions of the wet-lab experiments and programming were performed by Nkazi Nchinda under the supervision of Nate Cira. Data and results were interpreted by Nkazi Nchinda with assistance and guidance from Nate Cira.

Abstract

The tree of life lies at the heart of biology, but major gaps persist among bacteria. Attempts to identify these missing microbes face challenges in determining which organisms are poorly characterized and where to find them. Here, we have devised a bioinformatics-based pipeline for identifying novel organisms and assessing their relative abundance in different environments based on 16S sequences. Using data from GTDB, we validate that the 16S V4 region can be used to estimate the novelty of an organism's whole genome. Then, we apply the pipeline to 16S SILVA data, estimating how many organisms remain to be discovered at each taxonomic level. We also determine that V4 sequencing is likely to underestimate genome novelty relative to the full 16S. Next, we apply the pipeline to datasets from the Earth Microbiome Project, assessing the relative abundance of novel organisms in different environments. Our results indicate that soil samples contain the highest volume of novel bacteria, but the optimal environment for microbial discovery varies based on the desired taxonomic level of novel organisms and laboratory sequencing capacity. We then apply the pipeline to standardized samples collected from several environments, determining that salt marsh soil contains a high density of novel organisms. Lastly, we use the pipeline to enrich one marsh sample for novel organisms, assembling a novel Gracilibacteria genome in the process. This pipeline allows researchers to compare environments for microbial sequencing and enrich for novel organisms, speeding up the rate at which we discover novel bacteria.

Table of Contents

Acknowledgements 3
List of Contributions 4
Abstract 5
List of Tables/Figures 9
Glossary of Key Terms and Abbreviations 10
Chapter 1: Introduction 11
The Tree of Life 11
Defining Characterization 12
Level 1: 16S rRNA Sequencing 12
Level 2: Whole Genome Sequencing 15
Level 3: Culturing 17
The Challenge of Recovering Whole Genomes 18
A Specialized Approach 19
Chapter 2: Materials and Methods 23
Pipeline Overview 23
Data Sourcing 24
SILVA 24
Earth Microbiome Project 25
Collected Samples 25
Sequencing Adapter Removal and Demultiplexing 25
Forming Amplicon Sequence Variants 26
Sequence-Based ASV Filters 26
Sample-Based ASV Filter 27
Taxonomy Comparisons 27
Genome Taxonomy Database (GTDB) 27
SILVA 28
Nucleotide and RefSeq 28
Taxonomy-Based ASV Filters 28
Diversity Metrics 29
V4 Analysis 29
Calculating Number of Organisms Discovered in a Sample 30
Forming Clusters from ASVs 31
Pipeline Summary 32

Predicting Whole-Genome Novelty from the V4 using GTDB 32
Chapter 3: GTDB Results 34
Chapter 4: SILVA Results 36
Chapter 5: Earth Microbiome Project Results 38
Determining How Using the V4 Subunit Impacts ANI Matches 38
Determining How Varying V4 Read Lengths Impact ANI Matches 41
Evaluating Samples from Different Environments 42
Comparing Samples from Different EMP Environments 46
Chapter 6: Results from Collected Samples 52
Comparing Environmental Samples with Standardized Processing 52
Mini-Metagenomic Experiment 56
Locating a Gracilibacteria 56
Chapter 7: Discussion 58
GTDB 58
SILVA 59
Earth Microbiome Protocol 61
The Impact of the V4 Subunit Relative to the Full 16S 61
The Impact of Varying Read Lengths on ANI Matches 62
A Basis for Evaluating Samples from Different Environments 62
Comparing Different Environments 63
Collected Samples 66
Comparing Standardized Samples from Environments 66
Mini-Metagenomic Experiment 67
Conclusion 68
Chapter 8: References 69
Chapter 9: Appendix 74
Pipeline Code 74
SILVA Accession Numbers of Removed Entries 74
EMP Sample Information 75
EMP Environment Gallery 76
In Lab Materials and Methods 96
Data Sourcing 96
16S Library Preparation and Sequencing of All Sites 97
Mini-Metagenomic Preparation of Marsh 5 97

Whole Genome Library Preparation 98
Data Analysis 98
DCR Permit 99

List of Tables/Figures

Figure 1-1 | A depiction of how chimeric sequences form across two rounds of PCR.. 14
Figure 1-2 | Linear and log-scaled rank abundance plots depicting the sequencing data skew associated with metagenomic samples..... 16
Figure 1-3 | A depiction of how environmental samples are compared against GTDB. 20
Figure 1-4 | A schematic of the tree of life indicating the number of GTDB bacteria at each taxonomic level [43]. 21
Figure 2-1 | A visual overview of the bioinformatic workflow following sequencing..... 23
Figure 2-2 | An illustration of centroid-based clustering at an ANI threshold of 80%..... 31
Figure 2-3 | A detailed summary of the bioinformatic workflow following DNA sequencing.... 32
Figure 3-1 | Violin plots of the V4 ANI matches to GTDB corresponding to various taxonomic levels. 34
Figure 3-2 | Probability of matches to GTDB at a given ANI corresponding to each taxonomic level..... 35
Figure 4-1 | Results of comparing SILVA ASVs and Clusters to GTDB. 36
Figure 4-2 | A sample case demonstrating how ASVs may overestimate the abundance of novel microbes..... 37
Figure 5-1 | Comparison of the Average Nucleotide Identity of GTDB matches from full-length 16S sequences (~1400bp) and V4 sequences extracted by V-Xtractor (~85bp) and mothur (~253bp)..... 39
Figure 5-2 | Comparison of the Average Nucleotide Identity of GTDB matches from various-length components of the 16S..... 40
Figure 5-3 | Comparison of the Average Nucleotide Identity of GTDB matches from mothur and V-Xtractor. 41
Figure 5-4 | A demonstration of why the difficulty of recovering genomes can vary for samples with the same number of novel ASVs. 42
Figure 5-5 | Summary data from a single sample of soil from an Alaskan tundra ecosystem..... 44
Figure 5-6 | Summary data from a sample of slime from Catostomid fish in Colorado. 45
Figure 5-7 | The cumulative number of novel clusters predicted to be discovered at or below each ANI for individual samples at depths of 100, 1000, and 10000. 47
Figure 5-8 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for each environment (averaged from individual samples) at varying depths. 49
Figure 5-9 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for environment types (averaged from samples) at varying depths. 51
Figure 6-1 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for individual samples at varying depths. 53
Figure 6-2 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for environment types (averaged from samples) at varying depths.. 54
Figure 6-3 | A comparison of the cumulative number of novel clusters predicted to be discovered at each ANI or below for marsh samples relative to other samples at varying depths. 55
Figure 6-4 | Fraction of reads corresponding to a Gracilibacteria and other organisms from a sequencing run of a salt marsh sample. 57

Glossary of Key Terms and Abbreviations

Term	Abbreviation	Definition
16S ribosomal RNA	16S rRNA	A gene that can be used for inferring phylogenetic relationships between prokaryotes
Assembled Sequence Variant	ASV	A consensus DNA sequence derived from metagenomics that is expected to correspond to a particular microbe
Average Nucleotide Identity	ANI	The fraction of DNA bases that two organisms have in common following alignment (100 indicates that the genomes are identical at the regions being compared)
Earth Microbiome Project	EMP	A repository of 16S V4 sequences (with standardized preparation) submitted by researchers from across the globe
Genome Taxonomy Database	GTDB	A database that contains 16S sequences and whole genome sequences for organisms whose whole genomes have been assembled
Level 1 of Characterization		An organism whose 16S ribosomal RNA gene has been sequenced
Level 2 of Characterization		An organism whose whole genome has been sequenced
Level 3 of Characterization		An organism that has been cultured
Level of Novelty		The taxonomic level corresponding to a novel organism based on GTDB (e.g. phylum)
Novel		An organism that is currently at Level 1 of characterization but has not yet moved to Level 2
Operational Taxonomic Units	OTU	A cluster of closely related DNA sequences derived from metagenomics that are expected to correspond to a particular microbe
SILVA		A database of 16S/18S ribosomal RNA sequences
V4		One of 9 variable subunits of the 16S that can be used to compare microbes

Chapter 1: Introduction

The Tree of Life

At the heart of biology, the tree of life is a fundamental organizing structure [1]. Branching from the domains of Archaea, Bacteria, and Eukarya down to individual species, the tree relates together all known living organisms. However, the full scale of the tree remains unknown. Much of this uncertainty is due to prokaryotes, with estimates for the number of microbial species on Earth ranging from hundreds of thousands [2] to millions [3] to a trillion [4]. With a few recent exceptions [5], portrayals of the tree of life have tended to focus on evolutionary relationships near the tree's root [6] or well-classified organisms, particularly eukaryotes [7], rather than these prokaryotic gaps.

The lingering uncertainty about prokaryotes allows the tree of life to grow in leaps and bounds in response to new findings. For instance, in 2015, researchers identified over 35 new bacterial phyla from a single aquifer in Colorado, a group that may comprise over 15% of the bacterial domain [8]. These organisms present a drastically expanded view of the tree of life and comprise much of the current diversity on Earth [9]. With the acquisition of new whole genome sequences, the tree of life continues to expand rapidly [8], [10] with implications across multiple fields.

Three particularly relevant fields pertain to evolution, enzymes, and ecology. In the field of evolutionary history, filling in gaps in the tree of life with whole genome sequences helps to resolve the evolutionary origins of currently existing life. For example, in 2019, researchers identified a novel phylum of bacteria known as Thorarcheota, the closest known prokaryotic relative to modern-day eukaryotes [11], which further supports the evolutionary theory that eukaryotes developed from an archaeal cell engulfing a bacteria [12]. Researchers interested in

identifying enzymes with useful properties, such as Taq polymerase or CRISPR-Cas9, also greatly benefit from the acquisition of novel genomes. The discovery of novel genomes permits the discovery of the novel proteins contained within. Novel genomes also help to further ecological research on biogeochemical cycling and nutrient exchange in environments ranging from the oceans [13] to forests [14] to the human gut microbiome [15]. Whether interested in evolution, enzymes, or ecology, collecting new genome sequences can help researchers to develop new insights.

Defining Characterization

Constructing phylogenetic trees relies on the "characterization" of novel microbes, but there are multiple ways to characterize microbes. Three common levels of characterization include sequencing the portion of an organism's genome containing the 16S ribosomal RNA (rRNA) [16], sequencing of its whole genome [17], or evaluating the phenotype of a cultured isolate [18]. Throughout the text, these levels are referred to in shorthand as Level 1, Level 2, and Level 3, respectively. Modern efforts tend to focus on sequencing-based approaches [19], which are described below.

Level 1: 16S rRNA Sequencing

First championed by Woese in 1977 [20], the 16S is a subunit of the ribosomal RNA gene that can be used for inferring phylogenetic relationships between prokaryotes. The gene has both conserved regions—which can be used for primer targeting—and variable regions that evolve at different rates, allowing for comparison of prokaryotes at taxonomic levels ranging from domain to species [21], [22]. Woese was able to use the 16S region to distinguish Bacteria from Archaea [20], and in the time since, it has become the most sequenced taxonomic marker [23]. The Human Microbiome Project [24] was based on this region, and there are massive curated databases, such

as Greengenes [25] and SILVA [26], that store full-length 16S sequences from microbes sequenced by researchers across the globe.

In addition to the entire 16S, researchers also commonly sequence the V4 subunit of the 16S, a short variable region that can be used to compare bacteria at high taxonomic levels [27]. Modern high-throughput sequencing technologies have a maximum read length, limiting their ability to read the entire 16S at once. Thus, most studies rely on partial 16S rRNA sequences [27]. Though the V4 subunit contains less information than the full 16S, the length is more accommodating for modern sequencers. After collecting V4 reads, it is common to upload data to a shared database such as the Earth Microbiome Project [28].

Using the 16S and its subunits, researchers have found some success in resolving gaps in the tree of life. Clusters of similar sequencing reads can be grouped into Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs) that presumably originate from a single organism [29]. The majority of bacterial phyla have been discovered via these 16S-based efforts [30], [31]. However, these efforts face some limitations. Researchers analyzing sequences from the SILVA and Integrated Microbial Genomes databases noted that while 95% of full-length bacterial 16S sequences belong to an OTU that has been observed more than once [32], this categorization only describes approximately 30% of known bacterial OTUs [32]. In other words, the vast majority of full-length bacterial 16S sequences belong to a small fraction of organisms that have been repeatedly sampled. Though this leaves 70% of OTUs that have only been sequenced to a limited extent, researchers have noted that the rate of discovery of new OTUs seems to have slowed and a small number of studies are responsible for most new sequences [32]. These findings present what initially appears to be a saturation paradox. The rate of discovery for new OTUs seems to have plateaued, implying that most microbes have been discovered. However, the

majority of OTUs have only been observed once, implying that much more Level 1 characterization is required.

This paradox can partially be attributed to bias in sample selection and preparation. Though soil and aquatic samples are known to boast the most microbial biodiversity [33], these respectively make up only about 8% and 17% of full-length 16S sequences, with 55% of sequences coming from less-diverse host-associated environments [32]. Additionally, common primer sets used for 16S amplification have been shown to be biased in favor of certain organisms [34], and some prokaryotes have unusual features, such as introns, that prevent primers from binding [8]. A dedicated study found that at least 10% of organisms are missed by 16S primer sets [35]. In short, bias in 16S-based sequencing efforts limits their ability to fill gaps in the tree of life.

16S-based efforts also face a massive challenge from chimeric sequences [36]. As depicted in Figure 1-1, chimeras are a fusion of DNA from different organisms formed when a partially-extended piece of DNA reanneals to a foreign strand and is replicated during the PCR process. The resulting fusion strand may be misidentified as a novel microbe, artificially increasing the apparent diversity of the sample [36].

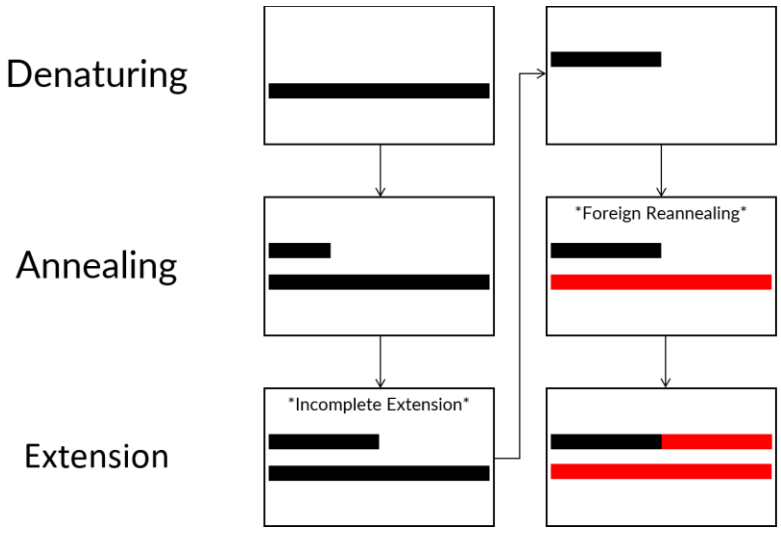


Figure 1-1 | A depiction of how chimeric sequences form across two rounds of PCR. Incomplete extension forms a short strand (black) that can reprime onto a different DNA strand (red).

These sequences have been shown to accumulate in public 16S databases [37], with conservative estimates suggesting that at *least* 5% of database records are chimeric or have substantial errors [38]. Unfortunately, chimeric errors can easily appear as highly distinct sequences, providing a challenge for estimates of microbial diversity and constructing accurate phylogenetic trees based on 16S sequences.

While algorithmic methods and a requirement for each 16S sequence to appear in more than one independent sample can help to mitigate chimeras, at present there are no viable strategies to completely prevent or detect chimeras [36]. Nevertheless, the 16S and its variable subunits remain popular as short, informative, inexpensive, and easy to sequence markers, with high conservation and per-base information content relative to other regions in prokaryotic genomes.

Level 2: Whole Genome Sequencing

While sequencing the 16S rRNA gene is a cost-effective way to analyze microbial environments, sequencing of the whole genome remains the most accurate method for identifying prokaryotes [39]. There are over 20 competing definitions of "eukaryote," and prokaryotic taxonomy is just as contentious [40]. Nevertheless, one common strategy for assigning taxonomy is to compare organisms using Average Nucleotide Identity (ANI)—the fraction of DNA bases that two organisms have in common [2]. For instance, an ANI of 100 indicates that 100% of bases match between organisms at the regions being compared and the sequences are identical. By aligning and comparing similar reference genes between organisms, ANI can be used to construct a taxonomic tree where organisms that are more closely related are assumed to share a higher ANI [41]. This ANI approach can even be applied to 16S sequences, though shorter reads and chimeras mean less taxonomic accuracy.

To obtain whole genomes, researchers sequence random short sections of the genome and computationally reassemble them to form full-length bacterial genomes [42]. Then, an organism's genome can be used to assign taxonomy—through a method such as the ANI of reference genes—or to explore questions involving regions outside of the 16S. These computational "metagenomic" approaches can leverage initial sample DNA from a bacterial isolate, a mixed environmental sample, or even an amplified single cell. Many of the resulting whole genome assemblies are available in public databases such as the Genome Taxonomy Database (GTDB) [43].

Unfortunately, these computational approaches face the issue of skewed species distributions. As described in early community ecology literature [44], the number of individuals of common environmental species can outnumber the number of individuals of rare species by several orders of magnitude. As shown in Figure 1-2, microbial communities are no exception, reflecting this same skew with most sequencing data mapping to a handful of common organisms.

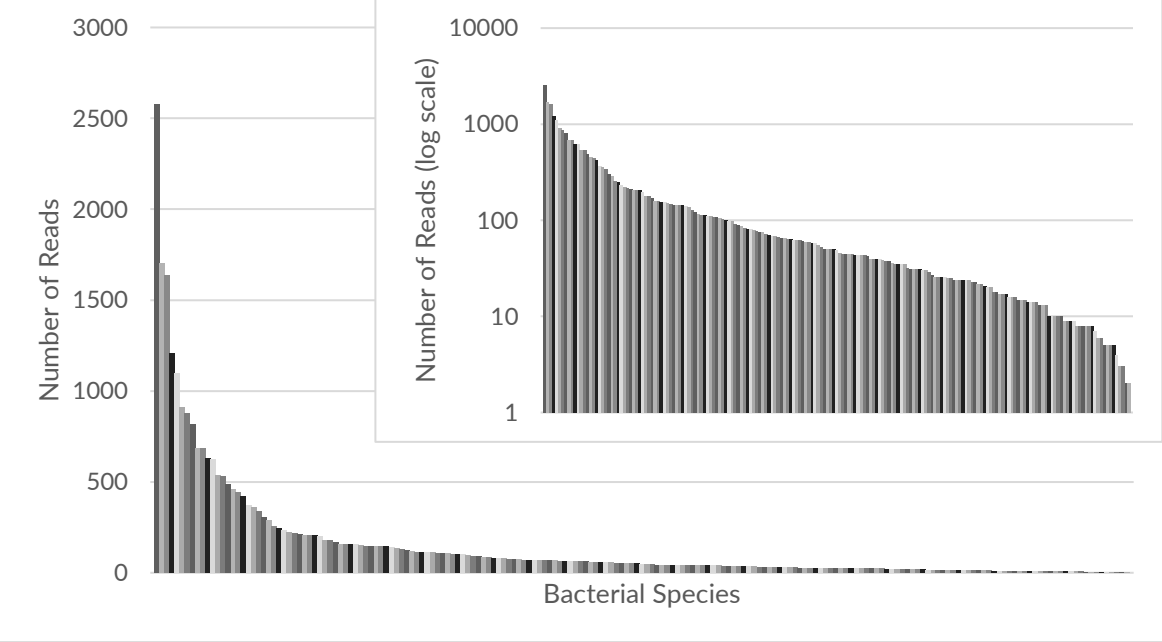


Figure 1-2 | Linear and log-scaled rank abundance plots depicting the sequencing data skew associated with metagenomic samples. Each bar represents a different bacterial species identified in a salt marsh sample. The distribution heavily favors a few organisms, with the 14 most common bacterial species making up >50% of the total reads, while the rarest 14 organisms make up 0.1% of the total reads.

Since most species are rare in typical environmental communities, many interesting novel organisms are likely to appear at low frequencies, making recovery of their genomes a costly process [45] and making them difficult to distinguish from background noise [46]. Traditional metagenomic approaches can reconstruct the genomes of novel organisms from such mixed communities, but the cost is high and the efficacy is low.

Multiple groups have developed single-cell approaches to whole genome sequencing that minimize the impact of environmentally skewed distributions on genome recovery [10], [47]. These microfluidic or "mini-metagenomic" approaches divide environmental samples into many subsamples containing one or a few bacterial cells each. These samples are processed and sequenced in parallel, rather than as a bulk group. Since each subsample is much less diverse than the overall sample, there is a decreased likelihood of accidentally co-assembling DNA from different starting organisms. This strategy means that high-quality genome reconstruction is even feasible for rare organisms.

Overall, while some large-scale efforts have been made to sequence whole genomes [31], there are far fewer whole genome assemblies than 16S sequences.

Level 3: Culturing

The culturing process classically involves growing bacteria on agar plates and isolating bacterial colonies. These living colonies can easily be sequenced or studied through techniques such as microscopy, Gram staining, and functional assays. This method is considered to be a gold standard since it provides large quantities of cells from a clonal population [19]. The clonal sample is particularly useful for microbe characterization since there is a surplus of DNA for sequencing. However, the vast majority of bacteria cannot yet be grown in a laboratory setting [48] due to a variety of potential reasons: inappropriate growth conditions [49], an excessively slow growth rate

[50], oxidative stress generated by the laboratory environment [51], missing pathways [52], stochastic awakening or acclimation [53], or the absence of necessary compounds produced by other members of the community [54]. As a result, over 88% of all cultured microbial isolates belong to just four phyla [10], and the majority of what is known about bacterial physiology stems from a small subset of easily-culturable, highly studied, medically-relevant organisms [55]. Due to this cultivation bias, genome sequencing efforts that are based on culturing reflect this same bias toward a few organisms [10]. In short, with culturing, we cannot find what we cannot grow. For this reason, most modern characterization efforts focus on Levels 1 and 2 [19].

The Challenge of Recovering Whole Genomes

Skewed species distributions associated with Level 2 mean that the blind spots remaining on the tree of life cannot merely be attributed to a lack of effort. Despite the aid of high-throughput DNA sequencing, recovering whole genomes from novel bacteria has remained a persistent challenge. There are many environments from which novel genomes could likely be recovered—such as saltwater, mines, and mosquitos [56]—but it would be useful to have a tool to prioritize these environments based on how many novel whole genomes they are expected to contain. This tool could be used to increase the efficiency of workflows such as culturing, metagenomics, or single-cell mini-metagenomics to maximize the likelihood that the genomes recovered belong to *novel* organisms for which we do not yet have whole genome assemblies.

A Specialized Approach

Researchers have used multiple levels of characterization for identifying novel microbes, but they still face two major limitations. First, the starting environmental samples play a major role in which bacteria can be discovered, but there remains a high degree of uncertainty about where to look to discover highly novel genomes that have not already been assembled. Second, even after collecting a sample, it remains difficult to determine which bacteria are both novel and *real*, rather than merely a processing artifact. Several fields of biology would greatly benefit from a tool that clarifies these questions and makes it easier to locate novel organisms.

The primary aim of this thesis was to develop a bioinformatics-based pipeline for identifying novel bacteria and ranking environmental sampling sites. In particular, we aimed to create a tool ("the pipeline") to examine 16S sequences (Level 1 of characterization), determine which DNA sequences belong to organisms that have been poorly characterized, and make corresponding recommendations about where to collect environmental samples for whole genome sequencing (Level 2) or culturing (Level 3). Rather than using the full 16S region, which is too long for high-throughput sequencing, we primarily rely on the V4 subunit as described earlier.

We rely on the Genome Taxonomy Database (GTDB)—a repository of whole genome assemblies—to determine which DNA sequences belong to poorly characterized microbes. In addition to whole genomes, the database contains the extracted 16S sequences for organisms whose whole genomes have been assembled. As depicted in Figure 1-3, by querying 16S sequences from a sample against the 16S sequences in GTDB, the pipeline can determine which microbes in the query sample do not match the 16S sequences of any known sequenced genome and are thus likely to be novel at some taxonomic level.

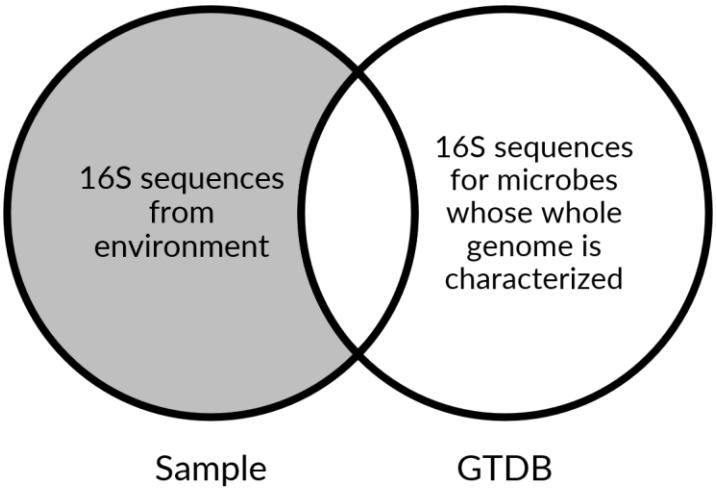


Figure 1-3 | A depiction of how environmental samples are compared against GTDB. Microbes in the shaded region match poorly to the database, indicating that the 16S sequence is known but the whole genome has not been extensively characterized.

The next step is to calculate the Average Nucleotide Identity (ANI) between the sample 16S and the 16S from the closest known sequenced genome to generate a quantitative estimate for the expected degree of novelty. As shown in Figure 1-4, organisms can be novel at a variety of taxonomic levels, such as a novel phylum or class of organisms. Rather than referencing these taxonomic levels, this thesis will primarily use ANI to describe the novelty of organisms in each sample for a variety of reasons. Primarily, the taxonomic levels are artificial constructs [57], so their placement is somewhat subjective. Furthermore, horizontal gene transfer among microbes complicates attempts at nomenclature [58]. For reference, some researchers have proposed ANI thresholds based on full-length 16S rRNA sequences of 75.0% for phylum, 78.5% for class, 82.0% for order, 86.5% for family, and 94.5% for genus [2]. However, since this thesis will primarily use ANI, this can be summarized as "a lower ANI indicates a higher degree of novelty."

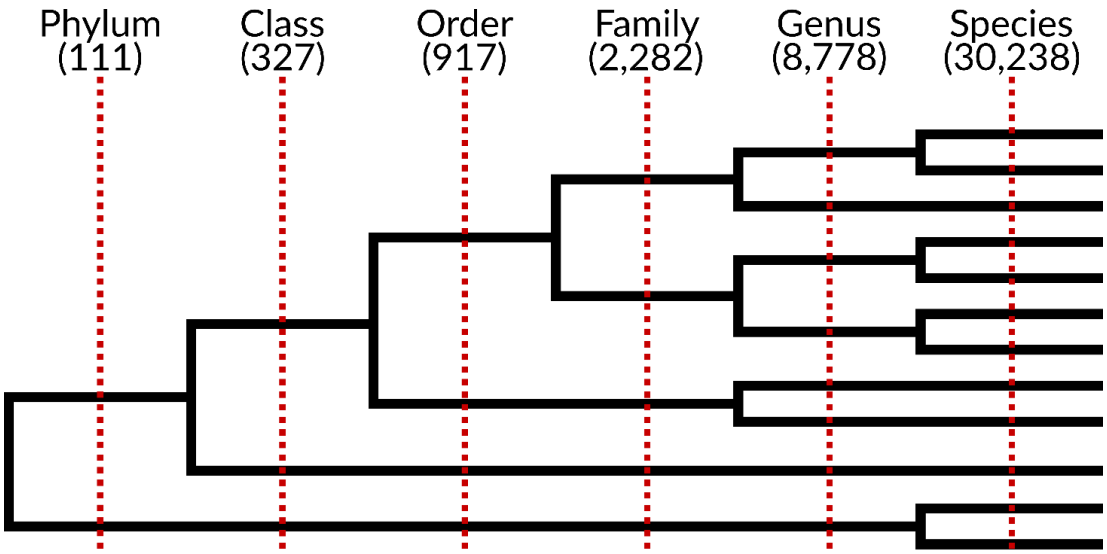


Figure 1-4 | A schematic of the tree of life indicating the number of GTDB bacteria at each taxonomic level [43]**.**

Following comparison to GTDB, samples with many novel 16S sequences that are highly different from any known sequenced genomes may represent good starting places for sequencing and culturing efforts that attempt to fill in blind spots on the tree of life. By comparing the number of novel organisms in different environments, the pipeline aims to indicate which environments are promising sampling sites for the identification of novel bacteria. The pipeline can also calculate the fraction of each sample that is composed of novel bacteria to determine how easy it would be to reconstruct their genomes. This is the first approach we know of that leverages this 16S reference database strategy to prioritize and compare the potential of different environmental sampling sites to harbor highly novel microbes. This strategy should aid in moving organisms from the level of known 16S sequences to known genomes and beyond.

When attempting to identify novel bacteria, it can be rather challenging to determine whether an unusual sequence is an authentic microbial variant or merely an error arising from the steps of processing. By the very nature of searching for novel bacteria, there is not a reference point with which to compare DNA sequences to ensure that they are real. Furthermore, chimeric

sequences remain a significant concern since these are impossible to eliminate but can appear highly novel. The pipeline attempts to err on the side of caution, potentially throwing out data that may belong to real organisms to minimize the influence of chimeras and other artifacts.

After detailing the pipeline's development, this thesis leverages four different datasets. First, we used data from GTDB to determine how feasible it is to predict the degree of novelty of a whole genome solely from the V4 subunit of the 16S. Second, we applied the pipeline to the SILVA database—a 16S sequence repository—to estimate how many microbes are at Level 1 of characterization (16S) and have yet to reach Level 2 (Whole Genome Sequencing). Third, we applied the pipeline to thousands of V4 reads from other researchers' environmental samples to determine which environments contained a high proportion of poorly characterized organisms (Level 1) and could benefit from future sequencing efforts. In the process, we investigated how read length—such as using the full 16S or varying length reads of the V4—impacted the predicted novelty of a sample. Fourth, we applied the pipeline to samples we collected to determine which sites our future experiments should prioritize. In addition, we conducted a mini-metagenomic 16S rRNA sequencing experiment on one of these samples, discovering and assembling a rare and novel Gracilibacteria genome in the process.

Chapter 2: Materials and Methods

This chapter describes the materials and methods that were involved in the development of the pipeline and its application to both environmental datasets and wet-lab samples. We were interested in developing a tool that could input high-throughput, 16S V4 sequencing data from any environment, remove wet-lab and sequencing artifacts, compare the cleaned data against several reference databases, and generate summary tables and figures characterizing each environment. The full description of the pipeline begins below. Apart from DADA2, which is based in R, all code was written in Python.

Pipeline Overview

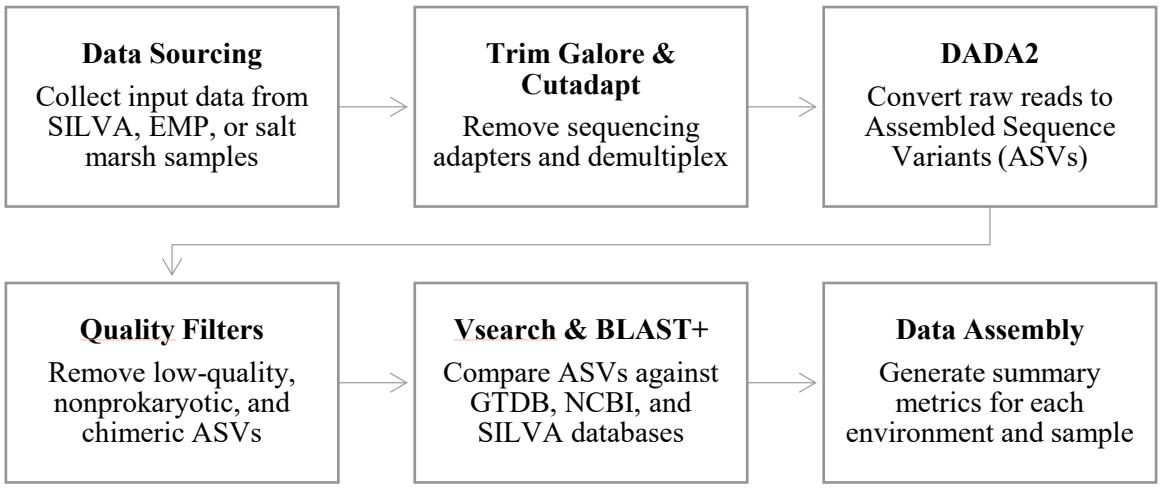


Figure 2-1 | A visual overview of the bioinformatic workflow following sequencing.

Most data sources in this thesis came directly from an Illumina MiSeq or HiSeq and were processed using all the steps above. However, SILVA data is already demultiplexed, quality-checked, and assembled into ASVs based on consensus sequences [26]. As a result, data from SILVA bypassed the demultiplexing, quality filtration, and DADA2 steps. For additional

flexibility, it is possible to enable or disable individual processing steps to meet a researcher's desired use case.

Data Sourcing

Pipeline input data was drawn from three sources: SILVA, the Earth Microbiome Project (EMP), and V4 sequencing data from samples that were collected from multiple sites.

SILVA

The Ref NR99 database of 16S/18S sequences was downloaded from SILVA Release 138.1, the latest version at the time of writing. NR99 indicates that the database was dereplicated at 99%, which was required to eliminate redundant sequences from the analysis. Using VSEARCH, the database was further dereplicated to 85%, which corresponds to families [2]. Sequences labeled as eukaryotes (18S) were removed using Biopython. 16S sequences labeled as chloroplasts were removed using Biopython to avoid erroneous eukaryotic matches. Comparison of the removed eukaryotes to the remaining database revealed that 69 remaining SILVA entries matched to removed eukaryotes with homologies of 37.4%-99%. To avoid eukaryotic contamination, these sequences were also removed. A full list of removed accessions is included in the Appendix under SILVA Accession Numbers of Removed Entries (Page 74). Using VSEARCH, the remaining sequences were compared against a database of all eukaryotic sequences that had been removed from the SILVA database in previous steps, and sequences that matched eukaryotes above 60% ANI were eliminated. (options— *--usearch_global --userfields query+target+id+alnlen+mism+gaps+qilo+qihi+tilo+tihi+trow --id 0.60 --maxhits 1*).

The remaining data was run through a "singleton filter" to help eliminate chimeras. Using a procedure based on Louca et al. [59], sequences were sorted by decreasing length and clustered

using VSEARCH (options— *–cluster_fast –centroids --id 0.75*) at a threshold of 75% ANI. Clusters with only one sequence were eliminated to minimize the risk of chimera contamination.

Earth Microbiome Project

Data from 19 studies was downloaded from the Qiita EMP repository [28], a collection of microbial samples from across the globe. Under the Earth Microbiome Protocol, samples cover the sample V4 region, use a standard primer set, and follow a standard library preparation workflow. Environmental data was chosen to include animal-associated, soil-associated, and marine samples in the analysis.

Collected Samples

Data from a new 16S mini-metagenomics experiment was included in the analysis. Most samples were collected from a variety of sites around Boston and Cambridge, including Belle Isle Marsh Reservation. Samples were also collected from the Hawaii coast. A full list of samples and the wet-lab protocol used for their preparation is included in the Appendix under In Lab Materials and Methods (Page 96).

Sequencing Adapter Removal and Demultiplexing

Trim Galore (a wrapper for Cutadapt [60] and FastQC) was used to remove sequencing adapters to prevent interference with taxonomic classification. Reads were not filtered based on length at this stage (options— *--length 0 --no_report_file --suppress_warn --cores 8*). Samples were then demultiplexed using Cutadapt. No barcode insertions or deletions were allowed, flanking N bases were trimmed, and a maximum error rate of 10% was permitted (options— *--no-indels --trim-n -e 0.1 --quiet)*. To ensure accurate adaptor removal, select reads were visually inspected with FastQC.

Forming Amplicon Sequence Variants

Based on the cleaned data, DADA2 [61] was used to create Amplicon Sequence Variants (ASVs), each of which theoretically corresponds to a unique microbe. Standard filtering parameters were used and lingering PhiX from sequencing was removed (options— *maxN=0, maxEE=2, truncQ=2, rm.phix=TRUE, compress=TRUE*). When inferring ASVs, demultiplexed samples were pooled to account for low read numbers present in some samples (options— *multithread=TRUE, pool=TRUE, selfConsist=TRUE*).

Sequence-Based ASV Filters

Kmer filtering was used to eliminate ASVs with low complexity, such as a continuous stretch of a single base. Mutations in the hypervariable V4 region should not have a positional bias, so a set of bases that repeats on a fixed interval likely suggests an artifact. The kmer word length was 10 bases, and Kmer-Counter was used to track the number of each kmer for ASVs. Using the SciPy stats module, Shannon entropy was calculated from the kmer counts, and sequences below 3.75 were eliminated. This cutoff was verified using an artificial dataset with manually-constructed low-complexity sequences. To prevent short primer or adapter sequences from entering the analysis, ASVs shorter than 65 bases were removed. EMP samples have a read length of 90-150 base pairs, which was far above this cutoff.

In case any ASVs containing Illumina adapters bypassed Cutadapt, an additional Illumina filter was implemented to discard reads containing forward (*TACGGCGACCACCGAGATCTAC*) or reverse (*CAGAAGACGGCATACGAGA*) adapter sequences.

Sample-Based ASV Filter

ASVs that did not appear in at least two different samples from the same study were eliminated. As described by Louca et al. [59], this filter was implemented to eliminate chimeric ASVs, though the apparent richness of some environments may have slightly decreased in the process.

Taxonomy Comparisons

Query samples were compared against four databases using VSEARCH or BLAST+ [62] at a minimum identity threshold of 60%. BLAST+ was used to compare query samples against the two NCBI sources: RefSeq and the Nucleotide database (options— *-task blastn -dust 'yes' -outfmt "10 qaccver saccver sskingdom stitle pident length mismatch gapopen qstart qend sstart send sseq" -perc_identity 0.60 -max_target_seqs 1*). VSEARCH was used to compare query samples against the GTDB and SILVA databases (options— *--usearch_global --userfields query+target+id+alnlen+mism+gaps+qilo+qihi+tilo+tihi+trow --id 0.60 --maxhits 1*)

Genome Taxonomy Database (GTDB)

The bac120_ssu_reps database of 16S sequences was downloaded from GTDB Release 95, the latest version at the time of writing [43]. GTDB contains both whole genome and 16S datasets for the same organisms, and 16S sequences were downloaded to enable comparison with the 16S from environmental samples. Since GTDB stores organisms whose whole genomes have been assembled, sequences that matched extremely well to a sequence in GTDB were assumed to belong to an organism that was well characterized at the whole genome level. Sequences that matched poorly to GTDB were considered to represent organisms that were less well characterized at the whole genome level, suggesting a higher level of novelty.

SILVA

The Ref database of 16S/18S sequences was downloaded from SILVA Release 138.1. To allow for the closest possible match, this database was not dereplicated to remove redundant sequences. Sequences labeled as eukaryotes (18S) were removed using Biopython to prevent erroneous eukaryotic matches. 16S sequences labeled as chloroplasts were also removed using Biopython. Chloroplasts are thought to have evolved from cyanobacteria [63], resulting in genetic similarities that complicate nomenclature and could potentially lead to eukaryotic matches.

Nucleotide and RefSeq

All sequences in the Nucleotide (nt) and RefSeq (ref_prok_rep_genomes) databases were downloaded using the update_blastdb Perl script in BLAST+. The Nucleotide database contained both prokaryotes and eukaryotes, while the RefSeq database was limited to prokaryotes. After December 2020, the local databases were no longer updated to ensure that all samples were compared in identical conditions. No additional processing was necessary for downloaded sequences. RefSeq matches can aid in manually confirming the identity of individual organisms but were not used to filter sequences.

Taxonomy-Based ASV Filters

Following BLAST+, ASVs that most closely matched a eukaryotic sequence in the Nucleotide database were discarded to prevent data contamination. After VSEARCH, ASVs that matched the SILVA database at below 60% similarity were discarded. SILVA represents an extremely thorough database of 16S sequences, and as noted by Louca et al. [59], ASVs whose closest match lies below this threshold are likely chimeric.

ASVs were additionally subjected to a “positive filter” to remove spurious or chimeric sequences. To be accepted, ASVs had to meet one of the following criteria: match a GTDB entry

with above 60% identity, match a SILVA entry with above 60% identity, or most closely align to a BLAST entry with "16S" in the title.

Diversity Metrics

For each sample, the Shannon entropy was calculated using the SciPy stats module. The remaining diversity metrics—richness, read count, unweighted mean and median ANI (ignoring read count), and weighted mean and median ANI (considering read count)—were computed in Python using read counts and ASVs from DADA2.

V4 Analysis

The Ref NR99 database was downloaded from SILVA Release 138.1 and prepared as described under Data Sourcing. The V4 region was extracted from full-length 16S sequences using two different methods: mothur and V-Xtractor [64], [65]. EMP primer sequences corresponding to the V4 were fed into mothur (*options—pcr.seqs; FWD: GTGYCAGCMGCCGCGGTAA; REV: GGACTACNVGGGTWTCTAAT*), which attempted to locate the targeted regions within the full-length sequence. This approach was meant to represent the full V4 region and flanking constant regions that would be obtained by EMP wet-lab protocols. V4 regions were also extracted from the full-length 16S using V-Xtractor, which attempts to locate the region using Hidden Markov Models. This approach was meant to represent short V4 reads (~85bp) associated with earlier versions of the EMP protocol. Full-length 16S sequences were grouped with their corresponding V4 regions extracted via mothur and V-Xtractor. Groups that did not contain V4 regions extracted via both methods were eliminated. 16S and both extracted V4 sequences were compared against the GTDB database as described in Taxonomy Comparisons (27).

Calculating Number of Organisms Discovered in a Sample

The number of organisms located in a sample was computed using a probability-based sum. In this context, a "sample" represents the environmental site that DNA was collected from for sequencing.

$$O = \sum_{n=1}^{E} 1 - \left(1 - \frac{x_n}{x_{tot}}\right)^M$$

Equation 2-1 | A probabilistic sum describing how many organisms a researcher can expect to find from sequencing based on the read counts of each organism.

In this equation, O represents the total number of ASVs found, E represents the number of different ASVs in a sample, x_n represents the number of reads corresponding to the *n'th* ASV, x_{tot} represents the total number of reads from all ASVs in the sequencing run, and M represents the theoretical maximum number of individuals that a sequencer can identify.

$\frac{x_n}{x_{tot}}$ represents the fraction of reads that belong to ASV n, which is the likelihood of selecting that ASV out of the environmental sample by random chance. Therefore, the portion of the equation within the parentheses represents the probability of *missing* an ASV after one sampling event. A sequencing machine can be used to identify up to M individual organisms, and each of these attempts represents a different sampling event. Raising the parenthetical expression to the power M provides the likelihood of missing a particular ASV after the sequencer has run its course. Subtracting this entire expression from one provides the likelihood that the sequencer has found ASV n. Summing the results across all ASVs provides the total number of ASVs a researcher can expect to find following sequencing. This same process can be repeated by substituting clusters in place of ASVs to determine the total number of clusters a researcher can expect to find.

This formulation assumes that the community is large enough that all samples are independent. It is also, of course, blind to additional ASVs that are not identified in the initial sample but could appear in the community with deeper sequencing.

In practical terms, if a sample contains *E* different ASVs in total and the sequencing machine can identify up to *M* individuals, a researcher can expect to find *O* distinct ASVs after randomly sampling members of the community.

Forming Clusters from ASVs

To avoid overestimating the number of organisms in environmental samples, similar ASVs were assembled into clusters using a VSEARCH centroid-based clustering algorithm (*options— –cluster_fast –iddef 4)*. "Centroid-based" indicates that the longest ASV was defined as the centroid of the cluster, and ASVs that matched to the centroid above a predefined ANI threshold were grouped into the same cluster. Then, the next longest ASV was defined as the centroid of a new cluster and the process repeated. This process is illustrated in Figure 2-2. At each ANI level, clusters were recreated using the new ANI threshold.

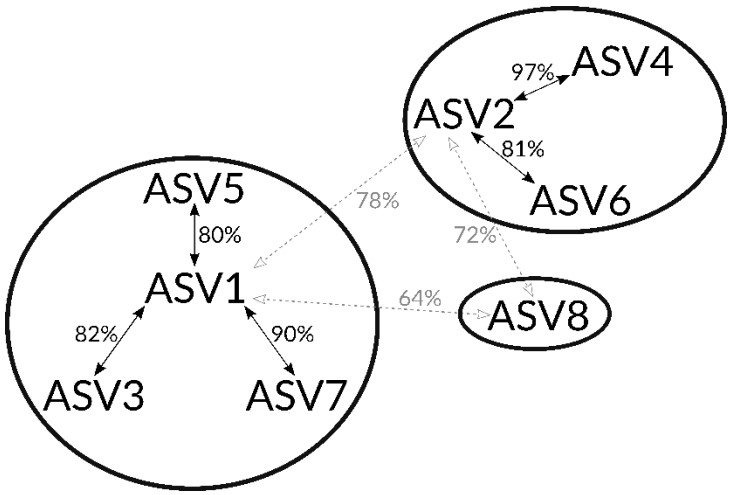


Figure 2-2 | An illustration of centroid-based clustering at an ANI threshold of 80%. ASVs that match to a centroid sequence (ASV1) above the ANI threshold are included within the cluster. ASVs that fall below the ANI threshold for one cluster (ASV4, ASV6) may align more closely to the centroid of another cluster (ASV2). ASVs that fall below the threshold for existing clusters become the centroid of a new cluster (ASV8).

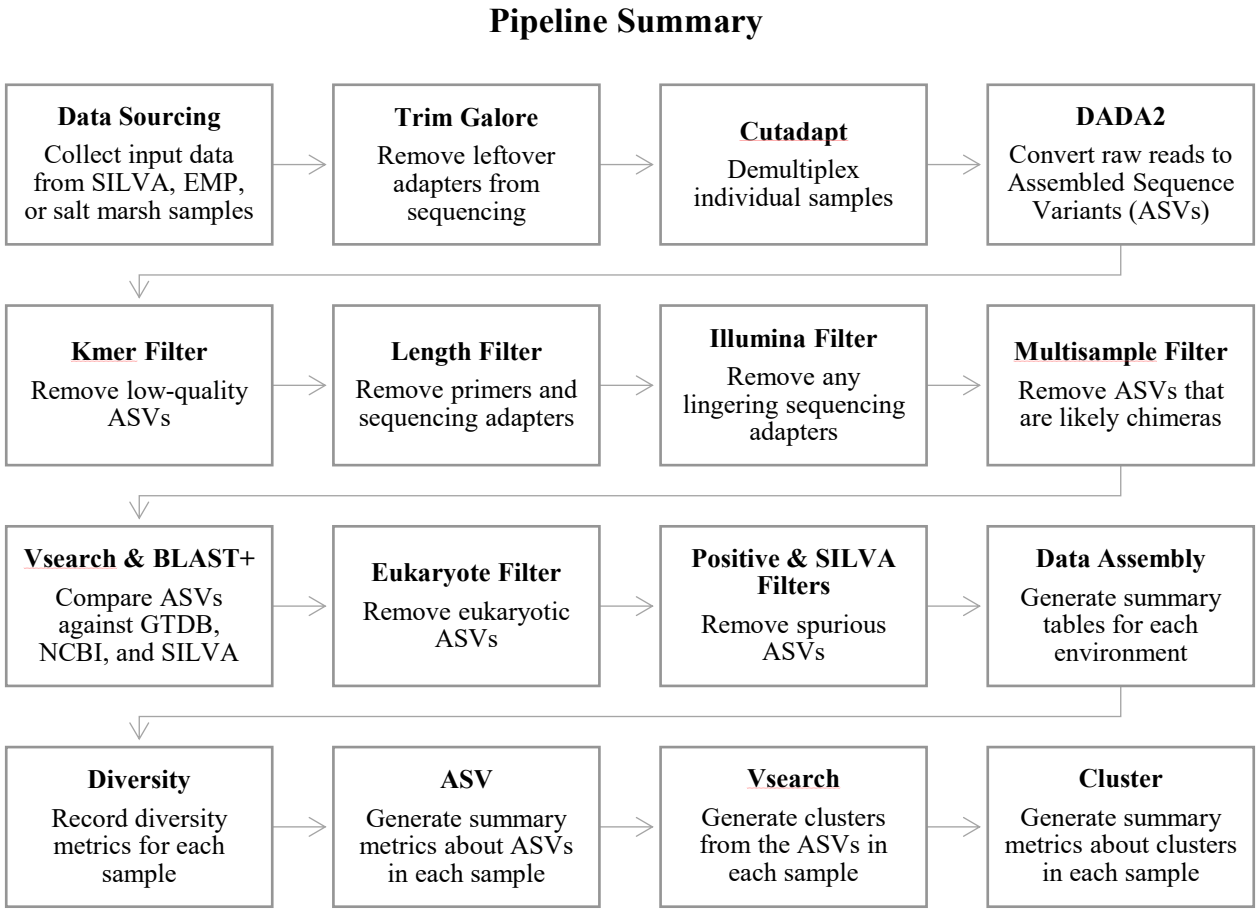


Figure 2-3 | A detailed summary of the bioinformatic workflow following DNA sequencing.

Predicting Whole-Genome Novelty from the V4 using GTDB

The bac120_ssu_reps database of 16S sequences was downloaded from GTDB Release 95. The full database contains ~195,000 bacteria and archaea whose whole genomes have been assembled and quality-checked using CheckM. EMP primer sequences corresponding to the V4 were fed into mothur [64] (*options—pcr.seqs; FWD: GTGYCAGCMGCCGCGGTAA; REV: GGACTACNVGGGTWTCTAAT*), which extracted 9,604 V4 regions with no errors in primer sequences permitted. VSEARCH [66] was used to compare V4 sequences to one another and obtain ANI values.

For each group at every taxonomic level (preassigned by GTDB based on reference genes from the whole genome), members of the group were removed from the database and the highest ANI match for each member to the remainder of the database was stored. For instance, all entries labeled within the Firmicutes phylum were removed from the database and compared to the remaining database to determine the closest ANI match. This process was repeated for every phylum. Then, it was repeated at all lower taxonomic levels.

Chapter 3: GTDB Results

The Genome Taxonomy Database (GTDB) is the most complete, phylogenetically consistent database currently available [67]. Genomes are classified from the domain to species level based on ANI between orthologous regions. We were curious how the ANI match of the V4 subunit to GTDB corresponds to the taxonomic novelty of the whole genome and investigated this question using 16S sequences from GTDB. By individually removing a group at a given taxonomic level (e.g. the phylum Firmicutes) and comparing its members to the remainder of the database, we aimed to determine the characteristic ANI match of V4 reads associated with each taxonomic level. The results are shown in Figure 3-1.

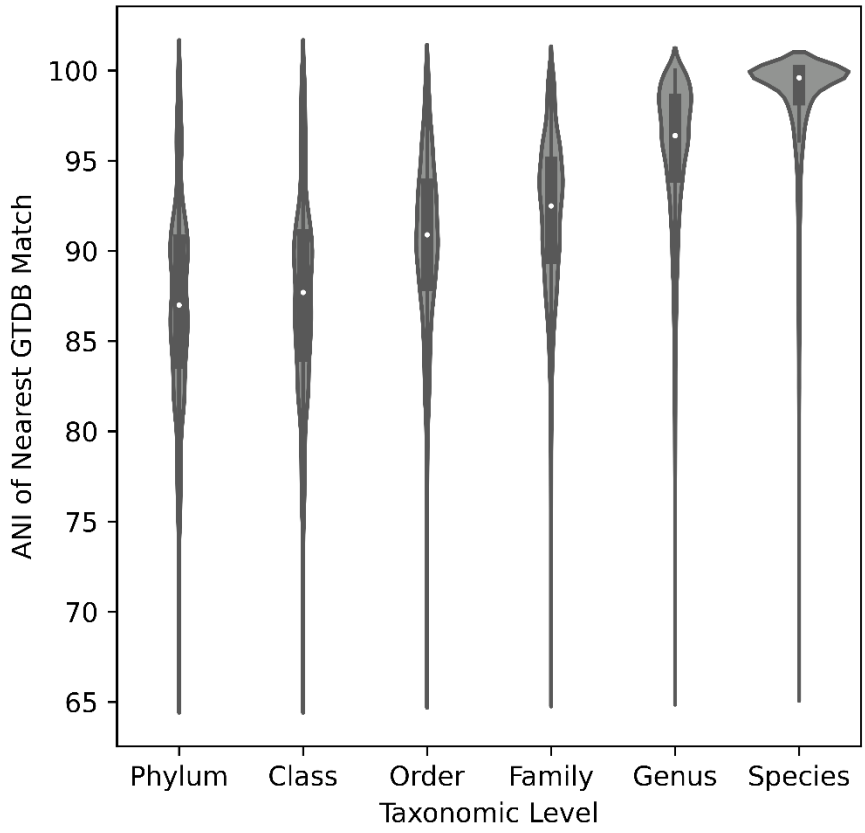


Figure 3-1 | Violin plots of the V4 ANI matches to GTDB corresponding to various taxonomic levels. The outer body of the violin plot represents the ANI distribution which surrounds a boxplot in dark grey.

When a phylum was removed from the database, its V4 reads tended to have the lowest ANI match to the remaining entries. The median ANI associated with taxonomic levels grew steadily while approaching the species level. For all taxonomic levels, there is a broad distribution with outliers that pass the median of the highest and lowest taxonomic levels.

For a given V4 ANI match to GTDB, its likelihood of corresponding to a particular taxonomic level is shown in Figure 3-2. Organisms with V4 ANI values in the range of 70-85 are likely to correspond to phylum and class with approximately equal measure. Larger ANI matches are more likely to correspond to lower taxonomic levels. At extremely low ANI values below 70, the phylogenetic classifications begin to break down. These results indicate that V4 ANI can be used to determine the novelty of the whole genome, primarily for organisms with an ANI at or above 70.

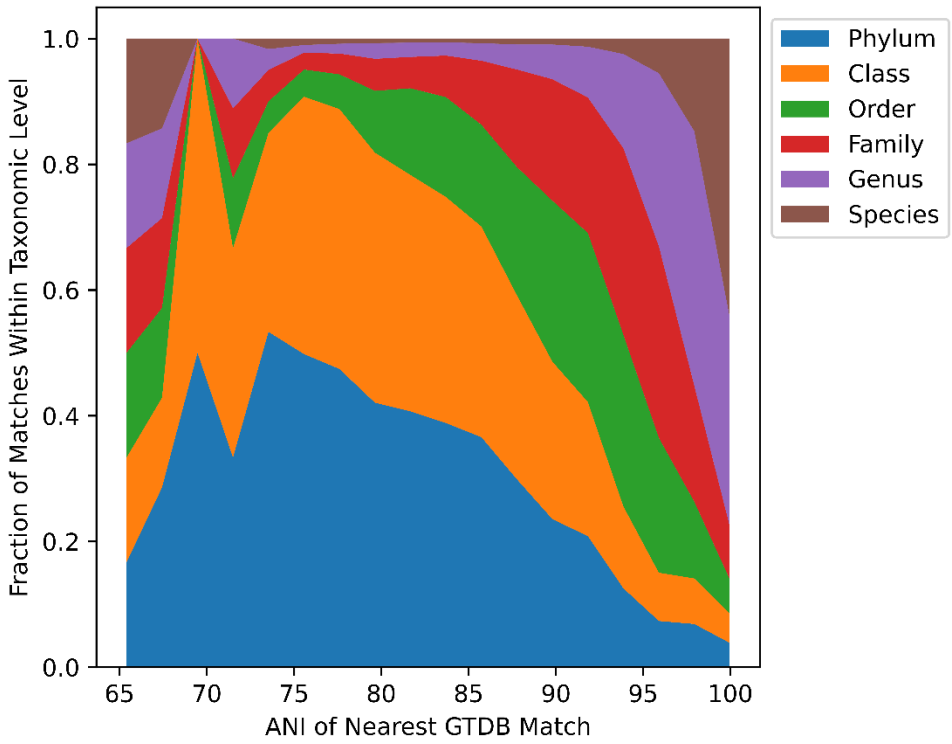


Figure 3-2 | Probability of matches to GTDB at a given ANI corresponding to each taxonomic level.

Chapter 4: SILVA Results

The pipeline relies on short reads of the V4 region to predict the level of novelty of the whole genome, and GTDB results indicated that the V4 could potentially be used for this purpose. Next, 16S sequences from SILVA (ASVs) were run through the pipeline to estimate how many organisms at each taxonomic level were at Level 1 of characterization (16S) and had yet to reach Level 2 (Whole Genome Sequencing). Organisms that match at a particular level or below can be considered as unique members at that taxonomic level. The results are shown in Figure 4-1 as a cumulative plot.

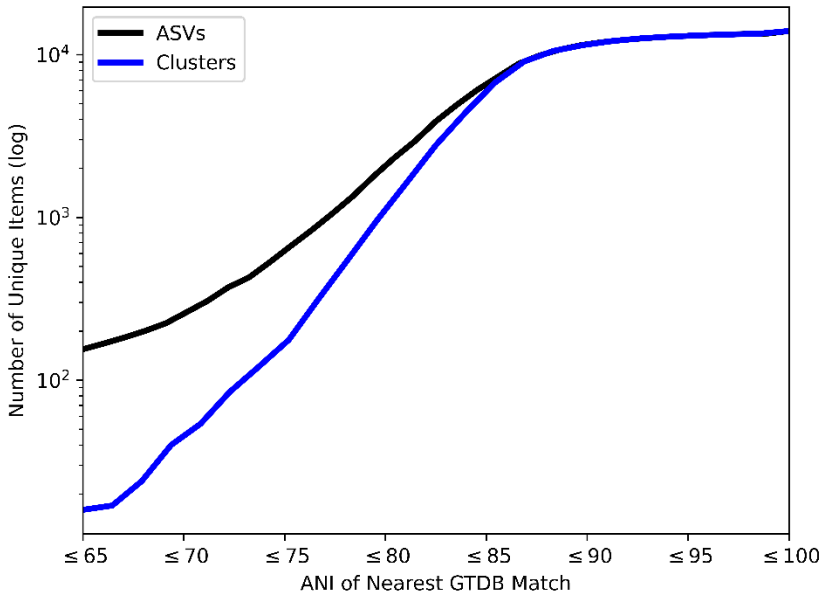


Figure 4-1 | Results of comparing SILVA ASVs and Clusters to GTDB. ANI match to the closest item in GTDB is shown along the x-axis. The plot is cumulative, so each tick indicates the number of ASVs/Clusters that match to GTDB at that level or below. Due to the large size of the SILVA database, sequences were clustered at 86.5% ANI to avoid redundancy before being compared to GTDB.

In addition to comparing SILVA sequences to GTDB directly, sequences were compared following clustering to avoid overcounting. As illustrated in the simple example in Figure 4-2, if multiple ASVs are compared to GTDB, they may all seem to be highly novel relative to anything

in the database. As shown, since each ASV matches GTDB at an ANI of 50%, the sample would appear to contain three ASVs that are novel at above the phyla level. However, since the ASVs are incredibly similar to one another, it would be more accurate to suggest that there is one new phylum with three members. In the example below, all three ASVs would be grouped into one cluster and counted as a single phylum.

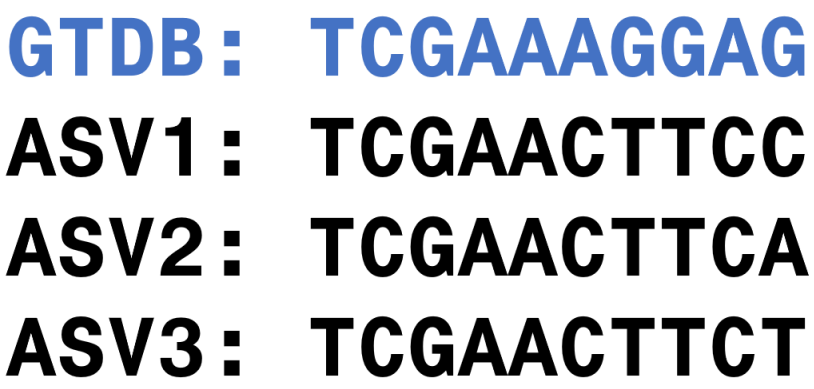


Figure 4-2 | A sample case demonstrating how ASVs may overestimate the abundance of novel microbes. Individual ASVs may be highly dissimilar from GTDB but closely resemble one another. Clustering these similar sequences together prevents overestimation.

The results of clustering are shown in Figure 4-1, and the estimated number of novel items at each taxonomic level decreases. Using Figure 3-2, it is possible to estimate the number of items corresponding to each taxonomic level based on ANI. For instance, at an ANI of 75, roughly 40% of the clusters shown on the plot are expected to be novel at the phylum level. Some caution is warranted since chimeric sequences are known to accumulate in 16S databases [37], and some may persist despite our efforts to remove them. Nonetheless, there appear to be many novel microbes at various taxonomic levels that have not yet reached Level 2 of characterization.

Chapter 5: Earth Microbiome Project Results

SILVA results provided an estimation of how many organisms at each taxonomic level were characterized at Level 1 without corresponding whole genome sequences. Next, we attempted to determine which Earth Microbiome Project environments contained the highest proportion of novel clusters and could potentially yield the most whole genomes with dedicated sequencing efforts. As with SILVA, the possibility of chimeric sequences within a sample hinders attempts to predict absolute numbers of how many novel microbes a sample will contain. However, since chimeras are not expected to form at drastically different rates between samples, relative comparisons between samples should be minimally affected.

Determining How Using the V4 Subunit Impacts ANI Matches

While SILVA contains full-length 16S sequences (~1400 base pairs), the EMP contains much shorter reads of the V4 region. To assess how using this subunit would impact the projected novelty of samples, the V4 was extracted from SILVA sequences via two different methods: V-Xtractor and mothur. V-Xtractor attempted to identify the V4 region using Hidden Markov Models (~85 base pairs) while mothur extraction was based on EMP primer sequences (~253 base pairs). These extracted V4 regions were classified using GTDB, and the ANI results were compared against SILVA to determine how the ANI of the V4 reads compared to the full-length 16S. As shown in Figure 5-1, V4 reads tended to match to GTDB with a higher ANI than full-length SILVA reads.

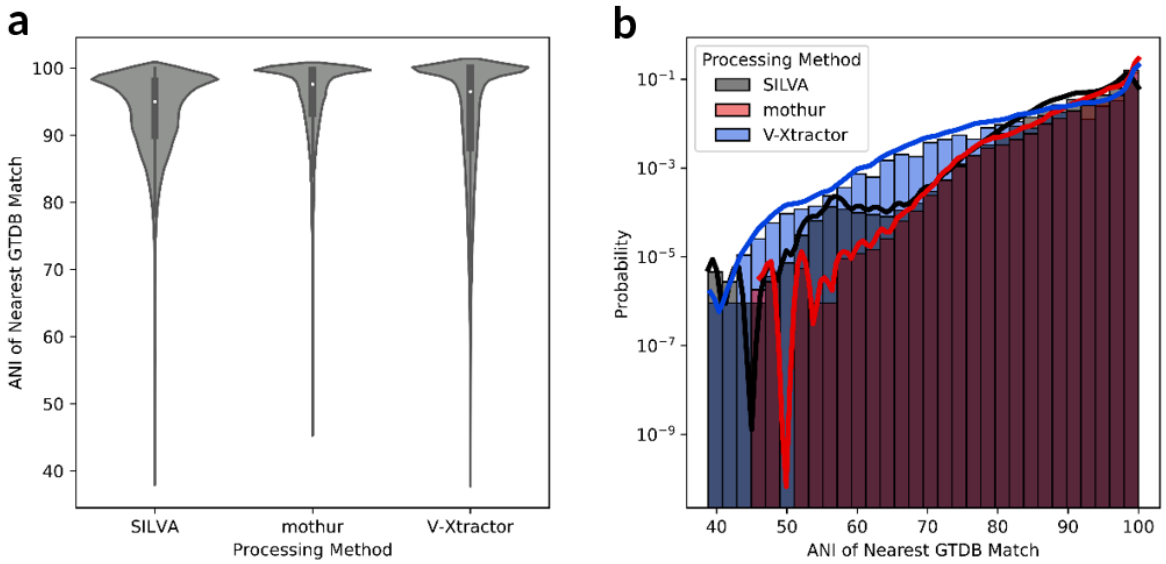


Figure 5-1 | Comparison of the Average Nucleotide Identity of GTDB matches from full-length 16S sequences (~1400bp) and V4 sequences extracted by V-Xtractor (~85bp) and mothur (~253bp). (a) The median of the extracted V4 regions is higher than the full-length SILVA 16S. **(b)** Histograms of the ANI matches provide a more granular view of the distribution of higher V4 ANI. Extraction of V4 regions was performed by Adarsh Singh while taxonomy comparisons were performed by Nkazi Nchinda. n=1,095,814 sequences.

After obtaining the general ANI distribution in Figure 5-1, we assessed to what extent the ANI for extracted V4 regions predicted the ANI of the corresponding full-length 16S. As shown in Figure 5-2, mothur results indicated that the GTDB results of primer-extracted V4 regions correlated with the full-length 16S with an R^2 of 0.58. Results from V-Xtractor indicated that the GTDB results of model-extracted V4 regions from the same samples correlated with the full-length 16S with an R^2 of 0.53.

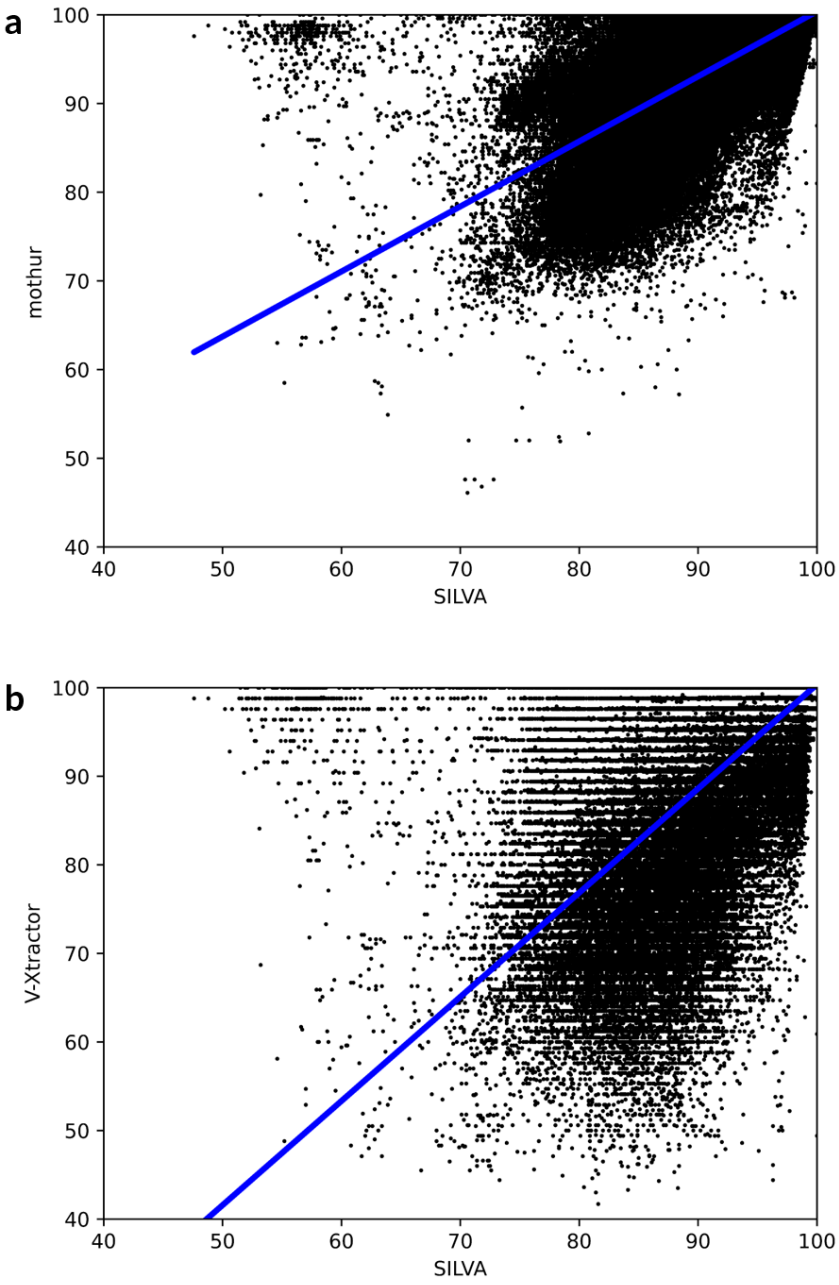


Figure 5-2 | Comparison of the Average Nucleotide Identity of GTDB matches from various-length components of the 16S. (a) Full-length 16S SILVA sequences against V4 regions identified by mothur ($y = 0.74x+27.04$; $R^2=.58$). **(b)** Full-length 16S SILVA sequences against V4 regions identified by V-Xtractor ($y = 1.18x-17.22$; $R^2=.53$). Extraction of V4 regions was performed by Adarsh Singh while taxonomy comparisons were performed by Nkazi Nchinda. n=1,095,814 sequences.

Determining How Varying V4 Read Lengths Impact ANI Matches

The standards of the Earth Microbiome Protocol have changed with time, so typical EMP sequence read lengths vary from 90 to 151 base pairs following processing. To determine how these varying lengths would affect the apparent novelty of samples, we compared ANI matches from mothur (~253 base pairs) and V-Xtractor (~85 base pairs) against one another. As shown in Figure 5-3, results indicated that reads of varying lengths from the same starting sample correlated with an R^2 of 0.77.

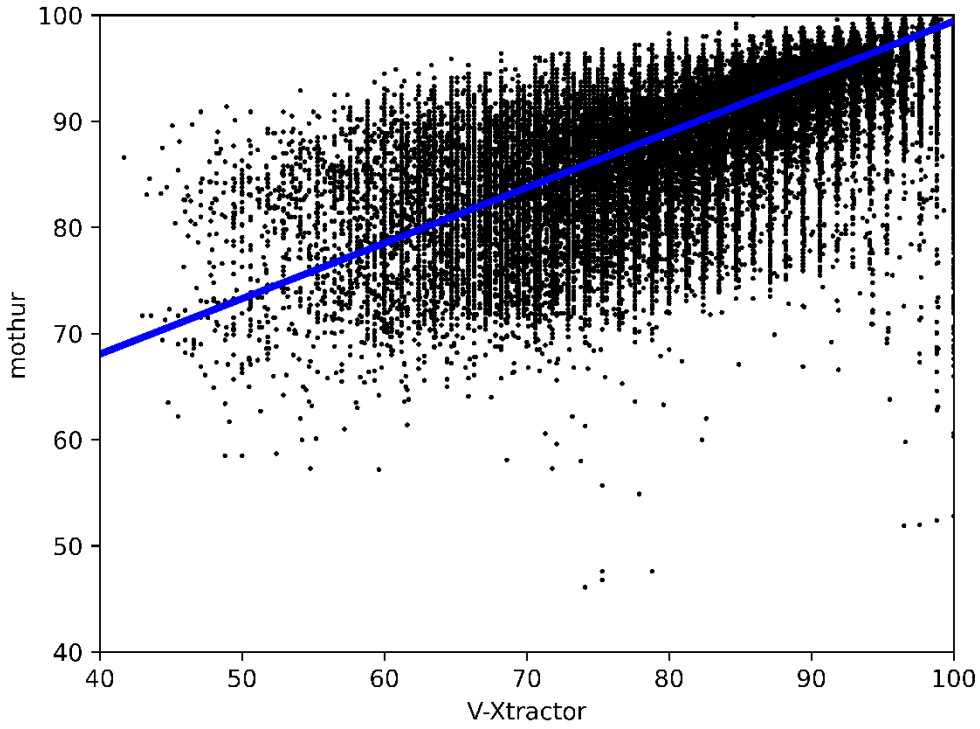


Figure 5-3 | Comparison of the Average Nucleotide Identity of GTDB matches from mothur and V-Xtractor (y = 0.52x+47.16; R^2=.77). Extraction of V4 regions was performed by Adarsh Singh while taxonomy comparisons were performed by Nkazi Nchinda. n=1,095,814 sequences.

Evaluating Samples from Different Environments

Following these two verification steps, 2480 samples from 19 EMP studies were run through the pipeline. Unlike with the SILVA data, sequences from such microbial community samples have an associated relative abundance that was a major consideration, since relative abundance reflects the practical ability to recover organisms at a given sequencing depth. As depicted schematically in Figure 5-4, even if two samples have the same number of novel ASVs, it is easier to recover the corresponding microbial genomes if the bacteria make up a larger fraction of the overall sample.

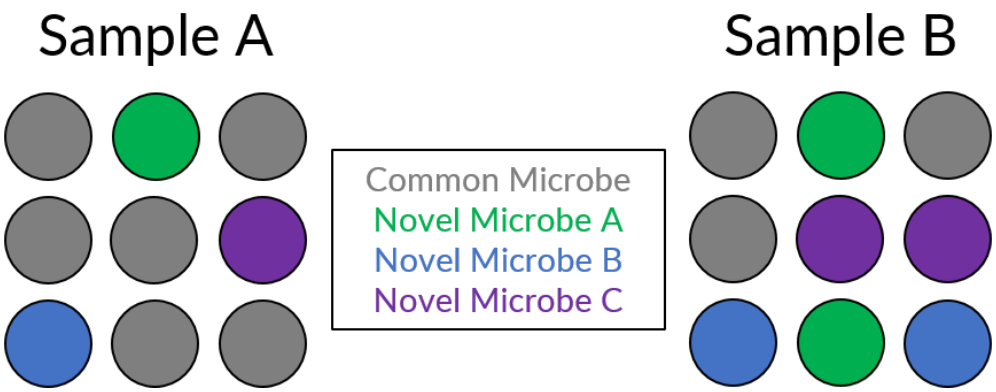


Figure 5-4 | A demonstration of why the difficulty of recovering genomes can vary for samples with the same number of novel ASVs. A researcher would likely select Sample B for further analysis since novel genomes make up a larger proportion of the sample and will thus require less effort to recover.

Figure 5-5 provides data of a single soil sample from an Alaskan tundra with 63,997 reads. In addition to the number of novel ASVs and clusters that a researcher can expect to find of each ANI similarity, the graphs indicate how this number varies as a function of sequencing depth (the maximal number of individuals sampled). The prediction of the number of organisms discovered is probabilistic, depending on relative abundances and the number of individuals sampled, so non-integer values are possible. The median genome size for a bacterium is 3.65Mb [68]. Therefore, for 100x genome coverage, an Illumina MiSeq run (up to 25 million reads of 2x300bp [69]) has a theoretical maximum of about 41 organisms. For each ANI level in Figure 5-5a-b, the number of

unique ASVs and clusters identified from a MiSeq run at each ANI would be between the depth values of 10 and 100. For an Illumina NextSeq run (up to 1.1 billion reads of 2x150bp [70]), the theoretical maximum depth is about 904 organisms.

In Figure 5-5a, the number of items located always drops with clustering. As sequencing depth increases, the number of novel organisms that can be identified in the sample increases accordingly. In Figure 5-5b, a similar trend is visible, but the image is non-cumulative. Most organisms tend to match to GTDB at a higher ANI. In Figure 5-5c, the fraction of clusters at each ANI appears nearly linear when the y-axis is log-scaled, indicating an exponential increase.

Figure 5-6 provides data from a single sample of Catostomid fish slime with 34,138 reads. At every depth, fewer organisms are projected to be found relative to the Alaskan soil sample. In Figure 5-6a, the lines appear to stack near higher ANI values as the depth increases. In Figure 5-6b, there are gaps corresponding to ANI values that had no sequence matches to GTDB. In Figure 5-6c, a constant fraction of the sample is composed of clusters below an ANI of 79, with most clusters matching GTDB at or above 80 ANI.

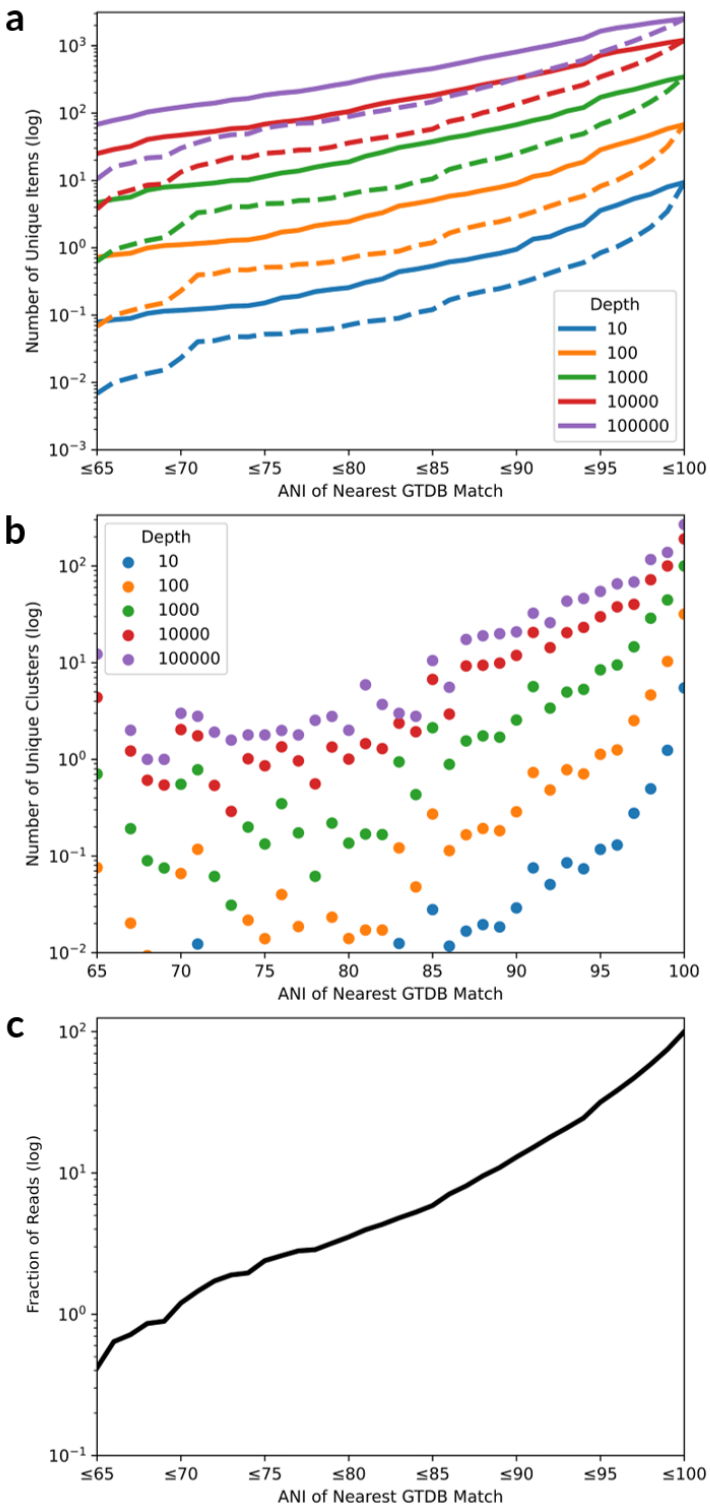


Figure 5-5 | Summary data from a single sample of soil from an Alaskan tundra ecosystem. (a) The cumulative number of novel ASVs (solid) and clusters (dashed) predicted to be discovered at each ANI or below for various sequencing depths. **(b)** The non-cumulative number of novel ASVs predicted to be discovered at each ANI for various sequencing depths. **(c)** The cumulative fraction of the sample composed of reads at each ANI or below.

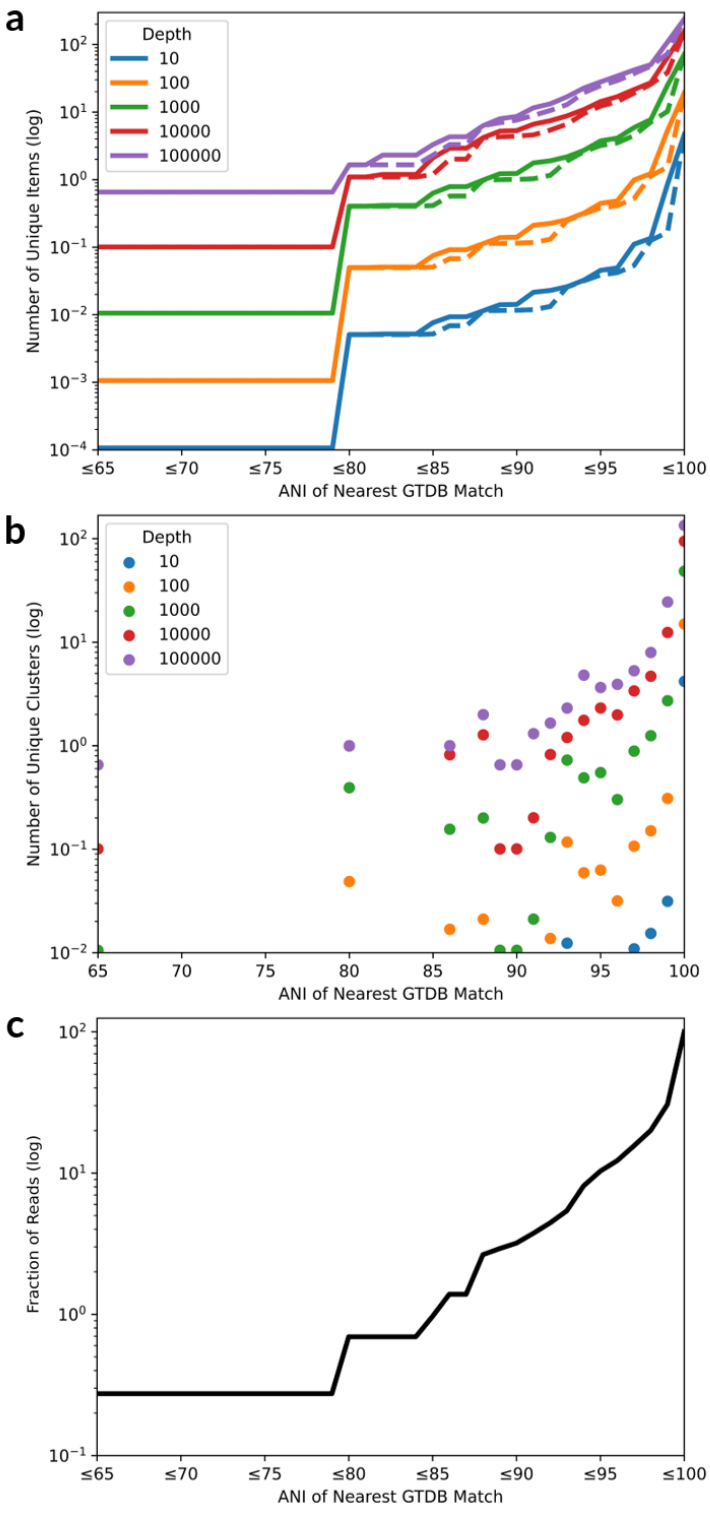


Figure 5-6 | Summary data from a sample of slime from Catostomid fish in Colorado. (a) The cumulative number of novel ASVs (solid) and clusters (dashed) predicted to be discovered at each ANI or below for various sequencing depths. **(b)** The non-cumulative number of novel ASVs predicted to be discovered at each ANI for various sequencing depths. **(c)** The cumulative fraction of the sample composed of reads at each ANI or below.

Comparing Samples from Different EMP Environments

After collecting data for individual samples, we compared the results across studies to determine which environments contained the most novel sequences. For each comparison, the theoretical maximum sequencing depth was held constant. Clusters were compared, rather than ASVs, to avoid overrepresenting the diversity of particular samples. Data from individual samples is shown in Figure 5-7. To avoid spurious fluctuations associated with extremely low read counts, samples with read counts below each specified depth are not displayed. This filter only affects a small number of samples. Due to the visual complexity of Figure 5-7, the figures that follow provide simplified cross-sections of the same data, while a comprehensive figure gallery of samples corresponding to each environment is in the Appendix under EMP Environment Gallery (Page 76).

Even at this resolution, it is evident that as the depth increases from 100 to 1,000 to 10,000 an increasing number of clusters can be identified at all ANI. The shape of sample lines also appears to fall into a few different paradigms. At a sequencing depth of 100 (Figure 5-7a), very few organisms are sampled at a low ANI. The resulting lines sharply increase as the ANI nears 100, which must represent all samples by nature of the cumulative plot. At higher sequencing depths, such as 10,000 (Figure 5-7c), the number of organisms at each level appears to grow almost linearly when plotted on a logarithmic scale.

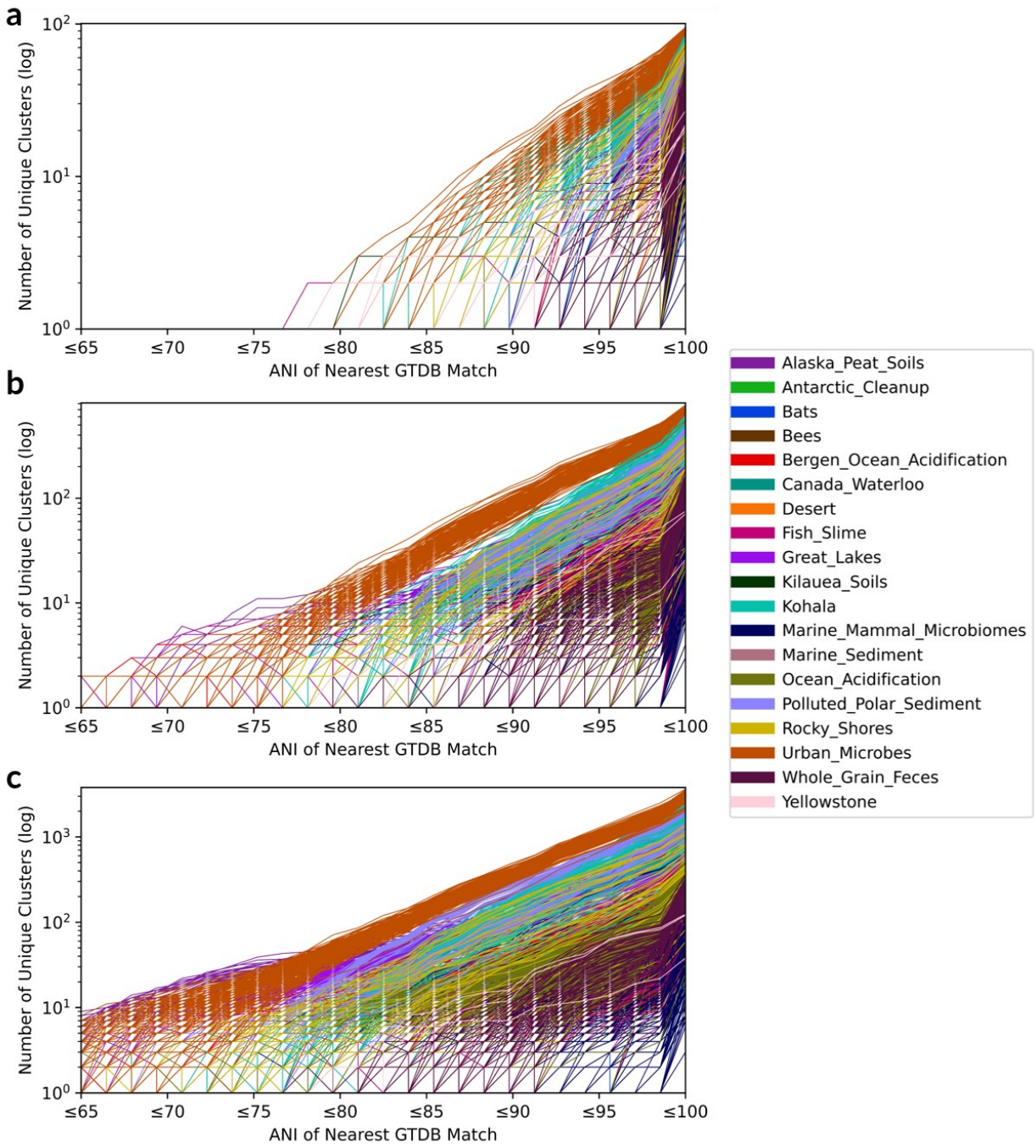


Figure 5-7 | The cumulative number of novel clusters predicted to be discovered at or below each ANI for individual samples at depths of 100, 1000, and 10000. (a-c) The cumulative number of novel clusters identified at each ANI when samples are sequenced at a sequencing depth of 100, 1000, and 10000, respectively. To avoid spurious fluctuations, samples with read counts below each specified depth are not displayed. All plotted values are rounded to the nearest whole number. n = 2480 samples.

In Figure 5-8, the samples corresponding to each environment are averaged. While averaging obscures potentially interesting variations between samples from the same environment, this step was included to aid with visibility. The resulting lines also minimize drastic fluctuations that may occur if one sample from an environment contains many novel microbes by chance.

The results vary with both the taxonomic level of interest and sequencing depth. At a sequencing depth of 100 (Figure 5-8a) and a low ANI, samples from the Kilauea volcano in Hawaii contain the most novel clusters, closely followed by samples from an urban study of Manhattan green roofs and city parks. As the ANI grows, the trend reverses with samples from the urban study containing more novel clusters. Moving further right, the urban study has the most clusters at all high levels of ANI, but samples from the Kohala volcano eventually surpass the Kilauea volcano. At a sequencing depth of 1000 (Figure 5-8b), samples from the urban study continue to contain the most novel clusters at most ANI values. However, for small ANI, samples from Alaskan peat soil contain the most novel clusters. At a sequencing depth of 10,000 (Figure 5-8c), the observed trend is similar to 1000. However, at high ANI, samples from polluted polar sediment surpass the number of clusters from Kohala soil samples.

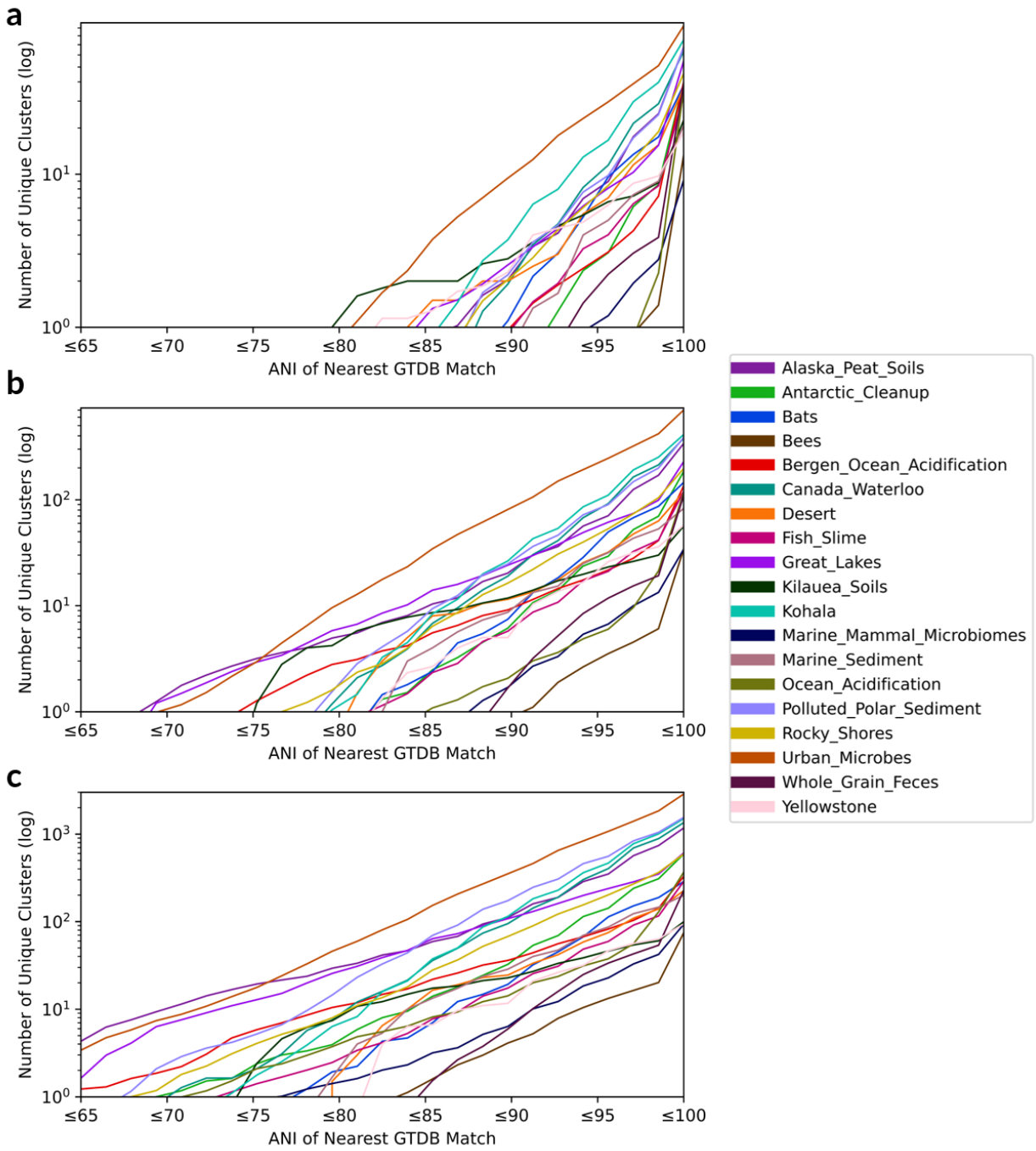


Figure 5-8 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for each environment (averaged from individual samples) at varying depths. (a-c) The cumulative number of novel clusters identified at each ANI when samples are sequenced at a sequencing depth of 100, 1000, and 10000, respectively. All plotted values are rounded to the nearest whole number, and samples with fewer reads than the specified depth are not included in the average.

Lastly, in Figure 5-9, samples are grouped and averaged based on the type of collection site. Soil and aquatic samples are known to boast the most microbial biodiversity [33], so we were curious whether having more types of microbes results in having more *novel* microbes. Again, this averaging process obscures potentially interesting variations between individual samples or environments.

At all depths, soil-associated samples contain the most clusters that are novel at every ANI level. At a sequencing depth of 100, water-associated and animal-associated samples present similar numbers of novel clusters across ANI levels. However, at higher sequencing depths, marine samples have more clusters than animal-associated samples for all ANI.

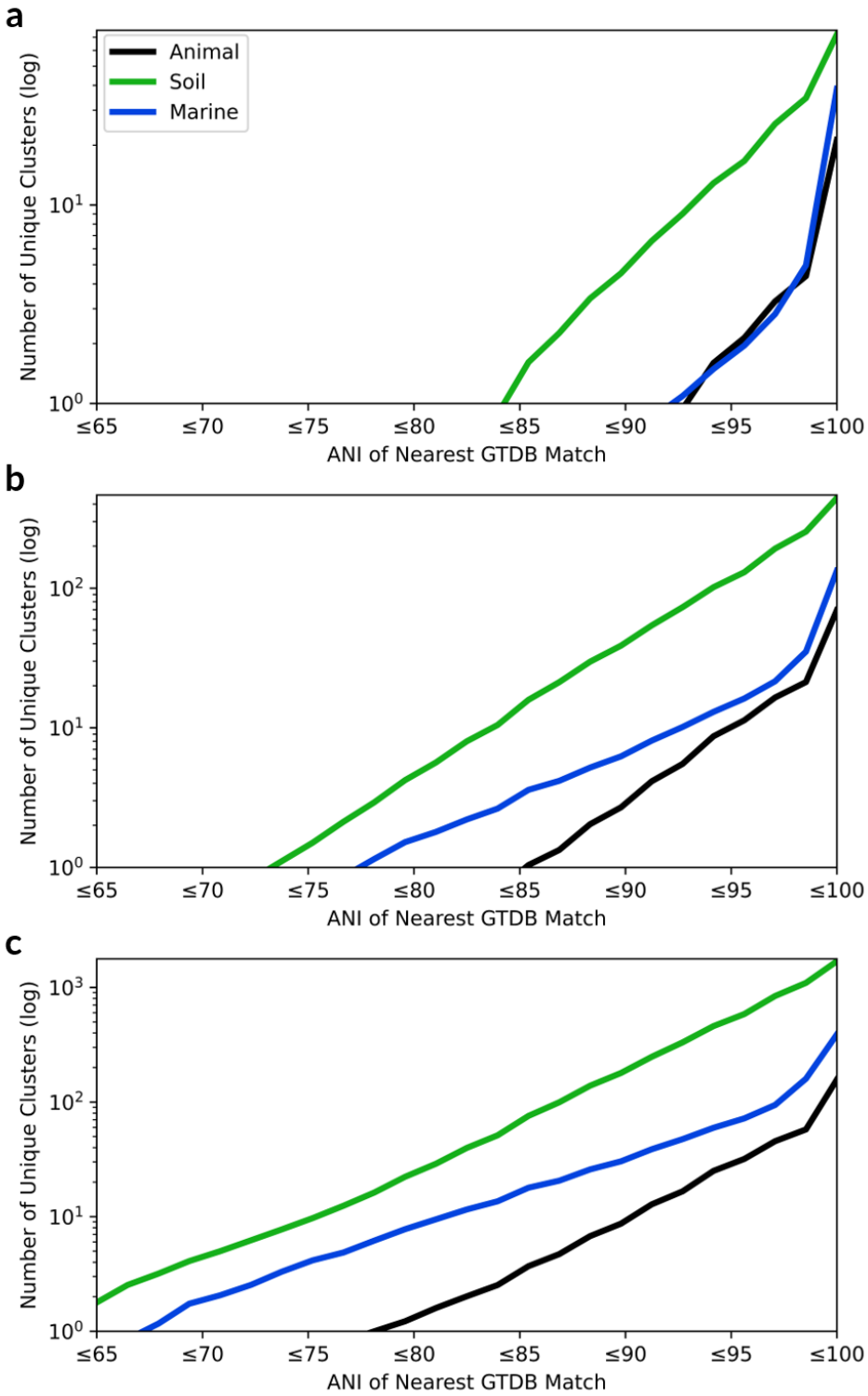


Figure 5-9 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for environment types (averaged from samples) at varying depths. (a-c) The cumulative number of novel clusters identified at each ANI when samples are sequenced at a sequencing depth of 100, 1000, and 10000, respectively. All plotted values are rounded to the nearest whole number, and samples with fewer reads than the specified depth are not included in the average.

Chapter 6: Results from Collected Samples

Comparing Environmental Samples with Standardized Processing

In addition to comparing environmental samples from other researchers' data, we applied the pipeline to samples that we had collected from various sites. Apart from the Hawaii coastal soils, all samples were collected at the same time, prepared identically, and sequenced in the same run with the same read length, eliminating much of the potential variability in read length and preparation between EMP samples. This experiment aimed to mimic how a researcher might use the pipeline in an experimental context and determine where to focus future in-lab experiments.

The results of comparing collected samples are shown in Figure 6-1. Relative to the EMP samples, the sample with the most novel clusters varies many more times across the range of ANI. More specific trends are visible after categorizing samples.

Figure 6-2 shows the results of averaging individual samples based on the starting environment. At a sequencing depth of 100, the moss sample has the most novel clusters across a wide range of ANI. As the depth increases, it is gradually surpassed by samples associated with soil. In general, samples collected from urban locations (cars, brick, etc.) tend to contain low numbers of novel organisms.

When comparing samples, it is also notable that the marsh samples frequently contain the highest numbers of novel clusters. As shown in Figure 6-3, this trend holds across most ANI and at all sequencing depths.

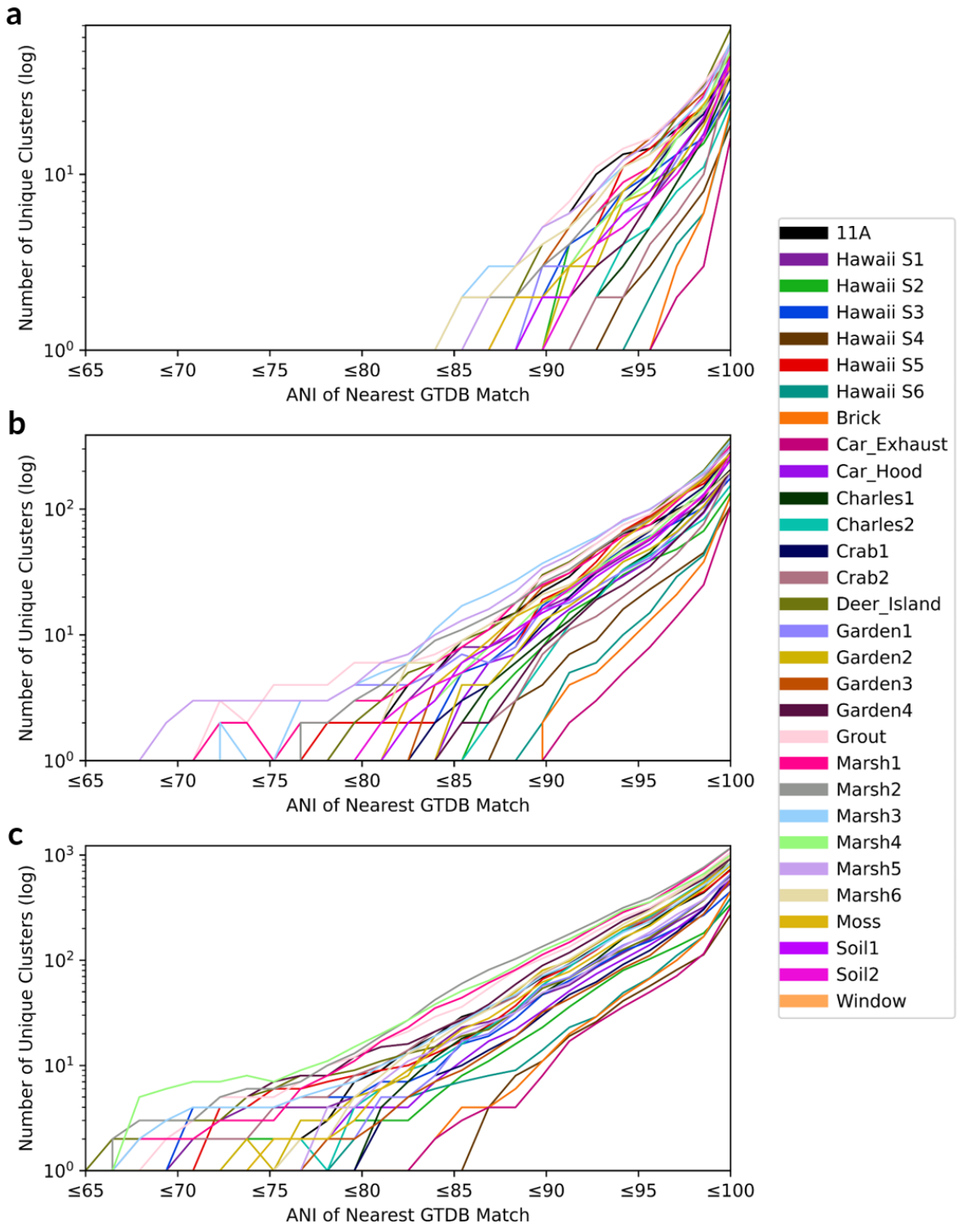


Figure 6-1 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for individual samples at varying depths. (a-c) The cumulative number of novel clusters identified at each ANI when samples are sequenced at a sequencing depth of 100, 1000, and 10000, respectively. All plotted values are rounded to the nearest whole number.

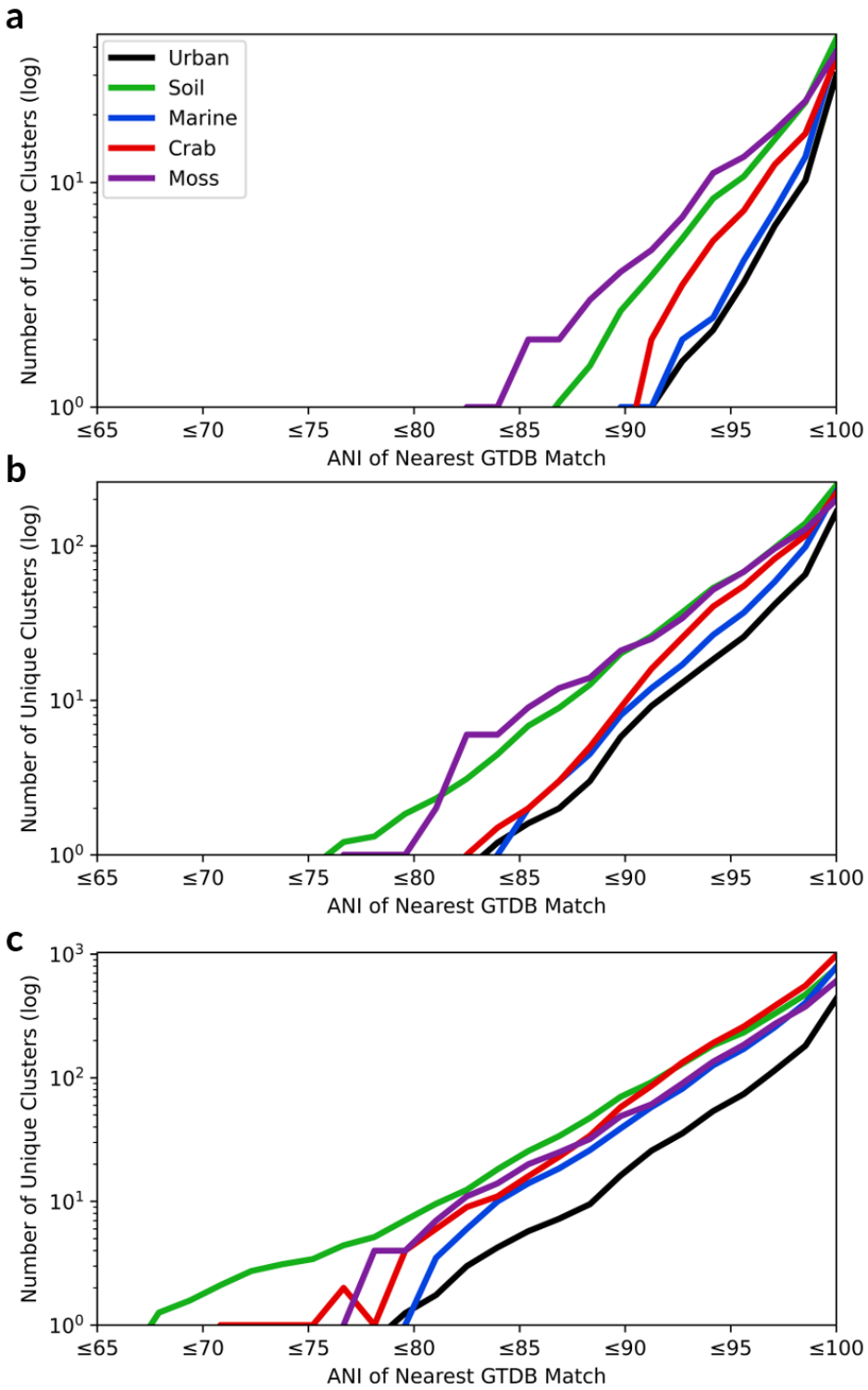


Figure 6-2 | The cumulative number of novel clusters predicted to be discovered at each ANI or below for environment types (averaged from samples) at varying depths. (a-c) The cumulative number of novel clusters identified at each ANI when samples are sequenced at a sequencing depth of 100, 1000, and 10000, respectively. All plotted values are rounded to the nearest whole number, and samples with fewer reads than the specified depth are not included in the average.

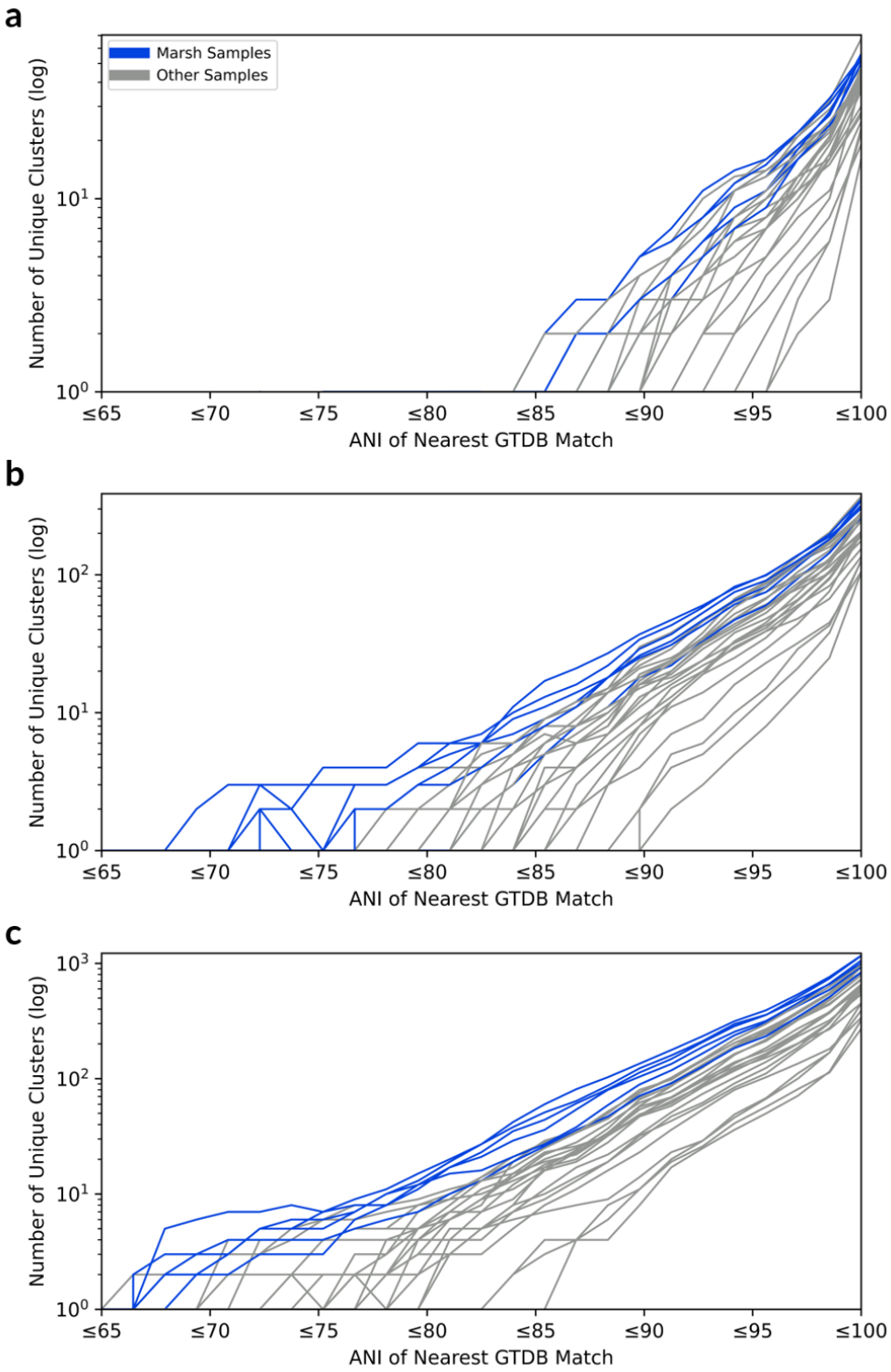


Figure 6-3 | A comparison of the cumulative number of novel clusters predicted to be discovered at each ANI or below for marsh samples relative to other samples at varying depths. (a-c) The cumulative number of novel clusters identified at each ANI when samples are sequenced at a sequencing depth of 100, 1000, and 10000, respectively. All plotted values are rounded to the nearest whole number.

Mini-Metagenomic Experiment

Since soil and aquatic samples contain the most microbial biodiversity [33], we decided to select one marsh sample for further analysis. In particular, we ran a mini-metagenomic experiment to attempt to characterize a highly novel organism. As described in Level 2: Whole Genome Sequencing (Page 15), mini-metagenomics leverages single-cell methods to minimize the environmental skew typically associated with environmental samples. Full details regarding sample preparation are included in the Appendix under Mini-Metagenomic Preparation of Marsh 5 (Page 97). The result was a 96-well plate with a small number of bacterial cells in each well.

Locating a Gracilibacteria

Though the pipeline was primarily used for comparing environmental samples, based on Figure 3-2, we theorized that it could also be used to find novel organisms in wet-lab samples. For instance, following 16S sequencing, samples with a V4 ANI above a certain threshold could be eliminated from consideration for Level 2 characterization. The remaining samples would have a lower V4 ANI on average and be more likely to contain novel organisms.

Following this enrichment method, we identified a well containing an organism that appeared to be highly novel at the class level. As shown in Figure 6-4, more than 55% of the V4 reads from the well corresponded to an organism called a Gracilibacteria. Within the original community, less than 1% of reads corresponded to the same organism. This strategy enabled us to isolate the novel organism in a well for whole genome sequencing.

Ensuing work by other members of the Cira Lab resulted in the assembly of the full Gracilibacteria genome. The organism was also noted to have introns, unusual codon usage, and a CRISPR/CAS system.

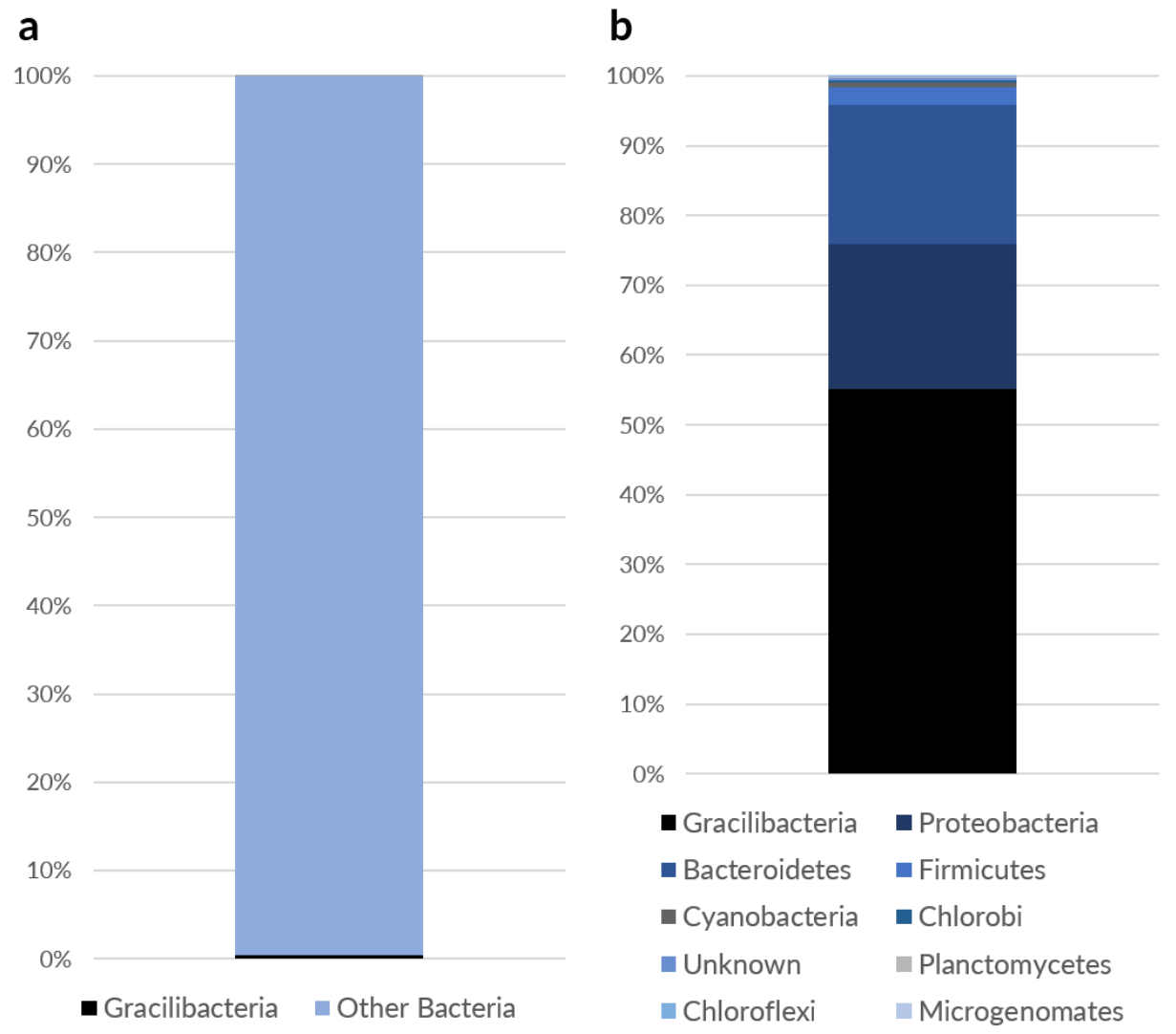


Figure 6-4 | Fraction of reads corresponding to a Gracilibacteria and other organisms from a sequencing run of a salt marsh sample. (a) Read fractions of the Gracilibacteria relative to the total number of reads from all sample wells. **(b)** Read fractions within the well where a Gracilibacteria reads make up most data.

Chapter 7: Discussion

GTDB

In Figure 3-1, we demonstrated that there is an association between V4 ANI and whole genome taxonomic classification. Though there was a broad spread in the ANI corresponding to a particular taxonomic level, the median ANI characteristically increases as the taxonomic levels decrease, suggesting that V4 ANI values carry information about an organism's novelty. This result suggests that the pipeline's strategy of assessing novelty from V4 ANI is viable, particularly when multiple organisms are analyzed simultaneously in a sample. The greatest ANI similarities occur between the phylum and class levels. As noted in A Specialized Approach (19), the taxonomic levels are artificial constructs [57] resulting in a high degree of subjectivity. Furthermore, as taxonomic levels increase, they tend to grow less coherent [71], which could partially account for the increased similarities between the phylum and class ANI.

Figure 3-2 illustrates how the V4 ANI can be used to estimate the novelty of organisms in practice. For instance, if an organism has a V4 ANI of 80, there is roughly a 40% chance that it is novel at the phylum level, 40% chance that it is novel at the class level, and 20% chance that it is novel between the order and genus levels. For extremely low ANI, the boundaries between taxonomic levels begin to break down. This effect could be due to contamination in the database from misassembled whole genomes that contain 16S sequences from the wrong organism. However, for ANI above 70, the V4 provides a robust way to predict the novelty of an organism's whole genome.

This ability to predict novelty is particularly useful in the context of comparing environments. When using the pipeline to compare samples from different environments, if a researcher knows the number of clusters at a given ANI (such as in Figure 5-5b), they can use

Figure 3-2 to map the ANI of discovered organisms directly to an estimated taxonomic level. In other words, rather than expecting to find "100 organisms that are novel at an ANI of 75," they might expect to find approximately 40 phyla, 40 classes, and 20 lower taxa.

Predicting novelty is also useful in the context of enriching for novel organisms. As briefly described in Locating a Gracilibacteria (56), V4 ANI can be used as a cutoff to increase the novelty of organisms selected. For instance, if a researcher has V4 reads from several experimental wells, they can remove wells with a V4 ANI of above 90% from consideration for whole genome characterization (Level 2). In the process, they would eliminate most organisms that are novel at the species level and greatly increase the likelihood of locating an organism that is novel at the phyla or class level. In other words, they would ensure that the organisms they are characterizing at Level 2 are much more likely to be novel.

Though we ran a mini-metagenomic experiment leveraging this enrichment strategy, future wet-lab work could help to determine the efficacy of V4-based enrichment strategies in practice. Depending on the results, these approaches could be applied to help identify highly novel organisms and resolve gaps in the tree of life.

SILVA

In Figure 4-1, we estimated at each taxonomic level how many organisms are characterized at Level 1 (16S) that remain to be characterized at Level 2. Using Figure 3-2, it is then possible to approximate how many organisms are novel at each taxonomic level. For instance, Figure 4-1 indicates that there are approximately 5000 clusters that match GTDB at an ANI of 85 or below. For ANI values below 85 in Figure 3-2, the average proportion of matches corresponding to phylum is 40%. Thus, we can expect approximately 2000 of these clusters to belong to a novel phylum. Given that a previously discovered group of 35 new phyla could comprise up to 15% of

the bacterial domain [8], this result suggests that if all 2000 clusters were to be sequenced at the whole genome level, the tree of life would drastically expand as a result. However, several effects suggest that we should not solely rely on this absolute number of clusters.

Multiple factors could lead to an overestimate. As noted earlier, it is impossible to state with complete certainty that novel clusters are not chimeras since these sequences are known to accumulate in 16S databases [37] and these sequences could easily appear highly different from known sequences. Additionally, the SILVA data is based on a full-length 16S comparison, but Figure 3-2 is based on V4 data. Results from Figure 5-1 suggest that the V4 is likely to underrepresent sequence novelty relative to the full-length 16S. In other words, a novel organism will likely have a higher V4 ANI than full-length 16S ANI. This trend means that if Figure 3-2 were recomputed with full-length 16S data, the plot would likely shift left, reducing the ANI associated with each taxonomic level. This effect would slightly decrease the estimated numbers of organisms that are highly novel. Thus, the use of V4-based Figure 3-2 for a full-length 16S calculation likely results in an overestimate.

Other factors could push the estimate in the opposite direction by removing real sequences that are novel. The inclusion of a stringent "singleton" filter [59] that threw out all clusters with one item likely discarded several novel sequences in addition to chimeric ones. Additionally, all chloroplasts (56,572) sequences were removed from the SILVA database to prevent the inclusion of any eukaryotes. Lastly, the database was pre-clustered, removing similar organisms that were potentially novel. These effects would result in an underestimate, but it is difficult to determine their magnitude relative to the factors that could lead to an overestimate.

Future research could build off this work by incorporating phylogeny. Though highly novel clusters have been identified, it remains unknown where on the tree of life these organisms are

located. Previous research has identified bacterial phyla that are underrepresented in sequencing efforts, such as Tenericutes and Spirochaetes [72]. A continuation of this project could investigate whether the highly novel clusters are on a branch adjacent to these underrepresented phyla or elsewhere in the tree. Researchers could also consider re-running this same experiment without de-replicating the SILVA dataset to get a more accurate indication of the upper limit of clusters.

Earth Microbiome Protocol

The Impact of the V4 Subunit Relative to the Full 16S

In Figure 5-1a, we demonstrated that V4 regions extracted via V-Xtractor and mothur tend to have higher ANI GTDB matches than the full-length 16S sequences from SILVA. This issue is especially pronounced with the primer-based V4 extraction from mothur. Ideally, the ANI distribution of the full 16S region would perfectly match the V4 region so organisms would appear equally novel regardless of whether the full 16S or solely the V4 was sequenced. However, since the EMP database contains high-throughput short reads of the higher-ANI V4, overall, EMP samples will appear to contain fewer novel microbes than the full 16S would suggest.

Figure 5-1b displays that the ANI distribution for V4 regions is much more compact than the full 16S. Prior research demonstrates that Illumina reads of 75-100 base pairs are sufficient to represent between-sample diversity [73]. However, a tighter ANI distribution would make it more difficult to discriminate between similar microbes, potentially leading to erroneous assessments of which organisms are novel. This issue is further demonstrated in Figure 5-2 by the moderate correlations between the nearest ANI match of SILVA and mothur (R^2=0.58) and SILVA and V-Xtractor (R^2=0.53). These correlations suggest that the pipeline has a limited ability to gauge the novelty level of an individual organism based on its V4 sequence relative to the full-length 16S.

The Impact of Varying Read Lengths on ANI Matches

We demonstrated in Figure 5-3 that short V4 regions extracted by V-Xtractor and longer V4 regions extracted by mothur are correlated with an R^2 of 0.77. This moderate correlation suggests that the varying read length of EMP samples has an impact on the novelty reported by the pipeline. As shown in Figure 5-1a, V4 reads from mothur (~253 base pairs) tended to have the highest median ANI and accordingly, the lowest novelty. The V4 reads extracted by V-Xtractor (~85 base pairs) did not match quite as highly. Results from V-Xtractor also had more outliers that matched to GTDB extremely poorly, which could represent V4 targeting errors in the Hidden Markov Model that do not appear with the primer-based approach. Overall, while reads of varying lengths will produce similar results, EMP samples with longer reads are expected to appear slightly less novel. This trend could partially be attributable to the inclusion of the constant region flanking the V4 in longer reads.

A Basis for Evaluating Samples from Different Environments

Figure 5-5 and Figure 5-6 demonstrate that the graph of ANI for individual samples can vary based on both the microbial population present and a researcher's choice of sequencing depth.

Both samples were processed on an Illumina HiSeq machine, which should theoretically produce the same amount of data in each case. However, if many more samples are pooled on a single HiSeq run, the number of reads corresponding to each sample decreases. This is likely what occurred with the fish sample in Figure 5-6. The 10,000 and 100,000 depth lines in 5-6a are extremely similar since even at an increased depth, the pipeline cannot identify microbes that were not picked up in the original sample of 34,000 reads.

Given that starting samples can vary drastically in sequencing depth, in some circumstances it makes sense to evaluate samples at lower depths. It has previously been shown

that 2000 single-end reads are required to recover the same relationships between samples as the full dataset [73], so EMP samples can contain anywhere from few thousand to hundreds of thousands of reads. In large part, the probability-based assessment should normalize this variance by setting a fixed limit on the depth—how many organisms can be identified by the sequencer. However, if the pipeline depth on the figure greatly exceeds the sequencing depth of the original sample, the number of ASVs and clusters will be underestimated. To ensure that low-read samples are not unfairly disadvantaged during cross-sample comparison, researchers using the pipeline should carefully consider the choice of sequencing depth or consider subsampling all environments. Researchers should choose the highest depth that they can afford to sequence to get the most accurate representation of what the sequencer will detect, but if lines of different depths appear to stack, this likely suggests that the starting sample was not sequenced deeply enough and a lower depth should be used for comparison.

Comparing Different Environments

We demonstrated in Figure 5-8 that the "optimal" sampling environment varies based on both the desired taxonomic level and sequencing depth. At some low ANI, Alaska peat soil samples appear to dominate. At higher ANI values, samples associated with the urban Manhattan study tend to perform better. Thus, the recommended place to look for microbes depends on the taxonomic novelty of what researchers are hoping to find.

The figure also demonstrates that the ranking of samples can vary based on the desired sequencing depth. If a researcher has a higher throughput machine, they will likely want to compare environments at a high sequencing depth to get an accurate representation of how many organisms they can expect to find. Among the samples we compared, typical sample read counts were usually above 100,000. To ensure headroom of a couple of orders of magnitude from the

sequencing depth and avoid disadvantaging lower-read samples, these environments can be compared using the depth of 1000. Therefore, among the analyzed EMP samples, to discover highly novel organisms and fill the gaps in the tree of life, samples from Alaska peat soils or Manhattan green roofs are likely good places to start.

In Figure 5-9, we noted that soil samples tended to contain the most novel clusters, followed by marine samples, and finally animal-associated microbes. In some ways, this result was unsurprising. Soil and aquatic samples are known to boast the most microbial biodiversity [33] but make up a minority of full-length 16S sequences in databases [32]. Thus, while there was no way to predict that soil from the Alaskan tundra or Manhattan green roofs would contain the most *novel* microbes, they were likely to contain a diverse community. A future study could investigate to what extent V4 novelty and microbial diversity correlate. We collected this information, but time constraints limited an in-depth analysis.

However, there are a few suggestions that our results may not be perfectly representative of these environments. To start, solid soil samples can have extreme short-distance heterogeneity [74] and these changes in texture can have significant impacts on which bacteria are present [75]. Though these samples appear to contain high numbers of novel clusters, there is no guarantee that a future sample of the same environments will recover the same organisms. Furthermore, the Manhattan study performs better than other samples to an unusual degree, even relative to other soil samples. Compared to other samples, no significant differences in read length were observed, and the probability-based pipeline should minimize differences due to read count. This behavior may indicate that there is an underlying structural factor from sample preparation at play. Such variations are inherent in relying on data from other researchers and helped to motivate our in-lab

sample preparation using standardized reagents and techniques across a wide variety of environments.

Future work in this area may seek to identify factors that explain the Manhattan study's unusual performance to ensure that they do not lead to the mischaracterization of certain environments. Researchers could further improve this analysis by cropping all EMP reads to a uniform length or discarding certain datasets to ensure that read length variability does not affect environment selection. It would also be worthwhile to include additional EMP datasets to conduct an even more comprehensive overview.

As an extension of this project, it would be valuable to determine the relative species abundance of organisms at varying levels of novelty. For instance, it would be useful to know if microbes that tend to match GTDB at lower ANI also tend to be rarer in environmental samples. If so, this would further promote the need for tools such as the pipeline that enrich for these organisms.

Additionally, Figure 5-8c demonstrates that, especially at higher sequencing depths, the relationship between organisms at each ANI appears to be roughly linear following a log transformation. A future study could evaluate whether it is possible to predict the number of organisms in a sample at low ANI based on the number at high ANI. For instance, based on the number of novel organisms at the genus and species level in a sample, this could allow a researcher to predict whether it may contain novel phyla that were missed in the initial sequencing run.

Overall, these results demonstrate the usefulness of the pipeline and provide another signal that environmental samples are a drastically under-sequenced resource.

Collected Samples

Comparing Standardized Samples from Environments

Comparing the EMP results in Figure 5-8 and collected sample results in Figure 6-1 suggests that sample processing may play a role in the number of novel clusters identified between samples. In Figure 5-8, a few samples are much more likely to dominate while in Figure 6-1 the results of different environments tend to be more similar. These differences in deviation could be due to differences in the samples themselves and sample selection. However, there may also be a component that stems from sample processing. In Figure 6-1, samples from various environments were processed simultaneously in the laboratory using the same workflow, eliminating much of the variability between EMP studies. This difference may account for the smaller variation in performance between environments.

When grouping environments in Figure 6-2, soil continues to perform very well at high depths, similarly to the EMP. As noted earlier, this is partially to be expected due to the high diversity associated with soil and water samples [33]. In the absence of the Manhattan sample, while soil continues to perform well, the difference is less extreme than prior results from Figure 5-9. Surprisingly, however, host-associated microbes—moss and horseshoe crab—outperformed marine samples from the Charles River and a sewage facility. Looking at Figure 6-1 indicates that the Charles River sample performed extremely poorly. There are multiple potential explanations for this trend, so it is difficult to assess whether the low number of novel clusters resulted from bias—such as during sample preparation—or an actual lack of novel clusters in the environmental sample.

We demonstrate in Figure 6-3 that salt marsh samples tend to perform well relative to other samples at all sequencing depths. These results strengthen the justification for our decision to select

a salt marsh sample for mini-metagenomics and characterization of an organism. They also emphasize that for future in-lab characterization efforts of novel microbes, salt marshes may be a useful site to consider.

Mini-Metagenomic Experiment

In Figure 6-4, we demonstrate how V4-based enrichment can be applied to existing mini-metagenomic methods to move organisms from Level 1 to Level 2. Using single-cell techniques, we were able to fluidically isolate several bacteria in individual wells, enabling the recovery of rare genomes in relatively pure form. The figure demonstrates that isolating an organism in a well can drastically increase the fraction of reads corresponding to the organism, from less than 1% to over 55%. V4-enrichment built on this premise, allowing us to determine that the bacteria in question was a Gracilibacteria that was likely to be novel at the phylum or class levels. Then, the organism could be selected for Level 2 characterization and other members of the Cira Lab could successfully reconstruct its genome.

This workflow exemplifies how the pipeline can be used for enrichment, and in our case, we discovered a rare, novel Gracilibacteria in the process. Other researchers could consider applying the pipeline to other samples to continue discovering novel organisms and resolving gaps in the tree of life.

Conclusion

In conclusion, we have developed a bioinformatics-based pipeline for identifying and quantifying novel organisms based on their 16S sequences. This approach was validated using GTDB data, which indicated that there was a strong link between 16S V4 ANI and whole genome taxonomy for ANI above 70. Leveraging this predictive ability, the pipeline can be used both to compare environmental samples and enrich wet-lab samples for novel bacteria. By applying the pipeline to 16S SILVA data, we were able to estimate the number of organisms at a range of taxonomic levels that are known at the 16S level of characterization but have not been sequenced at the whole genome level. We were also able to demonstrate that the V4 subunit underestimates organism novelty relative to the full 16S and this issue becomes more pronounced with longer V4 reads. By applying the pipeline to EMP datasets, we determined that soil samples have the highest likelihood of containing novel microbes, but individual sample selection should be based on the target level of taxonomic novelty and laboratory sequencing depth. Applying the pipeline to standardized samples from multiple environments also indicated that soil is likely to contain the highest proportion of novel microbes, with salt marsh samples performing particularly well. Lastly, a combination of mini-metagenomics and the pipeline was able to enrich one of these samples for organisms that were both rare and novel. The resulting assembled genome from a novel Gracilibacteria at the class level demonstrates the potential of this pipeline for identifying novel microbes and helping to resolve remaining gaps in the tree of life.

Chapter 8: References

[1] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 12, pp. 4576–4579, Jun. 1990.

[2] P. Yarza *et al.*, "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences," *Nat. Rev. Microbiol.*, vol. 12, no. 9, pp. 635–645, 2014.

[3] C. Pedrós-Alió, "Marine microbial diversity: can it be determined?," *Trends Microbiol.*, vol. 14, no. 6, pp. 257–263, Jun. 2006.

[4] K. J. Locey and J. T. Lennon, "Scaling laws predict global microbial diversity.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 21, pp. 5970–5, May 2016.

[5] L. A. Hug *et al.*, "A new view of the tree of life," *Nat. Microbiol.*, vol. 1, no. 5, p. 16048, May 2016.

[6] R. Gouy, D. Baurain, and H. Philippe, "Rooting the tree of life: The phylogenetic jury is still out," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1678. Royal Society of London, 31-Aug-2015.

[7] C. E. Hinchliff *et al.*, "Synthesis of phylogeny and taxonomy into a comprehensive tree of life," *Proc. Natl. Acad. Sci.*, vol. 112, no. 41, pp. 12764–12769, Oct. 2015.

[8] C. T. Brown *et al.*, "Unusual biology across a group comprising more than 15% of domain Bacteria," *Nature*, vol. 523, no. 7559, pp. 208–211, Jul. 2015.

[9] L. A. Hug *et al.*, "A new view of the tree of life," *Nat. Microbiol.*, vol. 1, no. 5, p. 16048, Apr. 2016.

[10] C. Rinke *et al.*, "Insights into the phylogeny and coding potential of microbial dark matter," *Nature*, vol. 499, no. 7459, pp. 431–437, Jul. 2013.

[11] L. Manoharan, J. A. Kozlowski, R. W. Murdoch, F. E. Löffler, F. L. Sousa, and C. Schleper, "Metagenomes from Coastal Marine Sediments Give Insights into the Ecological Role and Cellular Features of Loki- and Thorarchaeota," *MBio*, vol. 10, no. 5, Sep. 2019.

[12] J. M. Archibald, "Endosymbiosis and eukaryotic cell evolution," *Current Biology*, vol. 25, no. 19. Cell Press, pp. R911–R921, 05-Oct-2015.

[13] C. R. Benitez-Nelson, "The biogeochemical cycling of phosphorus in marine systems," *Earth Sci. Rev.*, vol. 51, no. 1–4, pp. 109–135, Aug. 2000.

[14] M. M. Brinson, "Decomposition and Nutrient Exchange of Litter in an Alluvial Swamp Forest," *Ecology*, vol. 58, no. 3, pp. 601–609, May 1977.

[15] L. V. Hooper and J. I. Gordon, "Commensal host-bacterial relationships in the gut," *Science*, vol. 292, no. 5519. American Association for the Advancement of Science, pp. 1115–1118, 11-May-2001.

[16] J. Jovel *et al.*, "Characterization of the gut microbiome using 16S or shotgun metagenomics," *Front. Microbiol.*, vol. 7, no. APR, p. 459, Apr. 2016.

[17] S. Chochua *et al.*, "Population and whole genome sequence based characterization of invasive group a streptococci recovered in the United States during 2015," *MBio*, vol. 8, no. 5, Sep. 2017.

[18] Z. Y. Zhang, L. P. Pan, and H. H. Li, “Isolation, identification and characterization of soil microbes which degrade phenolic allelochemicals,” *J. Appl. Microbiol.*, vol. 108, no. 5, pp. 1839–1849, May 2010.

[19] L. W. Hugerth and A. F. Andersson, “Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing,” *Frontiers in Microbiology*, vol. 8, no. SEP. Frontiers Media S.A., p. 1561, 04-Sep-2017.

[20] C. R. Woese and G. E. Fox, “Phylogenetic structure of the prokaryotic domain: The primary kingdoms,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 11, pp. 5088–5090, Nov. 1977.

[21] W. Ludwig and K. H. Schleifer, “Bacterial phylogeny based on 16S and 23S rRNA sequence analysis,” *FEMS Microbiol. Rev.*, vol. 15, no. 2–3, pp. 155–173, Oct. 1994.

[22] Y. Van de Peer, S. Chapelle, and R. De Wachter, “A quantitative map of nucleotide substitution rates in bacterial rRNA,” *Nucleic Acids Res.*, vol. 24, no. 17, pp. 3381–3391, Sep. 1996.

[23] W. Ludwig and H.-P. Klenk, *Bergey’s Manual® of Systematic Bacteriology*, 2nd ed., vol. 1. New York: Springer New York, 2001.

[24] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, “The Human Microbiome Project,” *Nature*, vol. 449, no. 7164. Nature Publishing Group, pp. 804–810, 18-Oct-2007.

[25] T. Z. DeSantis *et al.*, “Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB,” *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–5072, Jul. 2006.

[26] E. Pruesse *et al.*, “SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB,” *Nucleic Acids Res.*, vol. 35, no. 21, pp. 7188–7196, Dec. 2007.

[27] B. Yang, Y. Wang, and P. Y. Qian, “Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis,” *BMC Bioinformatics*, vol. 17, no. 1, p. 135, Mar. 2016.

[28] “EMP - Earth Microbiome Project.” [Online]. Available: https://qiita.ucsd.edu/emp/study/list/. [Accessed: 15-Mar-2021].

[29] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis,” *ISME J.*, vol. 11, no. 12, pp. 2639–2643, Dec. 2017.

[30] M. S. Rappé and S. J. Giovannoni, “The Uncultured Microbial Majority,” *Annual Review of Microbiology*, vol. 57. Annu Rev Microbiol, pp. 369–394, 2003.

[31] D. Wu *et al.*, “A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea,” *Nature*, vol. 462, no. 7276, pp. 1056–1060, Dec. 2009.

[32] P. D. Schloss, R. A. Girard, T. Martin, J. Edwards, and J. C. Thrash, “Status of the Archaeal and Bacterial Census: an Update.,” *MBio*, vol. 7, no. 3, pp. e00201-16, Jul. 2016.

[33] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, “Prokaryotes: The unseen majority,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 12. National Academy of Sciences, pp. 6578–6583, 09-Jun-1998.

[34] A. E. Parada, D. M. Needham, and J. A. Fuhrman, “Every base matters: Assessing small

subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples," *Environ. Microbiol.*, vol. 18, no. 5, pp. 1403–1414, May 2016.

[35] E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, and N. C. Kyrpides, "Metagenomics uncovers gaps in amplicon-based detection of microbial diversity," *Nat. Microbiol.*, vol. 1, no. 4, p. 15032, Feb. 2016.

[36] B. J. Haas *et al.*, "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons," *Genome Res.*, vol. 21, no. 3, pp. 494–504, Mar. 2011.

[37] P. Hugenholtz and T. Huber, "Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases," *Int. J. Syst. Evol. Microbiol.*, vol. 53, no. 1, pp. 289–293, Jan. 2003.

[38] K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman, "At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies," *Appl. Environ. Microbiol.*, vol. 71, no. 12, pp. 7724–7736, Dec. 2005.

[39] R. Ranjan, A. Rani, A. Metwally, H. S. McGee, and D. L. Perkins, "Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing," *Biochem. Biophys. Res. Commun.*, vol. 469, no. 4, pp. 967–977, Jan. 2016.

[40] R. Rosselló-Mora and R. Amann, "The species concept for prokaryotes," *FEMS Microbiol. Rev.*, vol. 25, no. 1, pp. 39–67, Jan. 2001.

[41] K. T. Konstantinidis and J. M. Tiedje, "Towards a genome-based taxonomy for prokaryotes," *J. Bacteriol.*, vol. 187, no. 18, pp. 6258–6264, Sep. 2005.

[42] J. Handelsman, "Metagenomics: Application of Genomics to Uncultured Microorganisms," *Microbiol. Mol. Biol. Rev.*, 2005.

[43] D. H. Parks *et al.*, "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life," *Nat. Biotechnol.*, vol. 36, no. 10, p. 996, Nov. 2018.

[44] F. W. Preston, "The Commonness, And Rarity, of Species," *Ecology*, vol. 29, no. 3, pp. 254–283, Jul. 1948.

[45] P. C. Blainey, A. C. Mosier, A. Potanina, C. A. Francis, and S. R. Quake, "Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis," *PLoS One*, vol. 6, no. 2, p. 16626, 2011.

[46] A. C. Howe, J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown, "Tackling soil diversity with the assembly of large, complex metagenomes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 13, pp. 4904–4909, Apr. 2014.

[47] F. B. Yu, P. C. Blainey, F. Schulz, T. Woyke, M. A. Horowitz, and S. R. Quake, "Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples," *Elife*, vol. 6, Jul. 2017.

[48] J. T. Staley and A. Konopka, "Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats," *Annu. Rev. Microbiol.*, vol. 39, no. 1, pp. 321–346, Oct. 1985.

[49] E. J. Stewart, "Growing unculturable bacteria," *Journal of Bacteriology*, vol. 194, no. 16. American Society for Microbiology Journals, pp. 4151–4160, 15-Aug-2012.

[50] K. Zengler *et al.*, "Cultivating the uncultured," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no.

24, pp. 15681–15686, Nov. 2002.

[51] T. Tanaka *et al.*, “A hidden pitfall in the preparation of agar media undermines microorganism cultivability,” *Appl. Environ. Microbiol.*, vol. 80, no. 24, pp. 7659–7666, Dec. 2014.

[52] K. J. Nye, D. Fallon, B. Gee, S. Messer, R. E. Warren, and N. Andrews, “A comparison of blood agar supplemented with NAD with plain blood agar and chocolated blood agar in the isolation of Streptococcus pneumoniae and Haemophilus influenzae from sputum,” *J. Med. Microbiol.*, vol. 48, no. 12, pp. 1111–1114, Dec. 1999.

[53] S. S. Epstein, “Microbial awakenings,” *Nature*, vol. 457, no. 7233. Nature Publishing Group, p. 1083, 26-Feb-2009.

[54] A. D’Onofrio *et al.*, “Siderophores from Neighboring Organisms Promote the Growth of Uncultured Bacteria,” *Chem. Biol.*, vol. 17, no. 3, pp. 254–264, Mar. 2010.

[55] J. C. Lagier, S. Edouard, I. Pagnier, O. Mediannikov, M. Drancourt, and D. Raoult, “Current and past strategies for bacterial culture in clinical microbiology,” *Clin. Microbiol. Rev.*, vol. 28, no. 1, pp. 208–236, Jan. 2015.

[56] E. A. Dinsdale *et al.*, “Functional metagenomic profiling of nine biomes,” *Nature*, vol. 452, no. 7187, pp. 629–632, Apr. 2008.

[57] R. Rosselló-Móra, “Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories,” *Environ. Microbiol.*, vol. 14, no. 2, pp. 318–334, Feb. 2012.

[58] J. M. Young, “Implications of alternative classifications and horizontal gene transfer for bacterial taxonomy,” *Int. J. Syst. Evol. Microbiol.*, vol. 51, no. 3, pp. 945–953, May 2001.

[59] S. Louca, F. Mazel, M. Doebeli, and L. W. Parfrey, “A census-based estimate of earth’s bacterial and archaeal diversity,” *PLoS Biol.*, vol. 17, no. 2, p. e3000106, Feb. 2019.

[60] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, p. 10, May 2011.

[61] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, “DADA2: High-resolution sample inference from Illumina amplicon data,” *Nat. Methods*, vol. 13, no. 7, pp. 581–583, Jun. 2016.

[62] C. Camacho *et al.*, “BLAST+: Architecture and applications,” *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–9, Dec. 2009.

[63] J. A. Raven and J. F. Allen, “Genomics and chloroplast evolution: What did cyanobacteria do for plants?,” *Genome Biology*, vol. 4, no. 3. BioMed Central, pp. 1–5, 03-Mar-2003.

[64] P. D. Schloss *et al.*, “Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities,” *Appl. Environ. Microbiol.*, vol. 75, no. 23, pp. 7537–7541, Dec. 2009.

[65] M. Hartmann, C. G. Howes, K. Abarenkov, W. W. Mohn, and R. H. Nilsson, “V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences,” *J. Microbiol. Methods*, vol. 83, no. 2, pp. 250–253, Nov. 2010.

[66] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, “VSEARCH: A versatile open source tool for metagenomics,” *PeerJ*, vol. 2016, no. 10, p. e2584, Oct. 2016.

[67] D. H. Parks, M. Chuvochina, P. A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz,

"A complete domain-to-species taxonomy for Bacteria and Archaea," *Nat. Biotechnol.*, vol. 38, no. 9, pp. 1079–1086, Sep. 2020.

[68] G. C. diCenzo and T. M. Finan, "The Divided Bacterial Genome: Structure, Function, and Evolution," *Microbiol. Mol. Biol. Rev.*, vol. 81, no. 3, Sep. 2017.

[69] "MiSeq Specifications | Key performance parameters." [Online]. Available: https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html. [Accessed: 21-Mar-2021].

[70] "NextSeq 1000 and NextSeq 2000 Sequencing Systems | Mid-throughput benchtop sequencing." [Online]. Available: https://www.illumina.com/systems/sequencing-platforms/nextseq-1000-2000.html. [Accessed: 21-Mar-2021].

[71] L. Philippot *et al.*, "The ecological coherence of high bacterial taxonomic ranks," *Nature Reviews Microbiology*, vol. 8, no. 7. Nature Publishing Group, pp. 523–529, 07-Jun-2010.

[72] P. D. Schloss, R. A. Girard, T. Martin, J. Edwards, and J. C. Thrash, "Status of the archaeal and bacterial census: An update," *MBio*, vol. 7, no. 3, Jul. 2016.

[73] J. G. Caporaso *et al.*, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. SUPPL. 1, pp. 4516–4522, Mar. 2011.

[74] G. Certini, C. D. Campbell, and A. C. Edwards, "Rock fragments in soil support a different microbial community from the fine earth," *Soil Biol. Biochem.*, vol. 36, no. 7, pp. 1119–1128, Jul. 2004.

[75] F. M. Seaton, P. B. L. George, I. Lebron, D. L. Jones, S. Creer, and D. A. Robinson, "Soil textural heterogeneity impacts bacterial but not fungal diversity," *Soil Biol. Biochem.*, vol. 144, 2020.

[76] J. St John, "jstjohn/SeqPrep: Tool for stripping adaptors and/or merging paired reads with overlap into single reads." [Online]. Available: https://github.com/jstjohn/SeqPrep#readme. [Accessed: 18-Mar-2021].

[77] "SMALT – Wellcome Sanger Institute." [Online]. Available: https://www.sanger.ac.uk/tool/smalt-0/. [Accessed: 18-Mar-2021].

[78] E. Bolyen *et al.*, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nat. Biotechnol.*, vol. 37, no. 8, pp. 852–857, Aug. 2019.

Chapter 9: Appendix

Pipeline Code

All pipeline code will be included in a forthcoming publication.

SILVA Accession Numbers of Removed Entries

Accession numbers of SILVA entries with high homology to removed eukaryotic sequences are below. Accession numbers of the 56,572 removed chloroplast sequences are available upon request.

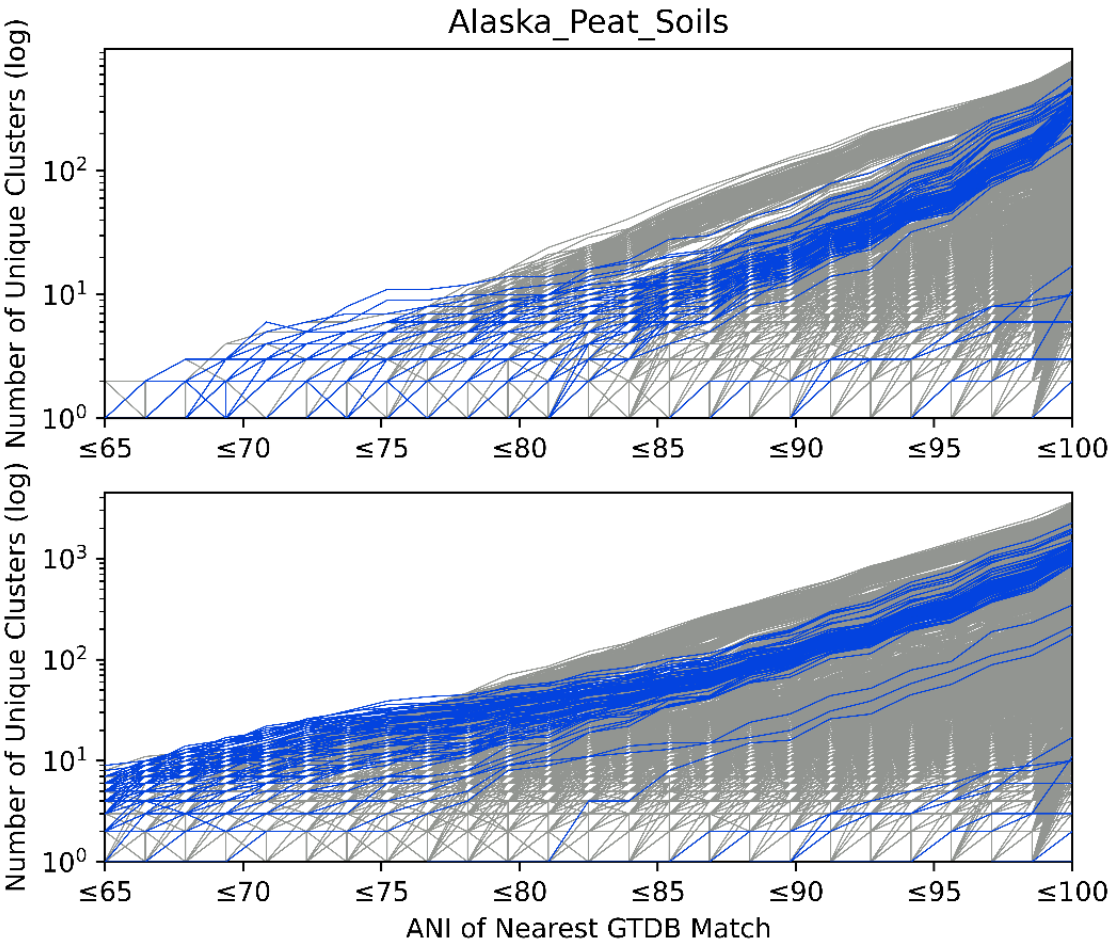
GDAH01000892.378.1918	FMPP01000014.85987.89090	CBTL0109237701.52.1488
KY764949.1.3982	JN935880.1.2291	KF738139.15057.16541
KY764921.1.3967	CMON01000010.50534.53692	GU905026.1.2297
KY764956.1.4000	ABCE01000043.54659.57378	JN935863.1.2157
KY764927.1.3979	JN935877.1.2206	EU478645.1.2163
KY764929.1.3970	ABDO02000010.33827.36483	KC353354.63785.65327
KY764931.1.3979	JN935881.1.2254	CCYC010538822.33353.35126
KY764930.1.3983	LT616956.1.2280	DQ984518.233103.235063
KY764922.1.3954	HG794416.1.1692	KP109804.1.2624
CP007479.257626.260370	GU905031.1.2316	AOTI010491221.2387.4311
JN935884.1.2303	AJ810554.1.2891	KX123363.1.2876
KY764919.1.3704	FIRE01000004.97119.99950	KR071121.97436.99296
JF731007.1.3023	JX015613.1.1495	MG996776.1.2263
CRPA01000007.15140.18272	CP010106.282981.285881	JN644756.1.2250
KX123359.1.3001	KU725488.45630.48552	DQ009461.1.2059
GU905029.1.2255	DQ009458.1.2068	GARE01001088.846.2317
EU714234.1.2252	CP002857.115758.118576	KU176938.115999.117675
EU859976.1.2259	FJ609188.1.2194	LN564836.1.1355
KP109803.1.2922	LT616955.1.2257	JN644755.1.2305
KY764920.1.3756	KU725478.45631.48554	EU937961.1.1492
CP001962.592399.595224	KX123350.1.3104	FPLP01002874.21.1502
CAQA01000070.1409.4877	GU905021.1.2288	FPLS01056723.9.1292
LS483396.118448.121363	FJ655918.1.2261	FPLS01021527.5.1367
KU725489.44345.47260	KU725492.38816.41706	JN935864.1.2290
CP001047.191404.193931	JN935878.1.2281	
CRGT01000004.76475.79474	FAOM01158642.67832.69764	

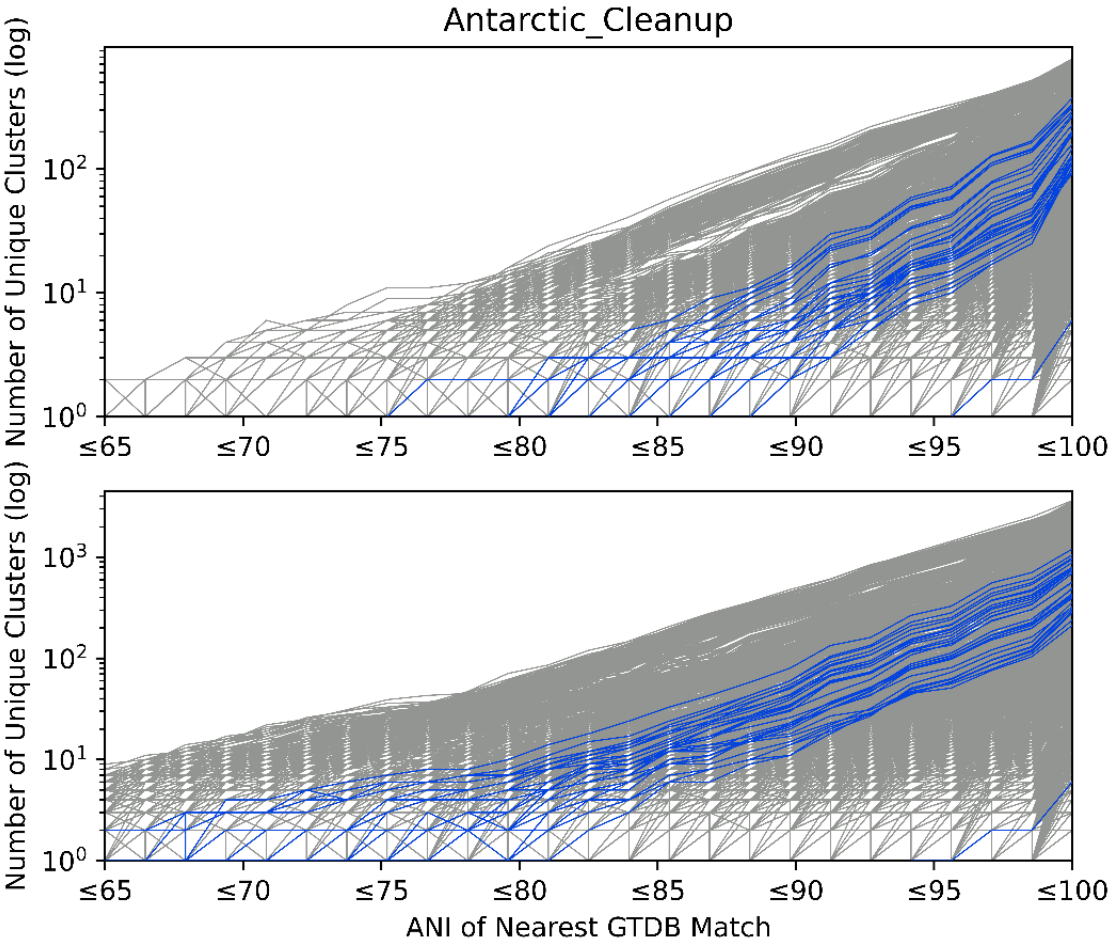
EMP Sample Information

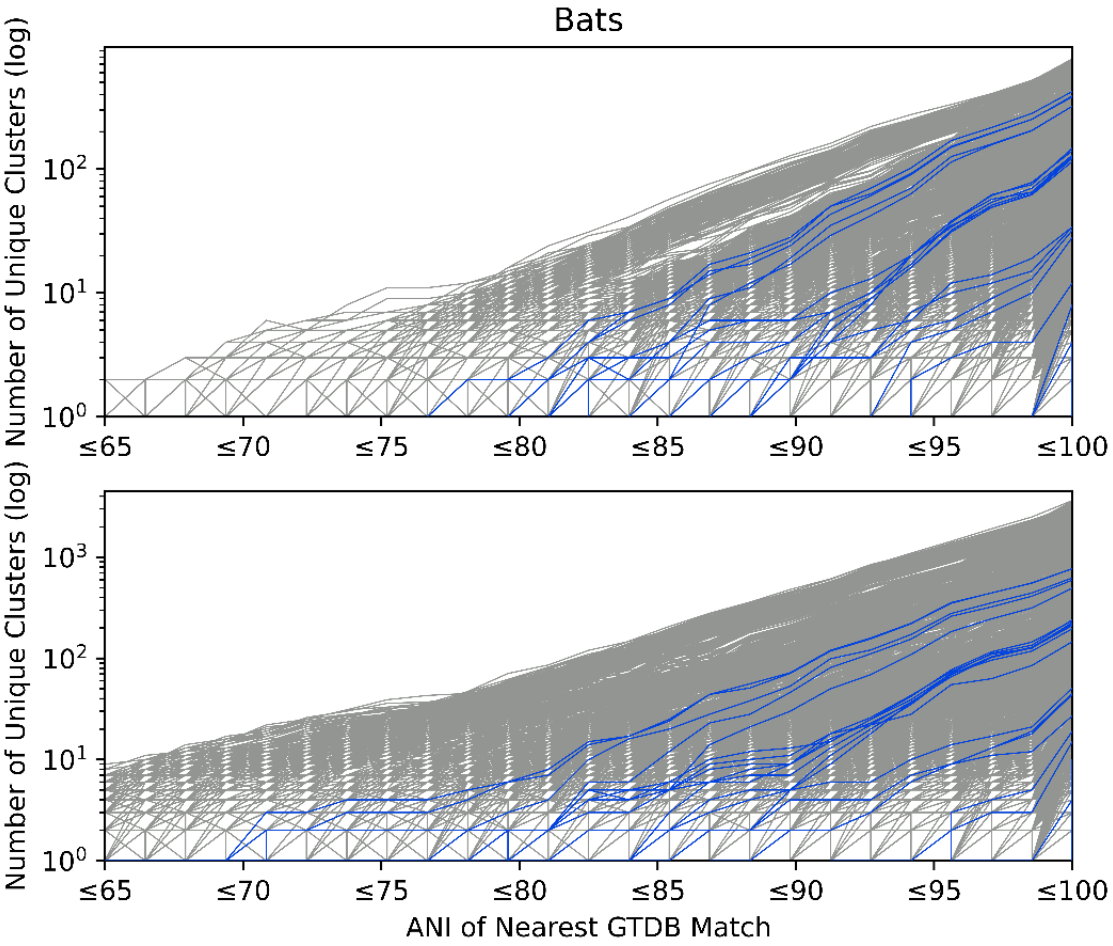
Alaska_Peat_Soils (Friedman Alaska peat soils - ID 1692; Soil)
Antarctic_Cleanup (Jurelivicius Antarctic cleanup - ID 776; Soil)
Bats (Comparison of the gut microbiome of phyllostomid bats, new world leaf nosed bats, that encompass a wide range of diets - ID 1494; Animal)
Bees (Microbiome of honey bees from Puerto Rico - ID 1064; Animal)
Bergen_Ocean_Acidification (Bergen Ocean Acidification Mesocosms - ID 1222; Water)
Canada_Waterloo (Canadian MetaMicroBiome Initiative samples from - ID 632; Soil)
Desert (Environmental metagenomic interrogation of Thar desert microbial communities - ID 829; Soil)
Fish_Slime (Microbiota of freshwater fish slime and gut from Catostomids in Colorado water system - ID 940; Animal)
Great_Lakes (Great Lake Microbiome – SID 1041; Water)
Kilauea_Soils (Kilauea geothermal soils and biofilms – ID 895; Soil)
Kohala (Hawaii Kohala Volcanic Soils – ID 1579; Soil)
Marine_Mammal_Microbiomes (Marine mammal skin microbiomes - ID 1665; Animal)
Marine_Sediment (Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin - ID 810; Soil)
Ocean_Acidification (Ocean acidification shows negligible impacts on high-latitude bacterial community structure in coastal pelagic mesocosms - ID 1235; Water)
Polluted_Polar_Sediment (Polluted Polar Coastal Sediments - ID 1198; Soil)
Rocky_Shores (The role of macrobiota in structuring microbial communities along rocky shores - ID 662; Soil)
Urban_Microbes (Urban stress is associated with variation in microbial species composition, but not richness, in Manhattan - ID 1674; Soil)
Whole_Grain_Feces (Ercolini whole grain feces - ID 1481; Animal)
Yellowstone (Yellowstone gradients - ID 925; Water)

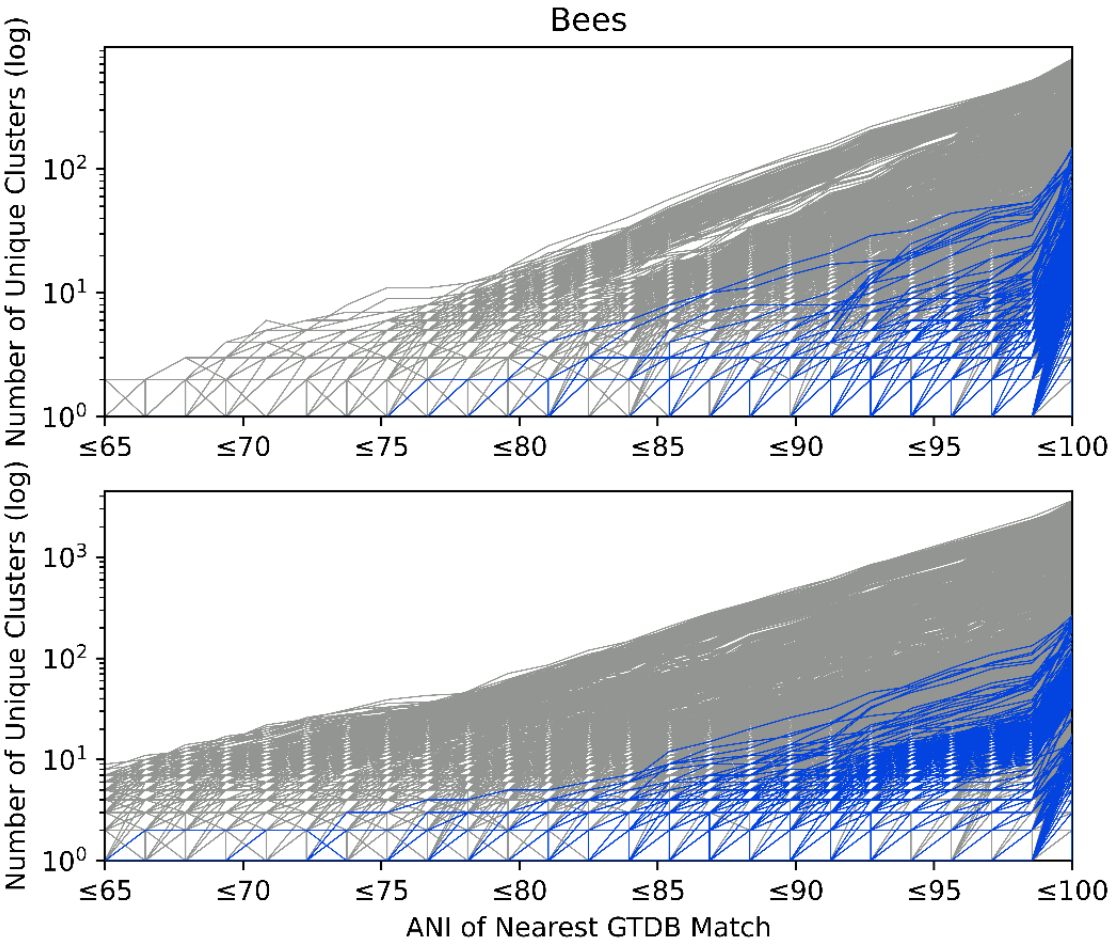
EMP Environment Gallery

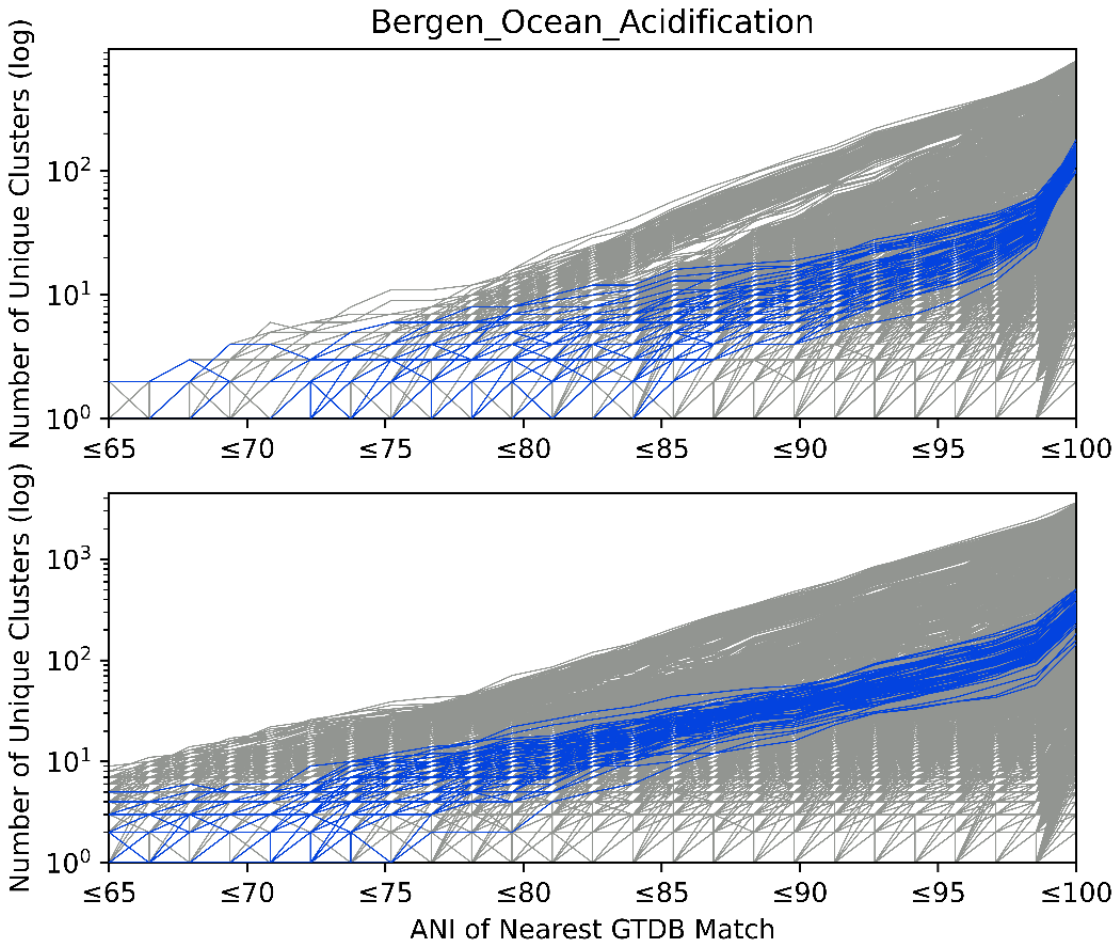
The following pages contain a gallery of figures corresponding to each of the 19 studied EMP environments. All figures are a variant of Figure 5-7. Samples from the specified environment are plotted in blue, while all other samples are plotted in grey. Additional data regarding individual samples is available upon request.

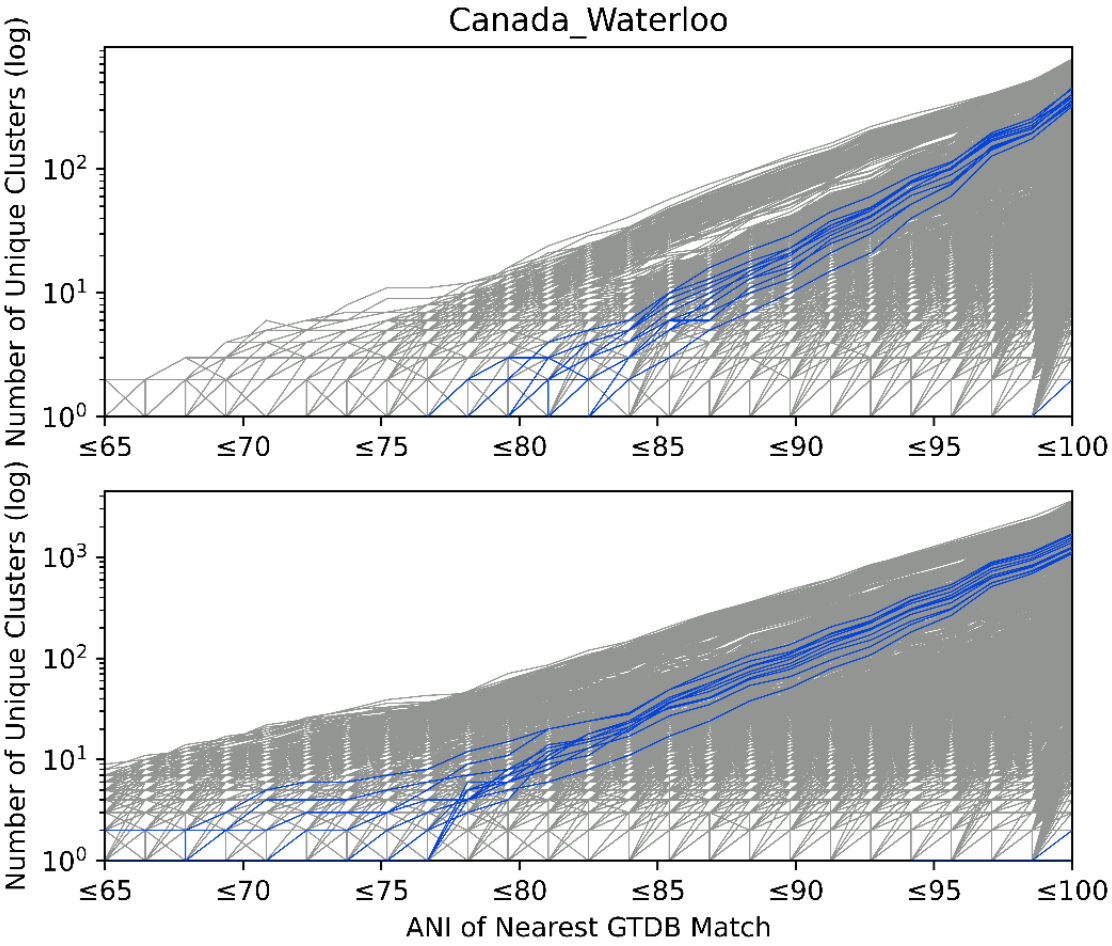

Alaska_Peat_Soils
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

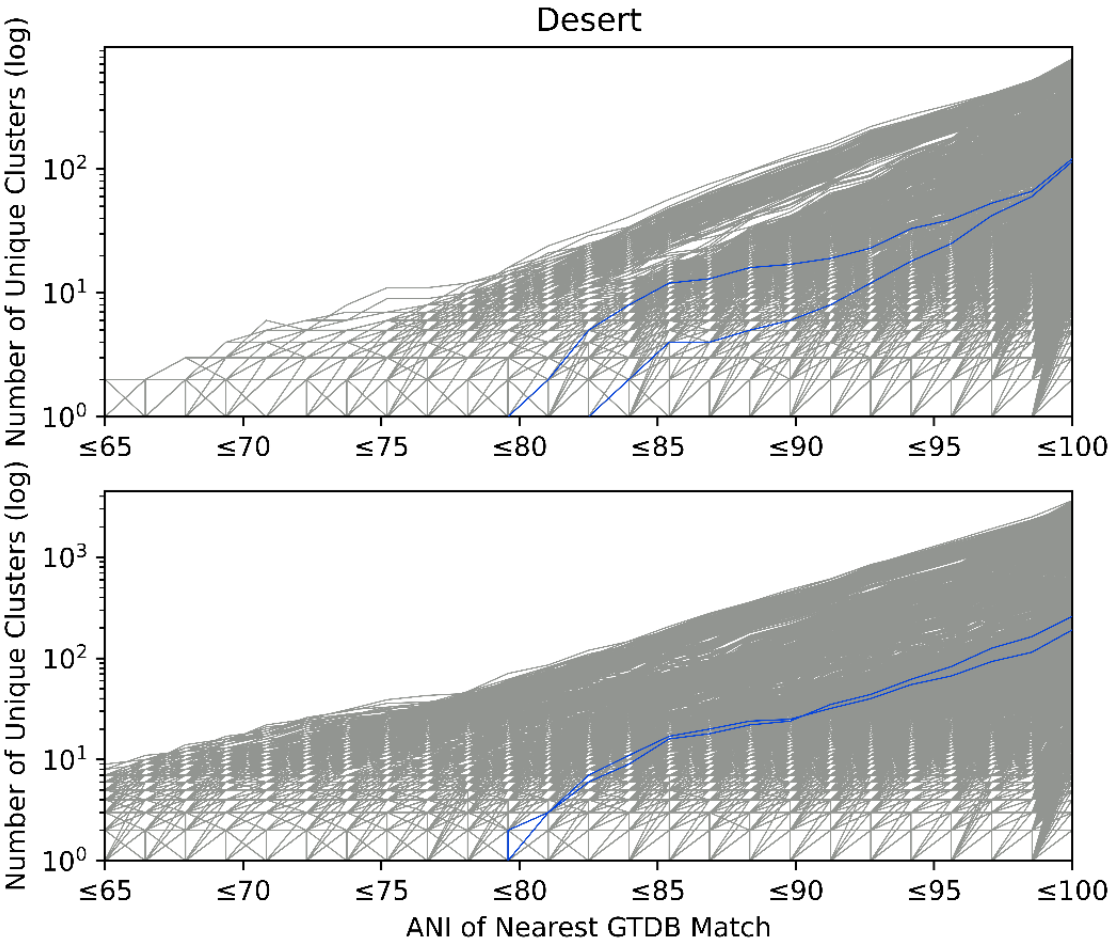

Antarctic_Cleanup
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

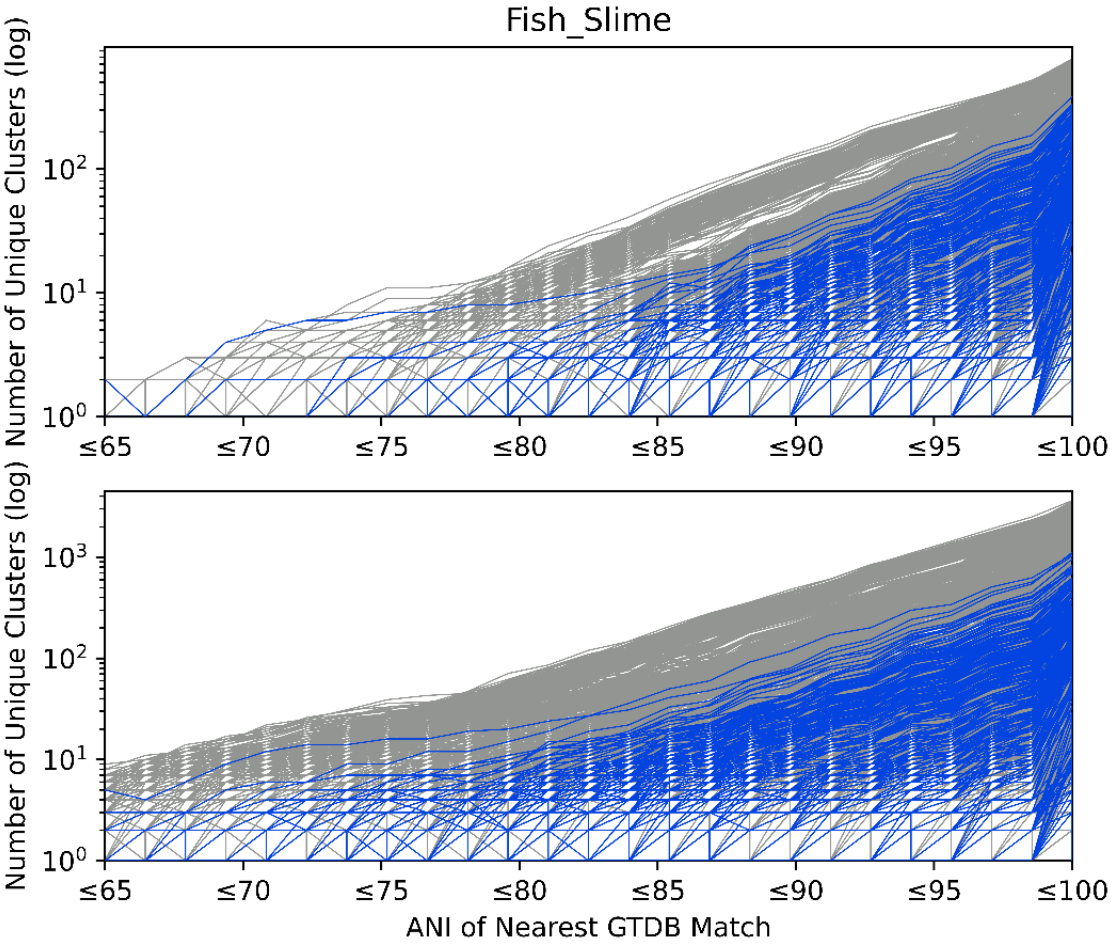

Bats
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

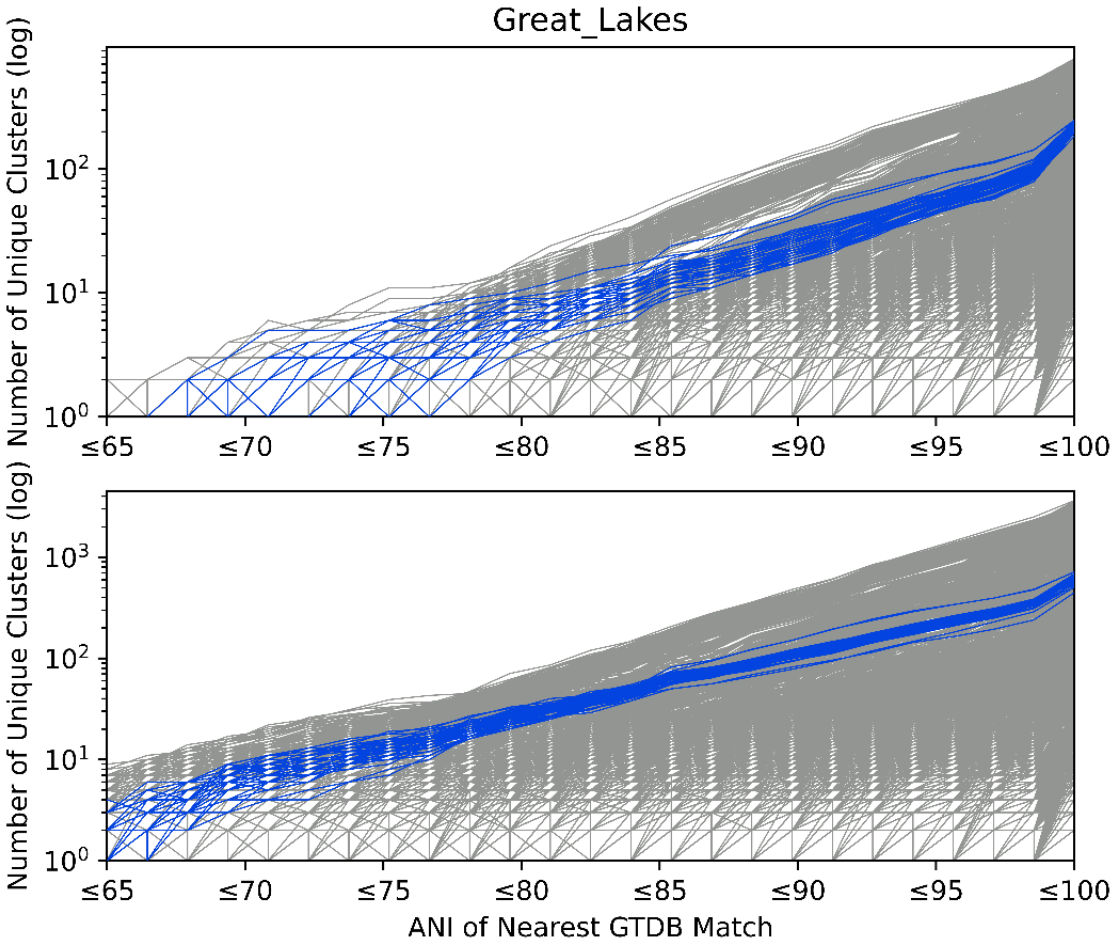

Bees
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

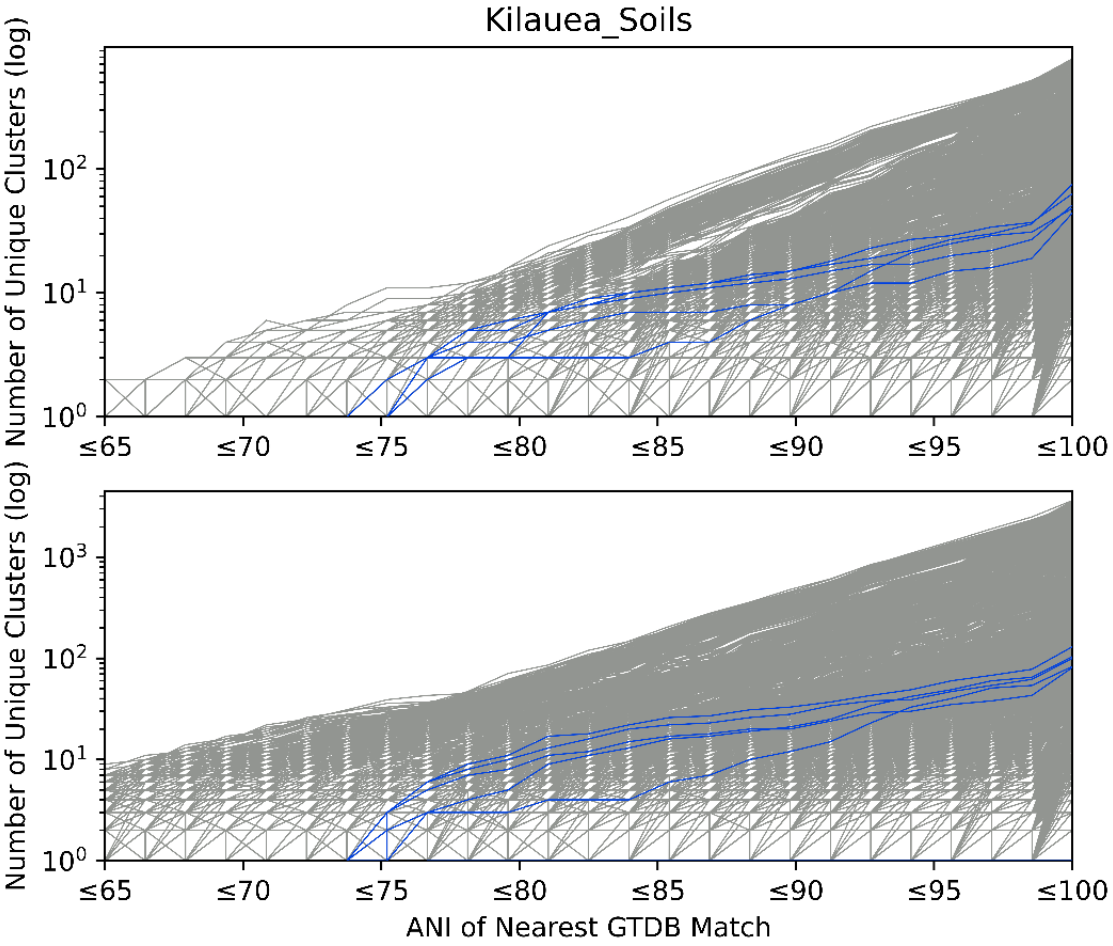

Bergen_Ocean_Acidification
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

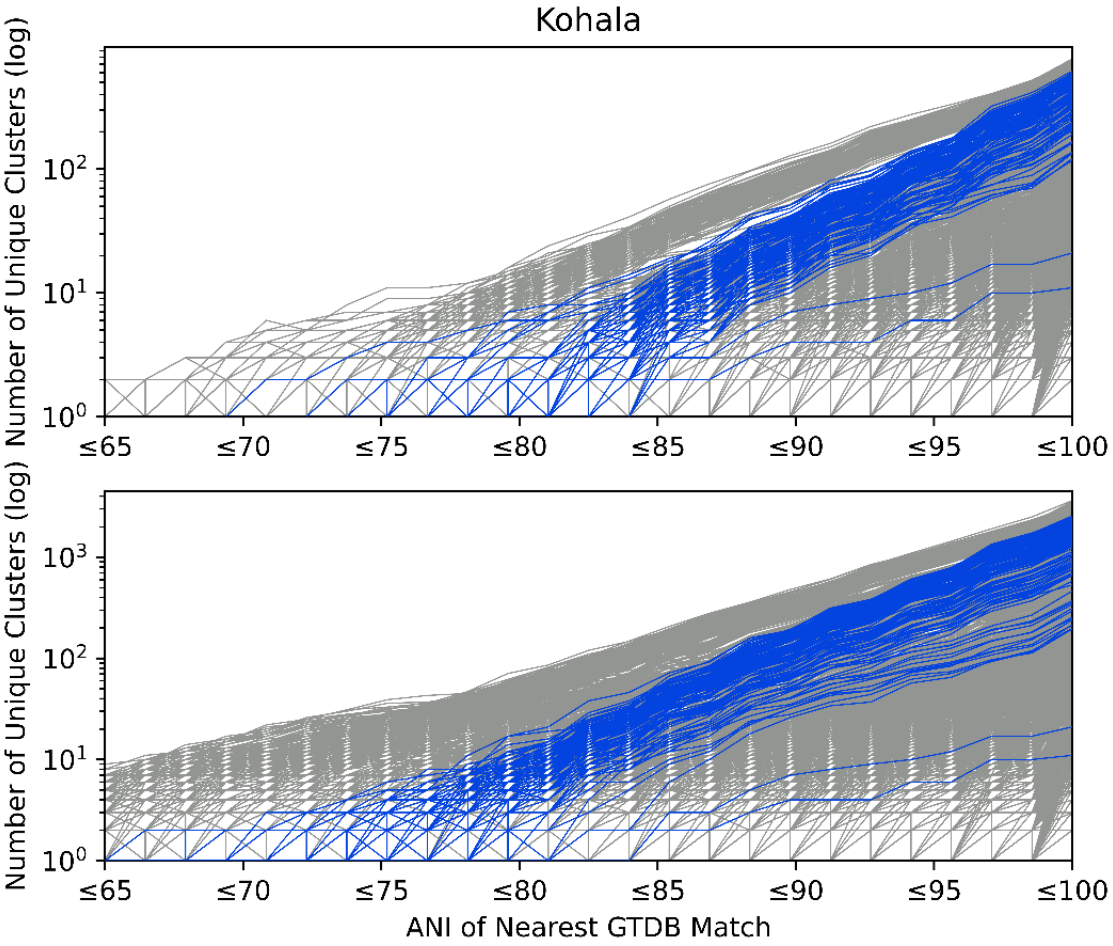

Canada_Waterloo
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

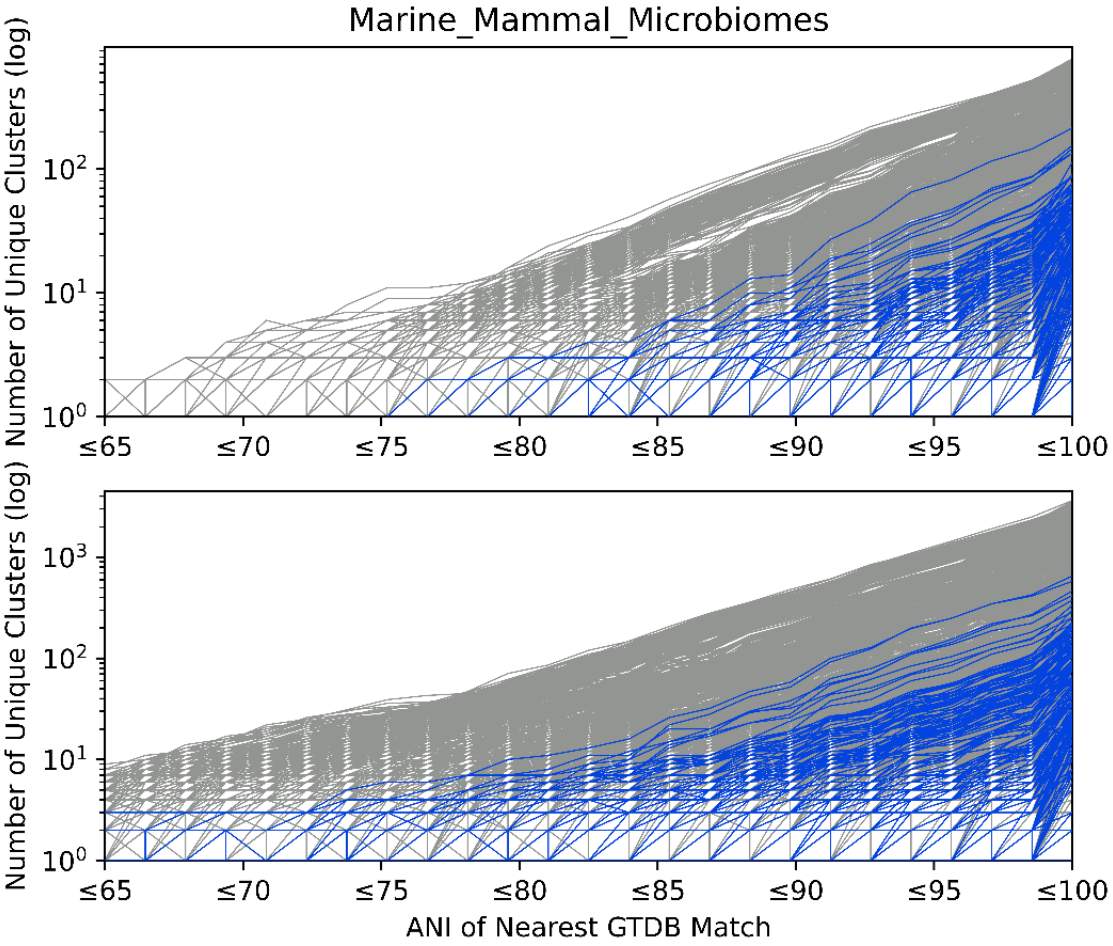

Desert
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

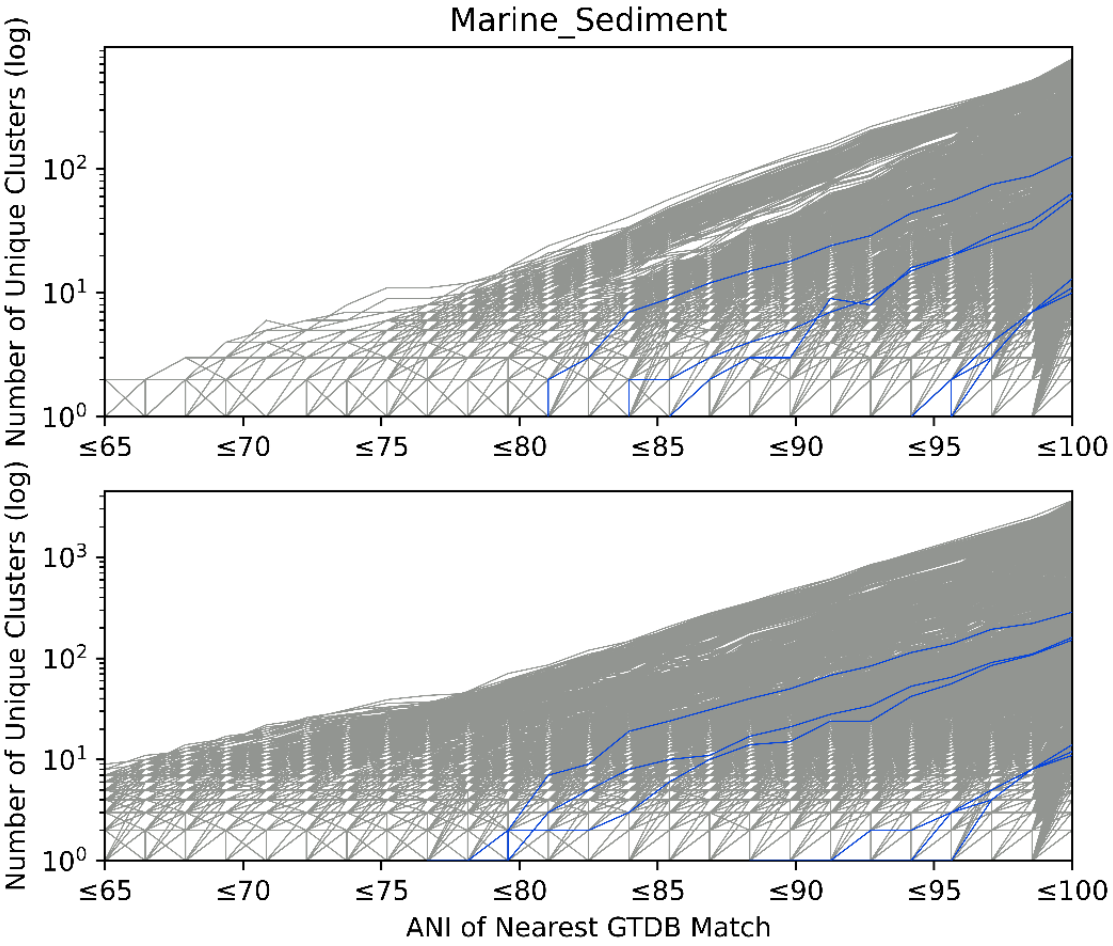

Fish_Slime
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

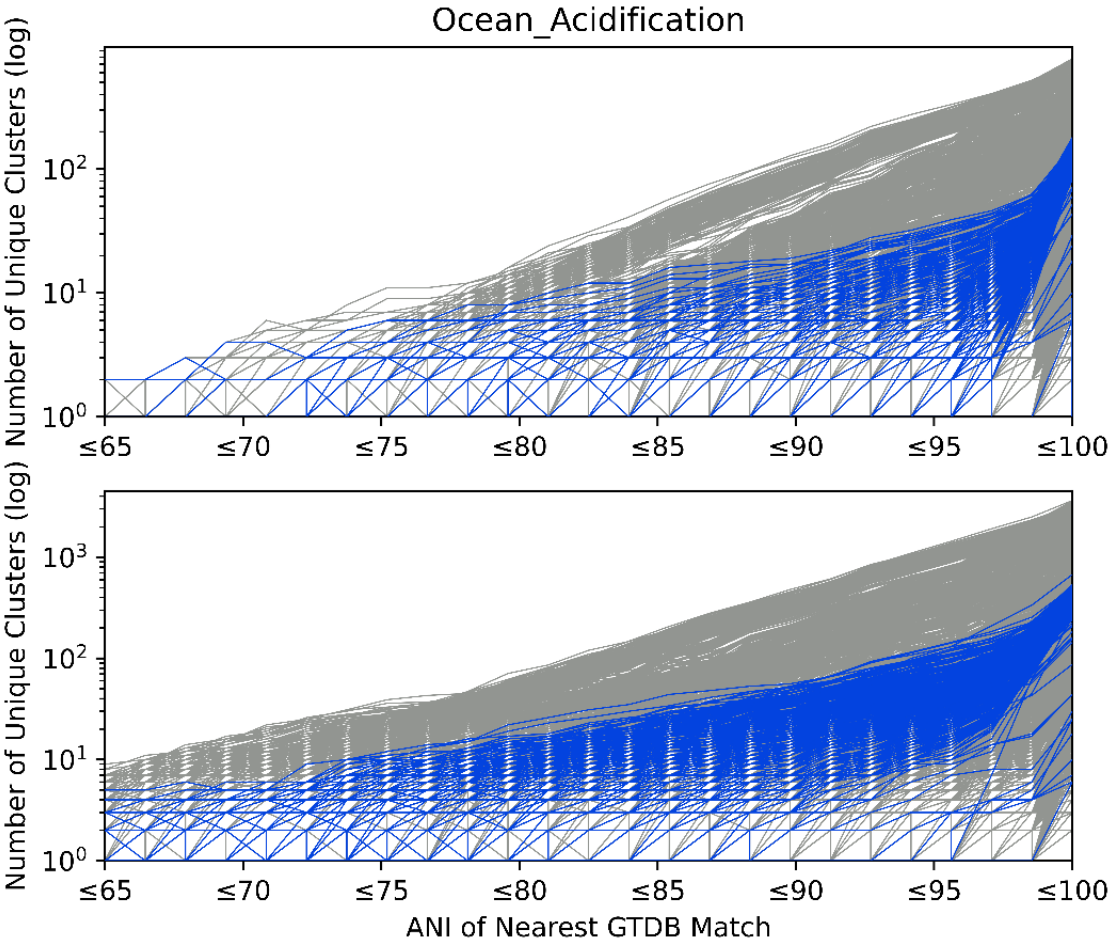

Great_Lakes
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

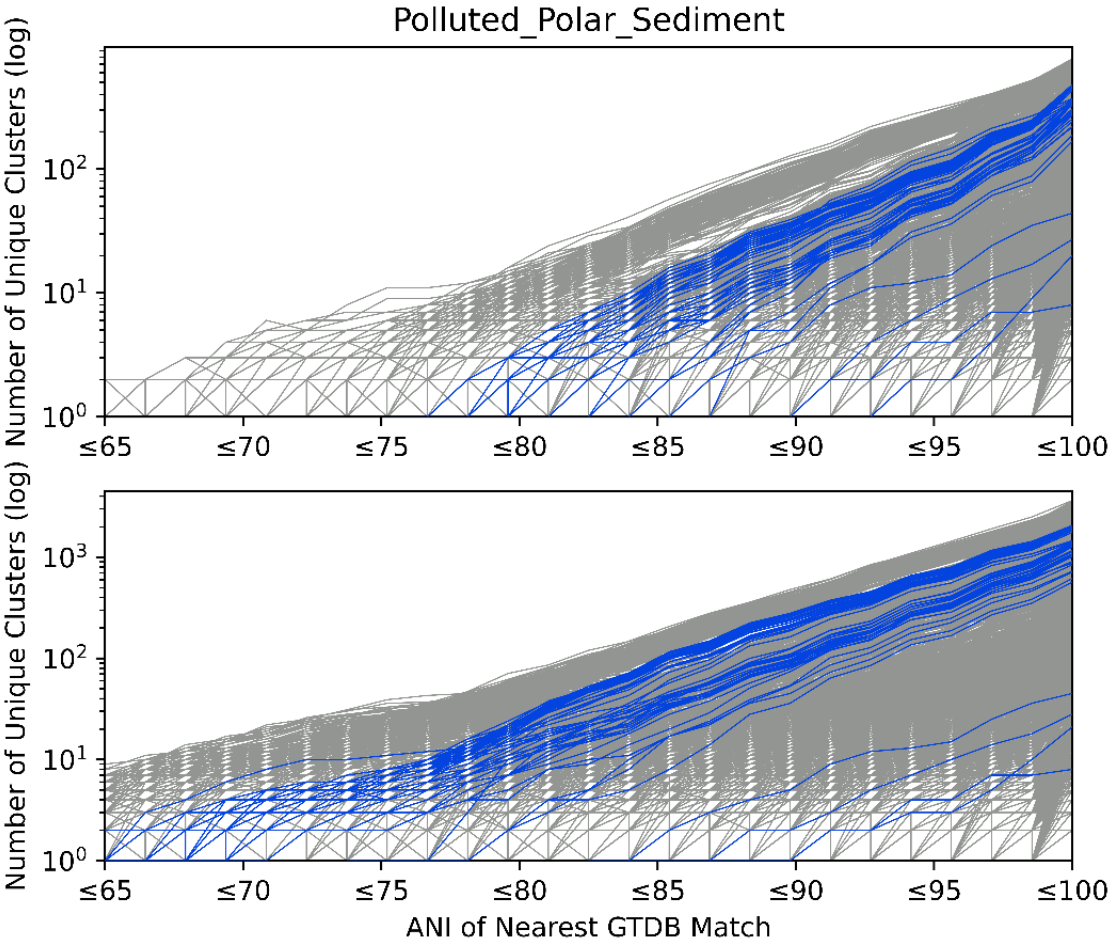

Kilauea_Soils
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

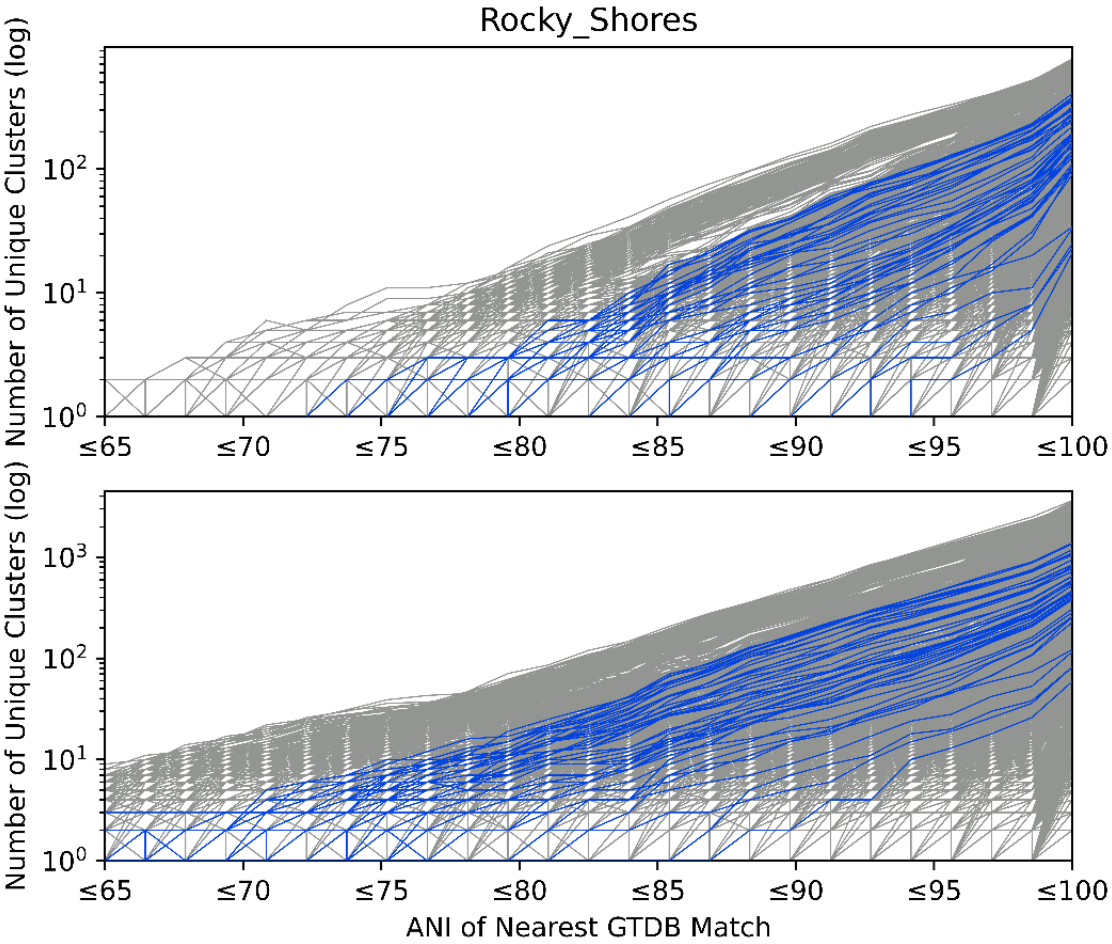

Kohala
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

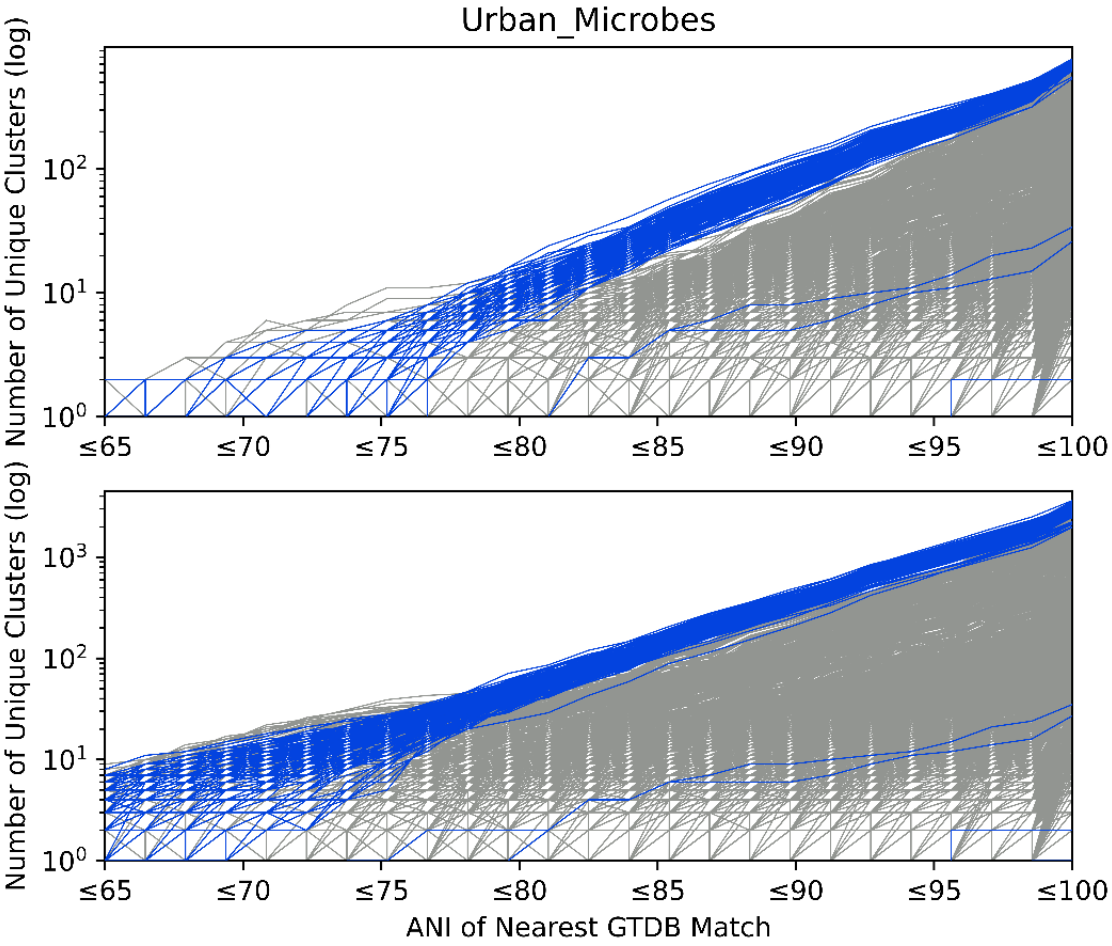

Marine_Mammal_Microbiomes
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

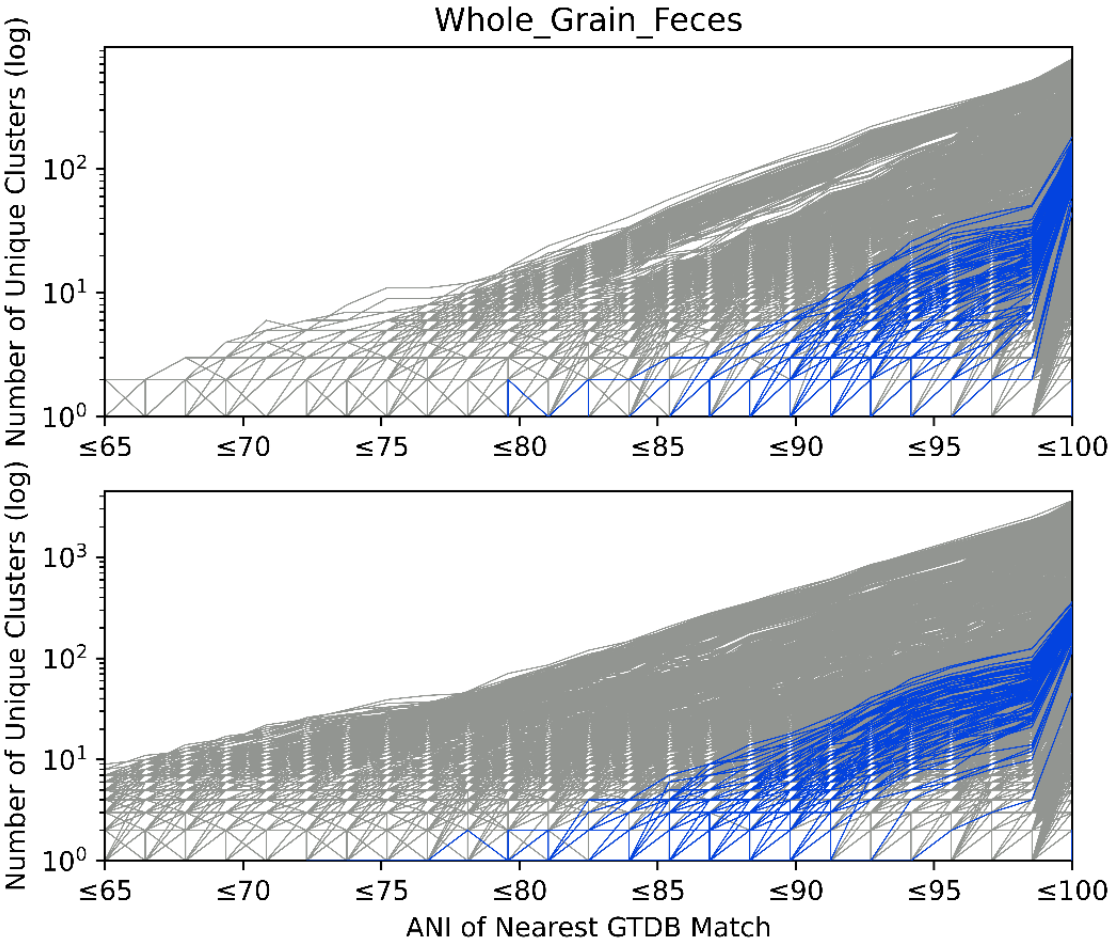

Marine_Sediment
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

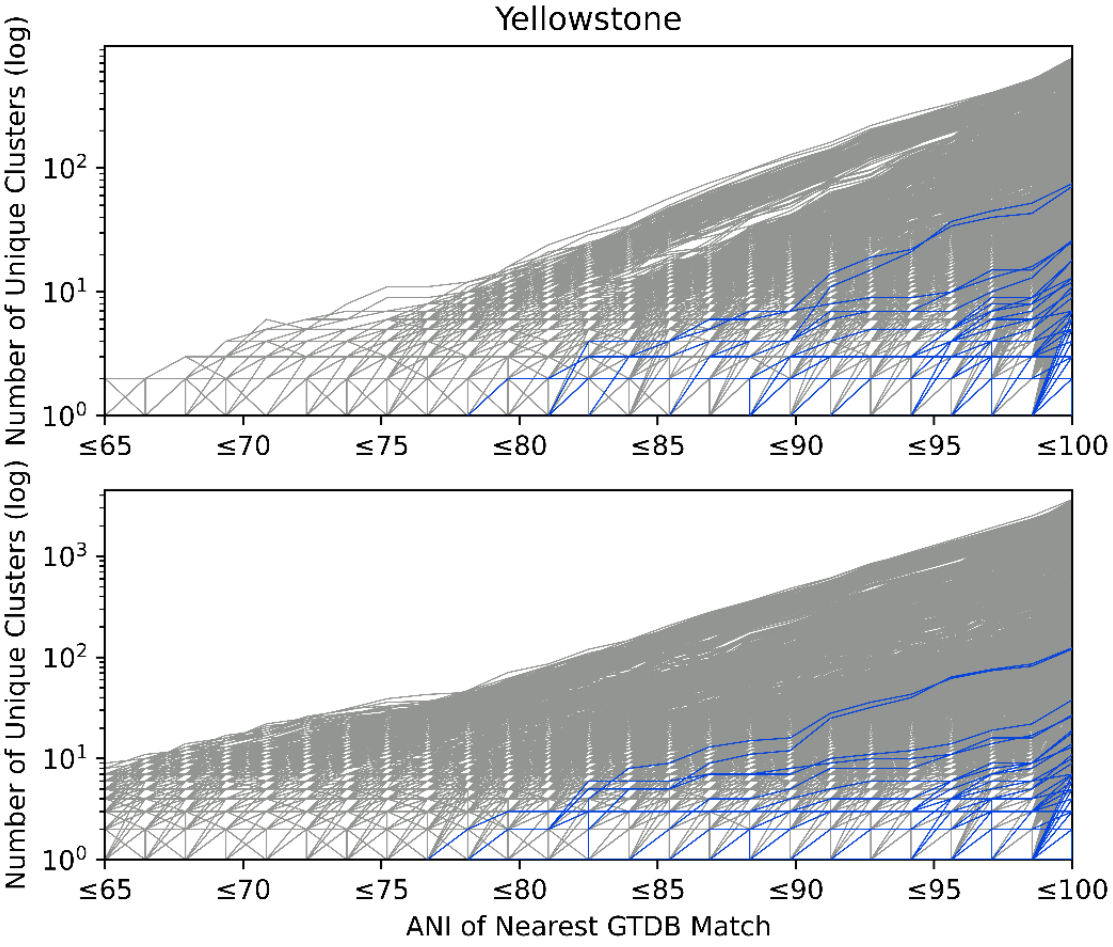

Ocean_Acidification
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match


Polluted_Polar_Sediment
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match


Rocky_Shores
Number of Unique Clusters (log)
10^0
10^1
10^2
10^3
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match


Urban_Microbes
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match


Whole_Grain_Feces
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match


Yellowstone
Number of Unique Clusters (log)
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
Number of Unique Clusters (log)
10^3
10^2
10^1
10^0
≤65
≤70
≤75
≤80
≤85
≤90
≤95
≤100
ANI of Nearest GTDB Match

In Lab Materials and Methods

Data Sourcing

Data was collected from the following locations:

11A (Elkhorn Slough; Soil)
Marsh 1-6 (Belle Isle Marsh Reservation; Soil)
Brick (Brick from The Rowland Institute; Urban)
Car_Hood (Car Hood Paint; Urban)
Car_Exhaust (Car Exhaust; Urban)
Charles1 (Sediment from The Charles River; Soil)
Charles2 (Water from The Charles River; Water)
Crab1 (Exterior Shell of a Horseshoe Crab; Crab)
Crab2 (Underside of A Horseshoe Crab; Crab)
Deer_Island (Liquid sludge from Deer Island Sewage Facility; Water)
Garden 1-4 (Indoor Garden Soils of The Rowland Institute; Soil)
Grout (Grout from the Rowland Institute; Urban)
Soil1-2 (Soil from Outside the Rowland Institute; Soil)
Window (Exterior Window of The Rowland Institute; Urban)
Hawaii S1-6 (Coastal Soils from Hawaii; Soil)

Hawaii Soils were collected on February 12, 2019, while all other samples were collected on July 3, 2018. Soil and water samples were collected in 15ml VWR centrifuge tubes. Surfaces were sampled with a sterile cotton swab soaked in MilliQ water. Swabs were deposited into centrifuge tubes after sample collection. Immediately after sample collection, samples were stored at 4C at the Rowland Institute to prevent DNA degradation. The same storage protocol was applied to Hawaii soil samples after transport to the Rowland Institute.

16S Library Preparation and Sequencing of All Sites

After swab samples were suspended in 1 mL water, DNA was extracted from all samples using the Qiagen DNeasy Powersoil Pro Kit (47014). PCR amplification of the V4 used DreamTaq Green PCR Master Mix (K1081) and the corresponding protocol. Due to the number of samples, the protocol used a custom multiplexing scheme with barcoded primers as suggested by [73]. Internal primers were manufactured by IDT and followed Earth Microbiome Project recommendations for 16S V4 targeting with the inclusion of a barcode for demultiplexing [28] (*FWD: 5' Primer ACACTCTTTCCCTACACGACGCTCTTCCGATCT Barcode NNNNN V4F GTGCCAGCMGCCGCGGTAA; REV: 5' Primer GACTGGAGTTCAGACGTGTGCTCTTCCGATCT Barcode NNNNN V4R GGACTACHVGGGTWTCTAAT)*. PCR products were cleaned using AMPure XP magnetic beads (A63880) at a 1x concentration. PCR was repeated using barcoded external primers to adapt the sequence to the flow cell (*FWD: 5' P5 AATGATACGGCGACCACCGAGATCTACAC Barcode NNNNNNNN Primer ACACTCTTTCCCTACACGACGCT; REV: 5' P7 CAAGCAGAAGACGGCATACGAGAT Barcode NNNNNNNN Primer GTGACTGGAGTTCAGACGTGTGCTCTTC)*. Then, PCR products were cleaned a second time with AMPure XP magnetic beads at a 1x concentration. Samples were sequenced on an Illumina MiSeq (2x300bp) at the Molecular Biology Core Facilities at Dana Farber.

Mini-Metagenomic Preparation of Marsh 5

The Marsh 5 sample was selected for mini-metagenomic preparation since soil and aquatic samples are known to contain the most microbial biodiversity [33]. The protocol was adapted from [47]. The sample was serially diluted to concentrations ranging from 10^{-2}-10^{-7} and aliquoted into wells for each concentration (n=16). Wells were processed using the Repli-g Single Cell Kit

(150343) at .25x scale to amplify the full genomes of any single cells present via Multiple Displacement Amplification (MDA). A portion of the MDA product was removed and run through the 16S Library Preparation protocol.

Whole Genome Library Preparation

A portion of MDA product was removed and run through the standard Nextera XT (FC-131-1024) protocol. Samples were sequenced on an Illumina MiSeq (2x300bp) at the Molecular Biology Core Facilities at Dana Farber.

Data Analysis

Paired-end reads were combined using SeqPrep [76]. Lingering PhiX reads from sequencing were removed using smalt [77] and demultiplexed using BioPython. Data was imported into Qiime2 [78] and denoised using DADA2 (*options— denoise-single –p-trim-left 0 –p-trunc-len 0*). Sequences were classified using the Greengenes classifier [25] (*options— feature-classifier classify-sklearn*).

DCR Permit

dcr
Massachusetts

RESEARCH PERMIT
Bureau of Planning, Design, and Resource Protection
251 Causeway Street, Boston, MA 02114

DCR RESEARCH ACCESS PERMIT #R-136

This permit allows access by the researcher(s) listed below to sites specified for this research, for the period and purpose stated. No unrelated access is allowed by this permit.

Researchers: Nate Cira, Shayandev Sinha, and Nkazi Nchinda of Rowland Institute at Harvard University

Project: **Characterizing New Salt Marsh Microbes**

Location: Belle Isle Marsh

Access period/Term: **October 2018 – September 2019**

1. At least 3 days prior to field work, notify Sean Riley, DCR Park Supervisor and the DCR Ecologist with date of on-site collections, via email.

2. Carry this permit & display in vehicle at all times while conducting research onsite.

3. On-site collections shall be limited to a maximum of 50 cubic centimeters of sediment or 50 ml of water to be collected from 5-10 sites within the yellow mapped areas as shown and described in the application. No on-site soil contaminants or chemicals regulated by DEP are to be analyzed under this permit.

4. Field crew shall be limited to 3 people and no heavy equipment or structures are permitted in the marsh. Impacts to vegetation and water quality shall be avoided.

5. All organisms and sediment collected at Belle Isle Marsh shall remain at Harvard Univ. or disposed of properly - these organisms are not permitted to be used for commercial use under this permit. This research is permitted for characterizing site soil properties and classifying microbes only.

6. Any state-listed rare species must be reported to NHESP & DCR Ecology Program within two weeks of observation and are not to be harmed or collected.

7. All survey data, findings, and a final report shall be submitted to the DCR Ecology Program within six months of permit expiration.

This permit may be revoked for failure to adhere to the above requirements or applicable DCR Rules and Regulations

Priscilla Geigis, Deputy Commissioner for Conservation and Resource Stewardship

****

DATE: 10/26/18 **ISSUED BY:** [signature]

COMMONWEALTH OF MASSACHUSETTS · EXECUTIVE OFFICE OF ENERGY & ENVIRONMENTAL AFFAIRS
Department of Conservation and Recreation
251 Causeway Street, Suite 600
Boston MA 02114-2119
617-626-1250 617-626-1351 Fax
www.mass.gov/orgs/department-of-conservation-recreation
Charles D. Baker, Governor
Karyn E. Polito, Lt. Governor
Matthew A. Beaton, Secretary, Executive Office of Energy & Environmental Affairs
Leo Roy, Commissioner, Department of Conservation & Recreation