Bridging Microbial Taxonomy and Function: Methods for More Comprehensive Microbiome Surveys

by

Robert Porch

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Biology

Dissertation Committee:

Brendan Bohannan, Chair

Stilianos Louca, Advisor

Lauren Hallett, Core Member

Melanie Spero, Core Member

Peter Ralph, Institutional Representative

University of Oregon

Spring 2026

© 2026 Robert Porch
This work is licensed under a Creative Commons
Attribution Non-Commercial 4.0 International License.



DISSERTATION ABSTRACT

Robert Porch

Doctor of Philosophy in Biology

Title: Bridging Microbial Taxonomy and Function: Methods for More Comprehensive Microbiome Surveys

Microbial ecology has undergone a dramatic transformation with the advent of high-throughput sequencing, particularly 16S rRNA gene sequencing and metagenomics, which together have enabled unprecedented insight into microbial diversity. While metagenomics enables direct inference of functional potential through gene content, 16S rRNA-based approaches remain widely used due to their scalability, cost-effectiveness, and well-established analytical tools. However, 16S rRNA amplicon sequencing primarily yields taxonomic composition, and consequently, linking taxonomic identity to functional potential remains a central challenge. Resolving this gap limitation requires novel, rigorous tools and methods for inferring metabolic functions from taxonomic data.

In my dissertation, I address these challenges, developing new tools and methods for linking microbial taxonomy and function, with a focus on 16S rRNA-based microbiome surveys. I build an exhaustive phenotypic database for thousands of cultured prokaryotic species, using evidence from the experimental literature to annotate the presence and absence of ecologically relevant metabolic functions. I demonstrate the utility of the database and integrate it in a novel software for inferring the functional capacities of taxa uncovered in 16S-based microbiome surveys.

In Chapter II, I develop the FAPROTAX2-db (Functional Annotation of PROkaryotic TAXa version 2) database. The database covers 13,574 bacterial and archaeal species across 101 functions. I explore the taxonomic and functional distribution of organisms represented in FAPROTAX2-db, test for biological relationships between function pairs in the database, and analyze the phylogenetic distribution of FAPROTAX2-db functions across a reference phylogeny.

In Chapter III, I develop FAPROTAX2-mapper, a software that leverages FAPROTAX2-db. This approach uses phylogenetic placement of query sequences and rigorous hidden-state prediction (HSP) algorithms to estimate the likelihood of functions for individual taxa. In contrast to FAPROTAX v1, FAPROTAX2-mapper explicitly models evolutionary relationships in its predictions, enabling more flexible inference. Cross-validation analyses demonstrate substantial improvements in predictive performance, and applications to empirical microbiome data show broader detection of functional potential.

The tools and methods I developed in my dissertation work advance the ability to link microbial taxonomy to function, and provide both a curated empirical foundation and a rigorously validated pipeline for functional inference in microbiome research.

This dissertation includes previously unpublished, coauthored material.

CURRICULUM VITAE

NAME OF AUTHOR: Robert Porch

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Rutgers University, New Brunswick, NJ, USA

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2026, University of Oregon
Bachelor of Science, Ecology, Evolution, and Natural Resources, 2020, Rutgers University

AREAS OF SPECIAL INTEREST:

Microbial ecology
Bioinformatics
Microbial evolution

PROFESSIONAL EXPERIENCE:

Teaching and Research Assistant, University of Oregon, 2021-2026
Instructor of Record, University of Oregon, 2025
NSF Research Experience for Undergraduates Scholar, University of California Berkeley, 2019
Undergraduate Tutor, Rutgers University-New Brunswick, 2018-2020
Research Assistant, Rutgers University-New Brunswick, 2018-2020

GRANTS, AWARDS AND HONORS:

Roger Locandro Award for Outstanding Student in Natural Resources, Rutgers University-New Brunswick, 2020

PUBLICATIONS:

Porch, R., Lambert, S. & Louca, S. FAPROTAX2-mapper: A command-line tool for phenotypic predictions in 16S rRNA microbiome surveys. (2026). *In submission at Nature Methods.*

Porch, R., Taylor, T.J., Nassib, A., Versteeg, L., Ellis, A.C., Keane, L.E., Roy-Hart, M., Grunkemeier, E., Stephens, P.C., Ryan, A.M., Williams, S., Qi, W., Loreto, O., Grover, L.M., Peairson, J., Simpson, E., Schnider, J.S., Smith, A.M., Pederson, B.A., Hoskins, J.A., Bonnet, L. & Louca, S. FAPROTAX2-db: A database of experimentally validated bacterial and archaeal metabolic functions. (2026). *In review at Nature Scientific Data.*

Korchagina, M.V., Mullin, C.E., Soufi, H.H., Lambert, S., Moran, I.G., **Porch, R.**, Albright, S.E., Montgomery, A.S., Jones, L.M., Malamoud, N., Jin, Q., Wood, A.M. & Louca, S. Genome-resolved metagenomic survey of 500 samples from 56 hot springs across the Western US. (2026). *Nature Scientific Data.* `https://doi.org/10.1038/s41597-026-07139-w`.

Soufi, H.H., **Porch, R.**, Korchagina, M.V., Abrams, J.A., Schnider, J.S., Carr, B.D., Williams, M.A. & Louca, S. Taxonomic variability and functional stability across Oregon coastal subsurface microbiomes. (2024). *Communications Biology*, 7(1): 1663. `https://doi.org/10.1038/s42003-024-07384-y`.

Meyer, K.M., **Porch, R.**, Muscettola, I.E., Vasconcelos, A.L.S., Sherman, J.K., Metcalf, J.E., Lindow, S.E. & Koskella, B. Plant neighborhood shapes diversity and reduces interspecific variation of the phyllosphere microbiome. (2022). *The ISME Journal* **16(5)**, 1376-1387. `https://doi.org/10.1038s41396-021-01184-6`.

ACKNOWLEDGEMENTS

I would first like to acknowledge my advisor, Stilianos Louca, for all his support over the past five years. He is a brilliant scientist who continuously pushed me to make my work as good as possible. I am grateful that he believed in me to take on such rigorous computational projects, despite coming from a different academic background. Much of the confidence I have gained as an independent researcher is owed to him. I also thank the members of my committee: Peter Ralph, Brendan Bohannan, Melanie Spero, and Lauren Hallett, the latter two of whom I had the pleasure of rotating with during my first year. To Brendan and Lauren, thank you so much for supporting me during a particularly difficult first year of graduate school. You helped me navigate an unexpected and scary landscape of uncertainty, and ultimately make the best decision of which research to pursue. An extra thanks to Lauren for also trusting in me to serve as instructor of record for her Data Visualization course, which gave me invaluable experience in teaching and pedagogy that I will take with me for the rest of my career.

Masha Korchagina and Hengameh Habibi-Soufi were the greatest peers I could have imagined working with. Having each other to lean on made graduate school all the more enjoyable. I will never forget our beautiful drives and camping trips at the Oregon coast and Yellowstone to do field work that others could only dream of.

I have an immense amount of gratitude for Sophia Lambert. She is one of the greatest mentors I could have asked for. She went above and beyond to chat with me about science and always reminded me to trust myself through the process

of research. To Sophia, thank you for helping me through so many important stages of my PhD. You are going to be an incredible PI one day.

Thank you to Kyle Meyer, one of my first research mentors when I was an undergraduate, for helping me discover my love for microbial ecology and influencing my choice to pursue graduate school. I had an incredible time doing fieldwork, DNA extractions, and sharing music with you.

I would also like to thank the instructors with whom I had the pleasure of working during my time at UO: Nicola Barber, Tobias Policha, Alice Naftaly, and Patrick Phillips. I have learned so much about effective teaching strategies thanks to my experiences with you all. A special thanks to Nicola Barber for helping me in my pursuits of a teaching career.

It has been an honor to have worked with so many curious and passionate undergraduate students, both in the classroom and in the lab. I thank all of the students whom I taught, for distracting me from the stress of my research and reminding me that learning is a collective lifelong practice. I especially thank the nearly 20 undergraduates with whom I worked on my first chapter. This work would not have been possible without their diligence, patience, and enthusiasm for microbial diversity.

To my amazing friends and roommates: Sophia Doerr, Nicole Sagui, and Anthony Dario (as well as their pets, Heidi and Ami). Living and spending my days in Eugene with you all has been unforgettable. I will always think fondly of our times together.

Lastly, I thank my family: my mom Linda, my aunt Debbie, my uncle Mark and my twin sister Angela. Leaving you and moving across the country was one of

the most difficult things I have done, but you all believed in me. I would not have finished my PhD without your constant support and love.

This dissertation is dedicated to my late grandmother, Mary Cotshott Ciambella.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 15

1.1. Overview 15

1.2. Background 16

1.2.1. Surveying microbial communities 16

1.2.2. Metagenomics 19

1.2.3. Marker genes: The 16S rRNA gene 20

1.3. Outline of dissertation 21

1.4. Summary 25

II. FAPROTAX2-DB: A DATABASE OF EXPERIMENTALLY VALIDATED BACTERIAL AND ARCHAEAL METABOLIC FUNCTIONS 26

2.1. Introduction 26

2.2. Methods 29

2.2.1. Data collection by species 29

2.2.2. Data collection by function 30

2.2.3. Computing correlations between FAPROTAX2 functions . . . 30

2.2.4. Phylogenetic analysis of FAPROTAX2 functions 31

2.3. Results 32

2.3.1. Data records 32

2.3.2. Data overview 32

2.3.3. Functional co-occurrences in FAPROTAX2-db 36

2.3.4. Phylogenetic distribution of FAPROTAX2-db functions . . . 39

2.4. Discussion 41

Chapter Page

III. FAPROTAX2-MAPPER: A TOOL FOR PHENOTYPIC PREDICTIONS IN 16S RRNA MICROBIOME SURVEYS 47

3.1. Introduction 47

3.2. Methods 50

3.2.1. Preparation of the underlying phenotypic data and reference phylogeny 50

3.2.2. The FAPROTAX2-mapper pipeline 50

3.2.3. Comparing accuracy between FAPROTAX2-mapper and FAPROTAX v1 59

3.3. Results 60

3.3.1. Comparison of functional predictions between FAPROTAX v1 and FAPROTAX2 60

3.3.2. FAPROTAX2 predictions versus metagenomes 66

3.4. Discussion 66

3.4.1. Advances in FAPROTAX2 66

3.4.2. FAPROTAX2 functions versus KEGG orthologs 68

3.4.3. Practical considerations and conclusions 70

IV. CONCLUSION 73

APPENDIX: APPENDIX 76

A.1. Additional Figures for Chapter III 76

A.2. Additional Tables for Chapter III 79

LIST OF FIGURES

Figure Page

1. Schematic diagram of microbial community surveys. 22

2. Taxonomic coverage of FAPROTAX2-db compared to FAPROTAX v1. 33

3. Overview of taxonomic and functional composition of FAPROTAX2-db. 35

4. Functional co-occurrences in FAPROTAX2-db 37

5. Reference phylogeny and distribution of FAPROTAX2-db functions . . . 40

6. Functionality of the FAPROTAX2-mapper pipeline. 52

7. Summary of internal cross-validation results for FAPROTAX2-mapper. 55

8. Functional composition of microbiome samples predicted by FAPROTAX2-mapper. 58

9. Number of ASVs with predicted presence for each function by FAPROTAX2-mapper vs. FAPROTAX v1. 62

10. Comparison of predictive accuracy between FAPROTAX v1 and FAPROTAX2-mapper. 64

A.1. Accuracy scores from internal cross-validation of FAPROTAX2. 76

A.2. Comparative predictive performance of HSP algorithms in FAPROTAX2-mapper. 77

A.3. Associations between abundances of KEGG functional groups and predicted FAPROTAX2 functions. 78

LIST OF TABLES

Table Page

1. **Phi coefficients of FAPROTAX2-db function pairs with the highest co-occurrences.** 38

2. **FAPROTAX2-db functions with the highest and lowest significant phylogenetic conservatism.** P-values were obtained using permutation tests from N=1000 permutations. . . . 41

3. **Summary of predictive accuracy of FAPROTAX2-mapper vs. FAPROTAX v1.** Scores are averaged across 10 rounds of cross-validation, as described in the Methods. Values presented in this table reflect the data visualized in Figure 10. 65

A.1. Description of microbiome samples used to demonstrate the utility of FAPROTAX2-mapper and compare functional predictions made by FAPROTAX2-mapper compared to FAPROTAX v1. 80

A.1. Table continued from previous page. 81

A.2. List of FAPROTAX2 functions mapped to KEGG functional groups and KEGG orthologous groups (KOGs). 82

A.2. Table continued from previous page. 83

A.2. Table continued from previous page. 84

A.2. Table continued from previous page. 85

CHAPTER I

INTRODUCTION

1.1 Overview

Over the past few decades, our understanding of microbial diversity has been transformed beyond what we once knew from cultures thanks to advances in sequencing technologies [1, 2, 3, 4]. This newly-appreciated diversity has opened an entire field of microbial ecological questions, including how microbial diversity shapes, and is shaped by, hosts and the abiotic environment. Many ecological questions require knowledge of both the taxonomic and functional composition of microbial communities. The taxonomic composition of microbial communities refers to the identity and relative abundance of the microorganisms present, typically resolved at levels such as family, genus, or species. This information allows researchers to characterize patterns of community structure, track shifts across environments or conditions, and infer ecological relationships such as competition, coexistence, and environmental filtering. However, taxonomic information alone does not directly reveal the functional capabilities of a community, or the processes it mediates.

In contrast, the functional composition of microbial communities describes the repertoire of metabolic traits and biochemical pathways encoded within those communities, such as polysaccharide degradation, nitrogen and methane cycling, or sulfur metabolism. Advances in shotgun metagenomic sequencing [1, 2] have enabled more direct assessment of functional composition by capturing the collective gene content of microbial assemblages, allowing researchers to link community structure to potential ecosystem processes. Despite these advances, marker-gene sequencing approaches, particularly 16S ribosomal RNA (16S

rRNA) gene sequencing, remain widely used due to their cost-effectiveness, scalability, and well-established analytical methods. However, 16S rRNA sequencing primarily provides taxonomic information a priori, as the marker gene itself is not functionally informative. As a result, connecting taxonomic composition to functional potential often requires additional inference. This highlights a central challenge in microbial ecology, and the improvement of which is the focus of my dissertation work.

A central limitation to linking taxonomic surveys to ecosystem function is the absence of a comprehensive, standardized mapping between microbial taxonomic identity to experimentally-validated phenotypes. While genomic methods continue to expand, functional annotations are often incomplete (e.g., focusing on physiological attributes rather than metabolic functions relevant to ecological processes), or are outdated. This demonstrates a need for a complementary knowledge base of cultured prokaryotic representatives, in which metabolic traits have been empirically demonstrated and can be reliably associated with taxonomic identities. Building a phenotypic database of cultured prokaryotic species addresses this gap by systematically curating presence-absence information for ecologically relevant functions across a broad taxonomic spectrum. Such a resource provides a high-confidence foundation for implementing inferential methods for linking taxonomy to function in microbiome surveys.

1.2 Background

1.2.1 Surveying microbial communities.

The surveying of microbial communities has undergone a profound transformation over the past several decades. Methods have largely shifted away from culture-based inquiry to high-throughput, sequence-based characterization of communities. Early

microbiology relied on cultivation, but it has become increasingly clear that the vast majority of microorganisms in the natural environment are not readily culturable under current laboratory conditions [5, 6]. Culture-independent approaches (e.g., marker-gene sequencing and metagenomics) have provided a promising solution, enabling researchers to characterize microbial diversity without cultivation biases. Environmental surveys of ribosomal RNA genes have revealed that cultivated organisms alone account for an incredibly small fraction of the total diversity of microbes on the planet. Indeed, newer syntheses of sequencing data have identified entire radiations of phyla of formerly unexplored microbial life [3], fundamentally reshaping our understanding of microbes and their underappreciated role in ecological processes.

Modern microbial community surveys typically begin with environmental sampling (e.g., soil, marine water, or host-associated microbiomes), followed by DNA extraction and high-throughput sequencing (Figure 1A). A dominant strategy has been the amplification and sequencing of relatively conserved marker genes. Among prokaryotes, one of the most popularly targeted marker-genes is the 16S rRNA gene. This gene encodes for a component of the 30S small subunit of the prokaryotic ribosome, and contains both conserved and hypervariable regions [7], which enables high-resolution inference across a breadth of taxa.

Sequencing of microbial communities produces large numbers of reads, which are short DNA sequences (typically a few hundred base pairs for amplicon sequencing, depending on the platform) derived from individual DNA fragments in the environmental sample. Each read represents a partial observation of a gene from a single microbial cell (or genome) present in the community. Following sequencing, reads undergo quality control, error filtering, and dereplication into a

set of unique sequences. Reads are then used to infer the taxonomic composition of the community through comparison with curated reference databases (e.g., the SILVA database and Greengenes [8, 9]).

A central challenge in these approaches is defining what constitutes a "taxon" from these aforementioned sequencing data. Often, sequences are clustered into Operational Taxonomic Units (OTUs), based on sequence similarity thresholds (commonly 97%), with each OTU serving as a proxy for a species-level grouping. This clustering step reduces sequencing noise and computational complexity, but at the cost of setting an arbitrary taxonomic resolution. More recently, methods have shifted toward identifying Amplicon Sequence Variants (ASVs) in samples [10], which resolve sequences down to single-nucleotide differences through error-correction models [11, 12, 13, 14] rather than similarity-based clustering. ASVs offer improved reproducibility and cross-study comparability, as they are not dependent on dataset-specific clustering thresholds [10, 15]. Both approaches ultimately produce abundance tables which serve as the basis for downstream ecological analyses, such as diversity estimation and ordination.

These methodological advances were motivated by foundational ecological questions that emerged over the past few decades, as molecular tools began to be applied to environmental microbiology. Seminal work emphasized the use of small-subunit rRNA genes to reconstruct microbial phylogeny directly from environmental samples [16, 17, 18, 19]. These studies revealed significant, previously unrecognized diversity in aquatic and terrestrial ecosystems. By the early 2000s, high-throughput sequencing and clone library approaches were being widely applied to address key ecological questions across diverse ecosystems, establishing microbial ecology as a sequence-dependent discipline.

1.2.2 Metagenomics.

While early surveys focused primarily on taxonomic composition via sequencing of marker genes, the development of metagenomics (i.e., the shotgun sequencing of all DNA within an environmental sample) was a major conceptual and technical advance. Unlike marker-gene sequencing, metagenomics captures the full genomic content of microbial communities, enabling direct inference of functional potential. This includes metabolic pathways, gene families, and even the reconstruction of partial or complete genomes (metagenome-assembled genomes, or MAGs).

As sequencing technologies grew sufficiently scalable and cost-effective, metagenomics became a transformative tool starting in the early to mid 2000s. Venter and colleagues [20] used metagenomic approaches in the Sargasso Sea system, uncovering hundreds of bacterial phylotypes and over one million genes that were previously uncharacterized. Similarly, a study from Francis et al. (2005) [21] identified the previously unappreciated importance of archaeal lineages in ammonia oxidation (i.e., a key first step in the nitrification process) in marine ecosystems. More environmentally agnostic studies, such as Brown et al. (2015) [22] and Hug et al. (2016) [3] used metagenomic methods to better understand uncultured lineages, identifying massive prokaryotic radiations in the tree of life. These studies have demonstrated that environmental shotgun sequencing could enable the discovery of vast reservoirs of novel genes, metabolic functions, and deep prokaryotic lineages, highlighting the degree of novelty of microbial function in natural environments.

However, metagenomics introduces additional analytical complexity. Assembly, binning, and annotation of metagenomic sequences require substantial computational resources and are sensitive to sequencing depth and community

complexity [23, 24, 25, 26]. This is especially important when regarding host-associated microbial communities. For instance, Pereira-Marques and colleagues [26] found that both host DNA and sequencing depth can dramatically affect the efficiency of microbial community profiling, particularly for less abundant microbial community members. Moreover, while metagenomics provides insight into functional potential, it does not directly measure gene expression or activity. Thus, metagenomics is often complemented with metatranscriptomics or metaproteomics for functional validation. Despite these challenges, metagenomics has become a central tool for linking microbial diversity to ecosystem function, significantly expanding the scope of microbial ecology beyond taxonomic surveys.

1.2.3 Marker genes: The 16S rRNA gene. Despite the power of metagenomics, marker-gene sequencing, particularly 16S rRNA amplicon sequencing, remains widely used in microbial ecology [27]. One primary reason for this is cost efficiency. Amplicon sequencing also requires far less sequencing depth than metagenomics, enabling the feasibility of large-scale studies with many samples. This makes it especially attractive for ecological surveys that prioritize replication, spatial coverage, or temporal resolution over genomic completeness.

In addition to cost, 16S-based methods benefit from long-established workflows and open-source analytical approaches. Pipelines such as QIIME2 [28], mothur [29], and DADA2 [12] provide robust procedures for sequence processing, taxonomic assignment, and diversity analysis. The widespread use of curated reference databases (e.g., SILVA, Greengenes) further supports consistent taxonomic annotation across studies. 16S datasets are therefore highly comparable and have accumulated into large public repositories.

Importantly, 16S sequencing is also well-suited for and popularly applied to answer many core ecological questions, including patterns of alpha and beta diversity [30, 31, 32], community turnover across environmental gradients [33, 34, 35], and responses to environmental perturbations [36, 37]. While it lacks direct functional resolution (Figure 1C), predictive frameworks (e.g., phylogenetic inference of traits) can begin to bridge this gap. Moreover, because many ecological processes are structured at higher taxonomic levels [38], 16S-based surveys often provide sufficient resolution for detecting biologically relevant patterns.

In practice, the choice between metagenomics and marker-gene approaches is inherently a trade-off between functional resolution and sampling breadth. Metagenomics is an ideal option for mechanistic and functional surveys, whereas 16S sequencing remains a scalable, cost-effective tool for broad ecological surveys. Both approaches therefore coexist as complementary strategies in modern microbial ecology, and each address distinct yet overlapping features of microbial communities.

1.3 Outline of dissertation

In my dissertation, I develop and assess the accuracy and utility of functional prediction methods that aid in marker-gene-based microbiological research. These novel approaches incorporate synthesized phenotypic information on the metabolic capacities of cultured prokaryotic organisms, and leverage this information by using rigorous phylogenetic algorithms to predict the presence of metabolic functions for user-provided query OTUs or ASVs, based on their phylogenetic relationship to reference taxa present in an underlying database. This work represents a substantial improvement in bridging the gap between microbial taxonomy and function, allowing users to explore ecological questions based on

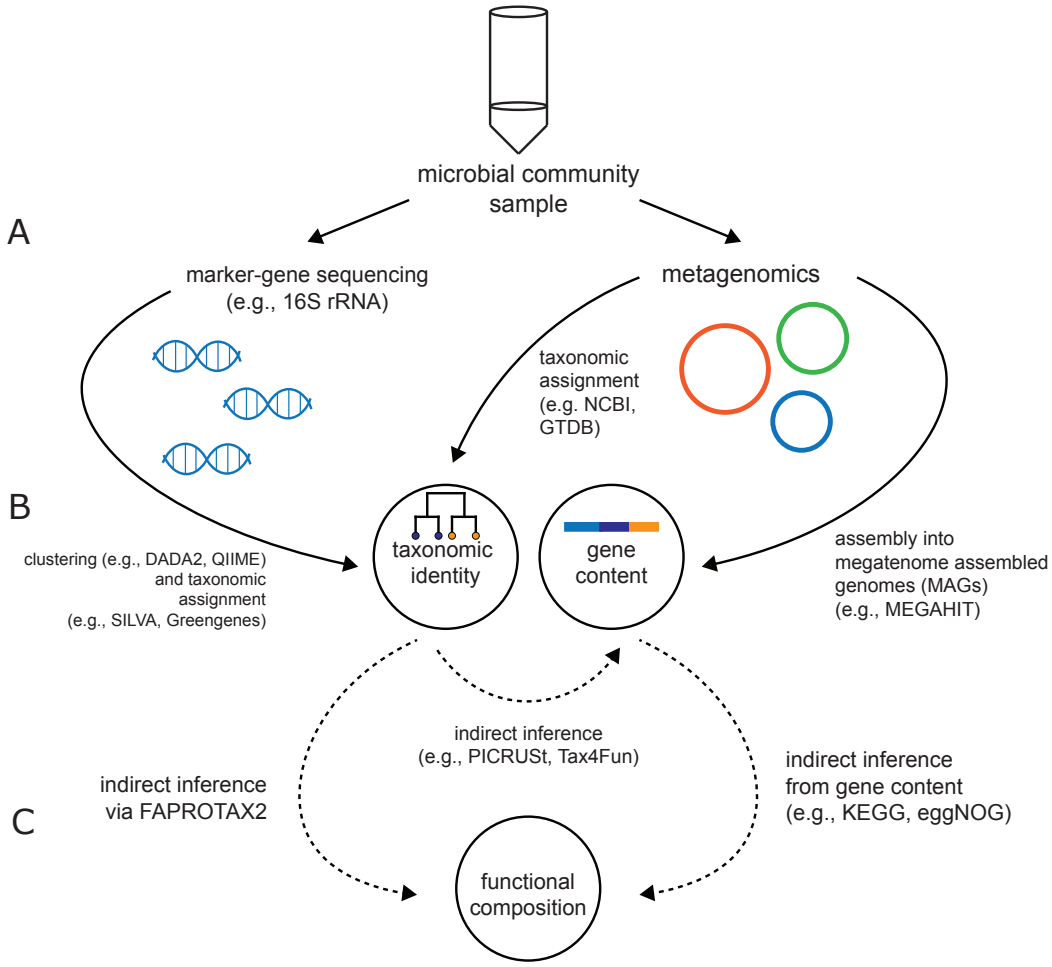


Figure 1. **Schematic diagram of microbial community surveys.** (A) Sequencing of microbial communities is typically done either through marker-gene sequencing (e.g., the 16S rRNA gene) or through metagenomics. (B) Under both marker-gene sequencing and metagenomics, sample processing and taxonomic assignment yield the taxonomic composition of the focal community. However, only under metagenomics is it possible to directly survey the gene content of the community. (C) Inference of functional composition. Under metagenomics, assignment of gene content to orthologs (e.g., KEGG, eggNOG) yields an inferred functional overview of the community. Under marker-gene sequencing, indirect inference is required, either by predicting gene content (e.g., PICRUSt or Tax4Fun), or inferring functional traits (e.g., FAPROTAX2). Resolving the latter is the focus of my dissertation work.

the composition and distribution of taxa and functions present in environmental microbiome samples.

Chapter II. FAPROTAX2-db: A database of experimentally validated bacterial and archaeal metabolic functions. In Chapter II, I develop a phenotypic database on the metabolic functions of cultured bacterial and archaeal species. The database primarily focuses on dissimilatory metabolic functions, including the capacity to cycle carbon, nitrogen, and sulfur compounds through respiration and oxidation. The data were collected from the primarily literature and syntheses of empirically validated functions, including the *Bergey's Manual of Systematics of Bacteria and Archaea* [39].

I first provide an overview of the data contained in the novel FAPROTAX2-db (Functional Annotation of PROkaryotic TAXa version 2) database. I show that many function records from the experimental literature are biased towards organisms in historically well-cultured phyla. I also demonstrate a marked improvement in taxonomic scope and resolution in FAPROTAX2-db versus FAPROTAX v1. Compared to FAPROTAX v1, all records in FAPROTAX2-db are stored at the species-level, providing improved robustness to novel taxonomic discoveries.

In Chapter II, I also explore biological relationships between functions by quantifying co-occurences across function pairs in FAPROTAX2-db. I find significantly correlated function pairs beyond those that are inseparably linked due to metabolic hierarchy or nestedness. Additionally, I explore the phylogenetic distribution of FAPROTAX2-db functions, showing considerable variation in the conservatism of functions along a reference phylogeny. The FAPROTAX2-db database is a central component to the methods I introduce in Chapter

III. However, beyond this use, FAPROTAX2-db has a variety of potential applications in microbiological research. For instance, the data can be used to model macroevolutionary processes across prokaryotic lineages to understand how the loss or acquisition of functions contributes to diversification of microbial organisms. Additionally, machine learning tools can leverage the data contained in FAPROTAX2-db to identify novel genetic proxies for these metabolic functions.

Chapter III. FAPROTAX2-mapper: A tool for phenotypic predictions in 16S rRNA microbiome surveys. In Chapter III, I leverage the phenotypic data collected in Chapter II to develop a novel functional prediction tool for 16S rRNA-based microbiome surveys. The new tool, FAPROTAX2-mapper, takes user-provided OTUs or ASVs, places the query organisms onto a reference phylogeny, and employs rigorous phylogenetic algorithms to compute the likelihood of each query organism being present or absent for functions in the underlying database.

Compared to the rigid framework of FAPROTAX v1, in which a query OTU or ASV is only predicted to exhibit a function if all members of its broader parent taxa are confirmed in the database to be present for the function, FAPROTAX2-mapper more carefully considers the evolutionary relationships between query OTUs/ASVs and reference taxa. I show that this new approach is a dramatic improvement using cross-validation and application to empirical microbiome data. My cross-validation analyses confirm substantial gains in predictive accuracy in FAPROTAX2 compared to FAPROTAX v1, and that with real microbiome samples, FAPROTAX2 predicts the presence of functions for considerably more ASVs than FAPROTAX v1. FAPROTAX2-mapper is therefore a promising new

tool for more meticulous predictions of functions in microbiome samples, with increased detection of functions and higher accuracy.

1.4 Summary

In my dissertation, I synthesize data and develop new approaches for linking microbial taxonomy to function, with a focus on 16S rRNA-based microbiome surveys. I explore the biological patterns of these data, assess the utility and accuracy of these new methods using rigorous cross-validation and application to microbiome data, and address outstanding limitations in the field regarding these approaches. My dissertation work delivers improved ways for inferring functional profiles of microbial communities, facilitating the testing of ecological, environmental, and clinical questions in microbiology.

Both Chapter II and Chapter III represent standalone, coauthored works that are pending publication. Chapter II is coauthored with Thomas J. Taylor, Arian Nassib, Liam Versteeg, Abbie C. Ellis, Lucy E. Keane, Mason Roy-Hart, Eva-Simone Grunkemeier, Piper C. Stephens, Abigail M. Ryan, Stella Williams, Wanrong Qi, Oliver Loreto, Lauren M. Grover, Jade Peairson, Elizabeth Simpson, Jared S. Schnider, Avery M. Smith, Benjamin A. Pederson, Jake A. Hoskins, Loie Bonnet and Stilianos Louca. Chapter II has been submitted and is in review in *Nature Scientific Data*. Chapter III is coauthored with Sophia Lambert and Stilianos Louca and has been submitted to *Nature Methods*.

CHAPTER II

FAPROTAX2-DB: A DATABASE OF EXPERIMENTALLY VALIDATED BACTERIAL AND ARCHAEAL METABOLIC FUNCTIONS

This chapter is coauthored with Thomas J. Taylor, Arian Nassib, Liam Versteeg, Abbie C. Ellis, Lucy E. Keane, Mason Roy-Hart, Eva-Simone Grunkemeier, Piper C. Stephens, Abigail M. Ryan, Stella Williams, Wanrong Qi, Oliver Loreto, Lauren M. Grover, Jade Peairson, Elizabeth Simpson, Jared S. Schnider, Avery M. Smith, Benjamin A. Pederson, Jake A. Hoskins, Loie Bonnet and Stilianos Louca. All authors contributed to literature search and data collection for the database. Stilianos Louca contributed to the manuscript and supervision of the data collection project. I managed the undergraduate team responsible for data collection, implemented all downstream analyses of data, and wrote most of the manuscript. The manuscript presenting this work is currently in review at *Nature Scientific Data*. This project was supported by US National Science Foundation grant 2135169.

2.1 Introduction

Bacteria and archaea (henceforth "prokaryotes" for brevity) are the most ancient and widespread organisms on Earth, occupying a remarkable range of ecological niches. Because of their diverse and often unique metabolic abilities, prokaryotes play a critical role in ecosystem functioning. Indeed, their metabolism is responsible for major biogeochemical cycles [40] and nutrient fluxes across a broad range of ecosystems [41, 42, 43]. Prokaryotic metabolism is also of increasing interest for its applications to medicine [44], bioremediation [45], and bioengineering [46].

The surge of culture-independent technologies over recent decades (e.g., 16S rRNA marker-gene sequencing and metagenomics) has transformed our understanding of the diversity of prokaryotes on the planet [47], revealing that most of earth's ecosystems are dominated by uncultured strains [5]. However, while we can now easily and cheaply identify thousands of prokaryotic species in the environment, interpreting these data requires a deep understanding of their associated ecological roles and metabolic functions. Currently, empirical studies of laboratory cultures are lagging behind the influx of genomic data [48]. While specific "proxy genes" are often used to infer phenotypes, this approach is frequently error-prone, as metabolic functions are not always explained by gene content alone [49, 50], and the prevalence of genes with unknown functions further limits predictive power [51, 52, 53]. Whether we aim to predict functions from entire genomes or infer them via taxonomic affiliation, effective application of this knowledge requires a systematic, machine-readable database of experimentally verified phenotypic data. Such a synthesis would not only provide "ground truth" necessary for predictive modeling, but could also refine gene-based inferences through large correlation analyses between gene content and phenotypes.

While useful databases of prokaryotic phenotypes currently exist [54, 55, 56, 57], they exhibit key limitations, including (1) a focus on higher-level taxa (i.e., above the species level), thus omitting important information on interspecific variation; (2) a lack of verified absence data, which adds uncertainty, introduces potential biases, and limits the robustness of functional predictions of uncultured strains; and lastly (3) a focus on morphological and physiological functions (e.g., cell shape, motility, and optimum pH and temperature) and broad-brush metabolism (e.g., metabolite utilization and production), and not on dissimilatory

metabolic functions relevant in the context of chemical fluxes (e.g., methane, nitrogen, and sulfur cycling, including chemolithotrophic metabolism). Improving upon the latter is especially important. For instance, a database with these functions can facilitate studies of the ecology of resident microbes in oceanic oxygen minimum zones, where nitrogen and sulfur metabolism drive biogeochemical fluxes [58, 59]. Similarly, a database covering methane-cycling prokaryotes (i.e., those involved in methanotrophy and methanogenesis) can strengthen investigations of nutrient cycles in the deep subsurface [60].

To address these aforementioned gaps, in this chapter we present FAPROTAX2-db, a comprehensive phenotypic database containing records for 101 empirically verified functions across 13,574 prokaryotic species in 45 phyla. We prioritized functions primarily reflecting physiological and metabolic capabilities, with a heavy emphasis on key transformations in biogeochemical cycling. Specifically, FAPROTAX2-db captures the capacity of organisms to: (1) use various electron acceptors for respiration (e.g., nitrate, nitrite, sulfate, manganese, and fumarate), (2) oxidize diverse electron donors (e.g., arsenite, methanol, ammonia, nitrite, and hydrogen), (3) perform fermentation, (4) photosynthesize (both oxygenic and anoxygenic pathways), (5) degrade environmentally and industrially relevant compounds such as pectin, lignin, and aromatic hydrocarbons, and (6) tolerate anoxic conditions (i.e., as facultative anaerobes or obligate anaerobes). We also included "umbrella" categories for higher-order functions (e.g., chemoheterotrophy and photoautotrophy) alongside non-metabolic traits of high biological importance, such as human pathogenicity and intracellular parasitism. All functional annotations in FAPROTAX2-db are based on experimental evidence synthesized from the primary literature, including the *International Journal of*

Systematic and Evolutionary Microbiology and *Bergey's Manual of Systematics of Bacteria and Archaea* [39]. In this chapter, we show summary statistics of the database (e.g., taxonomic coverage and functional frequency) and illustrate a few of its many potential applications.

2.2 Methods

2.2.1 Data collection by species. Phenotypic data were extracted manually and were limited exclusively to properly defined, cultured prokaryotic species for which the reported functions were supported by direct empirical evidence. Although the majority of functions in FAPROTAX2-db represent discrete metabolic capabilities, several categories (e.g., human pathogenicity and intracellular parasitism) were included, despite being more complex designations rather than well-defined functions per se. We deemed these essential for the utility of the database due to their extraordinary importance in clinical microbiome studies, and are henceforth also referred to as "functions" for simplicity. Records for all functions were obtained as follows. Taxonomic records were downloaded from the SILVA SSU database version 138.2 [8]. SILVA was selected not only for its extensive and curated taxonomic coverage, but also because it provides high-quality, full-length 16S rRNA gene sequences for each species, which served as the foundation for downstream phylogenetic construction. SILVA records were subset to include only Latin binomial names, as we presumed these records to indicate species that have been cultured and likely investigated experimentally. For these species, we obtained phenotypic information using the *Bergey's Manual of Systematics of Bacteria and Archaea* [39], or the primary description article documenting the discovery of the species, often published for prokaryotes in the *International Journal of Systematic and Evolutionary Microbiology*. Where

possible, we marked traits as present or absent, based on experimental evidence. For broader traits (e.g., chemoheterotrophy or photoautotrophy), we often deduced presence or absence based on implicit information. For instance, if an organism was reported to ferment using an external carbon source, then chemoheterotrophy was inferred as present. Similarly, if an organism was reported to be obligately chemoorganotrophic, then photoautotrophy was inferred as absent. Function states were marked as "unknown" in the case of a function for which no experimental evidence was identified. We henceforth refer to this method of data collection as "species-guided data collection."

2.2.2 Data collection by function. In order to mitigate the potential underrepresentation of functions that were seldom encountered in the literature used for species-guided data collection, we additionally searched for reports of species exhibiting functions, one function at a time (i.e., as opposed to scanning the literature species-by-species and searching for multiple functions at a time). In this approach, functions (and keywords related to the function) were used to identify relevant primary sources from the literature. We then examined these articles (in addition to the *Bergey's Manual of Systematics of Bacteria and Archaea* [39]) for relevant phenotypic evidence to list prokaryotic species capable of the function. For example, to curate records of plastic degradation, we queried a recently published database of plastic-degrading microorganisms [61], and cross-referenced the listed species with primary literature sources to confirm functional assignments.

2.2.3 Computing correlations between FAPROTAX2 functions. For function pairs with sufficient overlapping data, we quantified co-occurrence by computing the Phi coefficient (i.e., the Pearson correlation coefficient estimated

for binary variables) [62]. Functional annotations were converted from categorical presence/absence labels (P/A) to numeric values (1/0), respectively, and unknown states were excluded. We restricted our exploration to function pairs sharing at least 30 species with verified presence or absence records to ensure adequate statistical support for estimating association. For each qualifying function pair, Phi was computed using all species with non-missing annotations for both functions, and statistical significance was assessed via permutation testing [63, 64] with 1000 permutations to generate an empirical null distribution. The resulting symmetric matrix of Phi coefficients and associated p-values was used for downstream visualization.

2.2.4 Phylogenetic analysis of FAPROTAX2 functions.

Phylogenetic conservatism was quantified for each function across a reference tree built from matching 16S rRNA gene sequences from SILVA version 138.2 [8], using the curated SILVA multiple sequence alignments, and inferred via IQ-TREE2 [65] with automatic model selection (Figure 5). For each function, we estimated phylogenetic conservatism using the consenTRAIT depth [66], which gives the mean phylogenetic depth (τ_D) of clades in which $\geq 90\%$ of descendant tips are positive for the function. Singleton clades (i.e., single-tip clades meeting the $\geq$ 90% criterion) were retained in the calculation to capture shallow phylogenetic clustering. The statistical significance of observed τ_D values was assessed using a permutation test (N = 1000), in which function labels were randomly shuffled across the tips of the phylogeny while preserving overall function prevalence, generating a null distribution of τ_D under random phylogenetic association.

2.3 Results

2.3.1 Data records. The data contained in FAPROTAX2-db are stored and accessible on a permanent designated website (`http://www.loucalab.com/archive/FAPROTAX2`). FAPROTAX2-db consists of two tab-separated files. The first one, "FAPROTAX2-db_master.tsv," contains all cultured species records, all 101 functions, and their corresponding trait states, cited literature sources, and notes. Notes are typically provided when the trait state of a species was inferred from other evidence (e.g., phototrophy being absent, given that the organism was reported as an obligate chemoorganotroph). The second one, "function_metadata.tsv," contains metadata for all 101 functions within FAPROTAX2-db, including their definitions, and examples of evidence used from the literature to mark the functions as present or absent.

2.3.2 Data overview. The FAPROTAX2-db database contains phenotypic records for 13,574 cultured prokaryotic species across 45 phyla, 112 classes, 262 orders, 567 families, and 3,013 genera (Figure 2A). The phyla with the greatest representation in the database are Pseudomonadota, Actinomycetota, Bacillota, and Bacteroidota (formerly Proteobacteria, Actinomycetes, Firmicutes, and Bacteroidetes, respectively), with nearly 90 percent of species records belonging to these four phyla. They are also the most data-rich phyla in the database; that is, they contain the greatest number of total functional annotations in the database (Figure 2B). The high representation of these taxa is expected, as the vast majority of cultured strains fall within one of these four phyla [67, 68]. In fact, other databases of prokaryotic functions are also dominated by organisms belonging to these phyla [57, 56, 55].

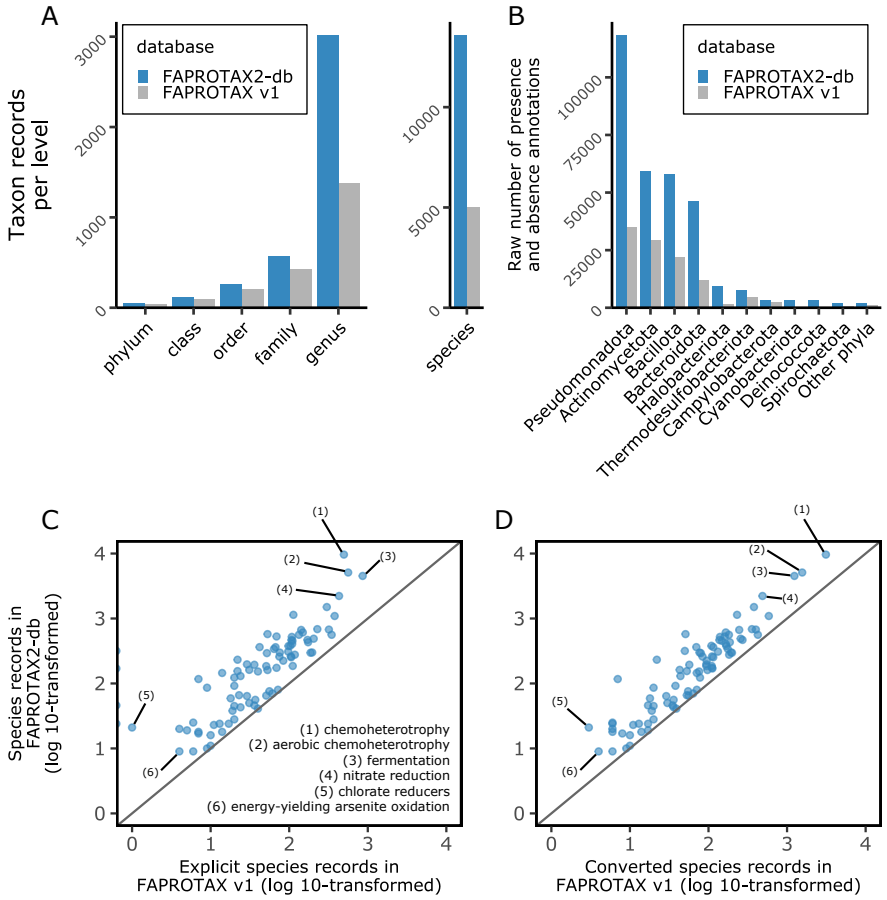


Figure 2. **Taxonomic coverage of FAPROTAX2-db compared to FAPROTAX v1.** (A) Number of unique taxa, at each rank, covered by records in FAPROTAX2-db versus FAPROTAX v1. Because functions in FAPROTAX v1 are annotated at variable taxonomic resolutions (e.g., family-, genus-, or species-level), counts at a given level include taxa represented either explicitly at that level, or implicitly through higher- or lower-level annotations. (B) Number of raw functional annotations (i.e., records of presence and absence), across different phyla. (C) Number of explicit species records for each function in FAPROTAX v1 versus FAPROTAX2-db. (D) Number of species records (after converting higher-level taxon records in FAPROTAX v1 to list all species members of that taxon) in FAPROTAX v1 versus FAPROTAX2-db. For (C) and (D), each point corresponds to a function that exists in both FAPROTAX v1 and FAPROTAX2-db. Note that in panel (C), functions with no explicit species records in FAPROTAX v1 have points that appear fixed to the vertical axis, due to an undefined log10(0).

Among the 101 functions in FAPROTAX2-db, those with the greatest richness of data (i.e., the highest number of annotations for both presence and absence) are chemoheterotrophy, aerobic chemoheterotrophy, fermentation, nitrate reduction, and gelatin degradation (Figure 3A). By contrast, the functions with the lowest number of annotations are energy-yielding arsenite oxidation, iron-oxidizing anoxygenic photoautotrophy, arsenate detoxification, methanogenesis by reduction of methyl compounds with H_2, and arsenite oxidation for detoxification (Figure 3A). These results indicate substantial heterogeneity in annotation across functions within the database.

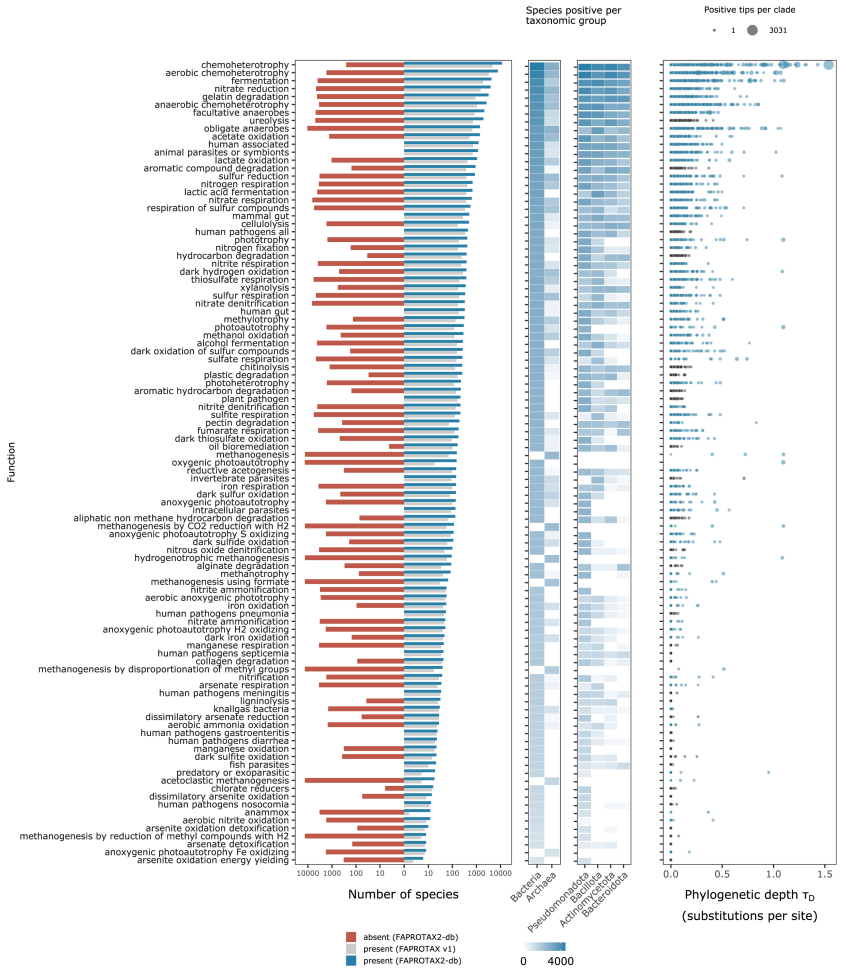


Figure 3. **Overview of taxonomic and functional composition of FAPROTAX2-db.** (A) Number of presence (teal, right-oriented) and absence (red, left-oriented) records in FAPROTAX2-db, compared to presence records in FAPROTAX v1 (light grey, right-oriented). Panels (B-D) correspond only to data in FAPROTAX2-db. (B) Number of species present for each function in the prokaryotic domains in FAPROTAX2-db. (C) Number of species present for each function in the four largest bacterial phyla in FAPROTAX2-db. (D) Dotplot depicting the phylogenetic depth τ_D (computed using the consenTRAIT depth [66]) of functions in FAPROTAX2-db. Each dot represents a positive clade (i.e., a clade in which $\geq$ 90% of tips exhibit the function), and the size of each point scales with the number of tips in the clade. Dots colored teal indicate functions with statistically significant (i.e., $p < 0.05$) phylogenetic signal (i.e., against a null expectation of randomly-assigned presence and absence annotations across the phylogeny) via permutation test with N=1000 permutations.

2.3.3 Functional co-occurrences in FAPROTAX2-db.

To demonstrate the use of FAPROTAX2-db for exploring biological relationships, we explored the co-occurrences of a selection of functions across species by computing their Phi coefficients [62], which can be interpreted as a Pearson coefficient between two binary variables. We found that, as expected, several function pairs exhibited perfect or near perfect associations ($\Phi \approx 1$), reflecting hierarchical or metabolically inseparable relationships. For example, nitrification (the aerobic oxidation of ammonia to nitrate, via nitrite) and the nested aerobic ammonia oxidation (to nitrite) were nearly perfectly correlated ($\Phi = 0.98$, $p < 0.05$) (Table 1), which is consistent with the fact that aerobic ammonia oxidation constitutes a core component of the nitrification pathway. Similarly, strong positive associations were observed among closely related but non-nested sulfur oxidation and reduction pathways (Figure 4A, Table 1). For example, dark sulfur oxidation and dark thiosulfate oxidation have $\Phi = 0.97$ ($p < 0.05$), and thiosulfate respiration and sulfate respiration have $\Phi = 0.80$ ($p < 0.05$). Similarly strong co-occurrences can be observed between respiration of nitrate and nitrite (Figure 4A). We also found that some functions related to the degradation of complex, structural polysaccharides exhibited intermediate co-occurrence. For instance, pectin and alginate degradation have $\Phi = 0.67$ ($p < 0.05$), and pectin degradation and xylanolysis have $\Phi = 0.59$ ($p < 0.05$) (Figure 4B).

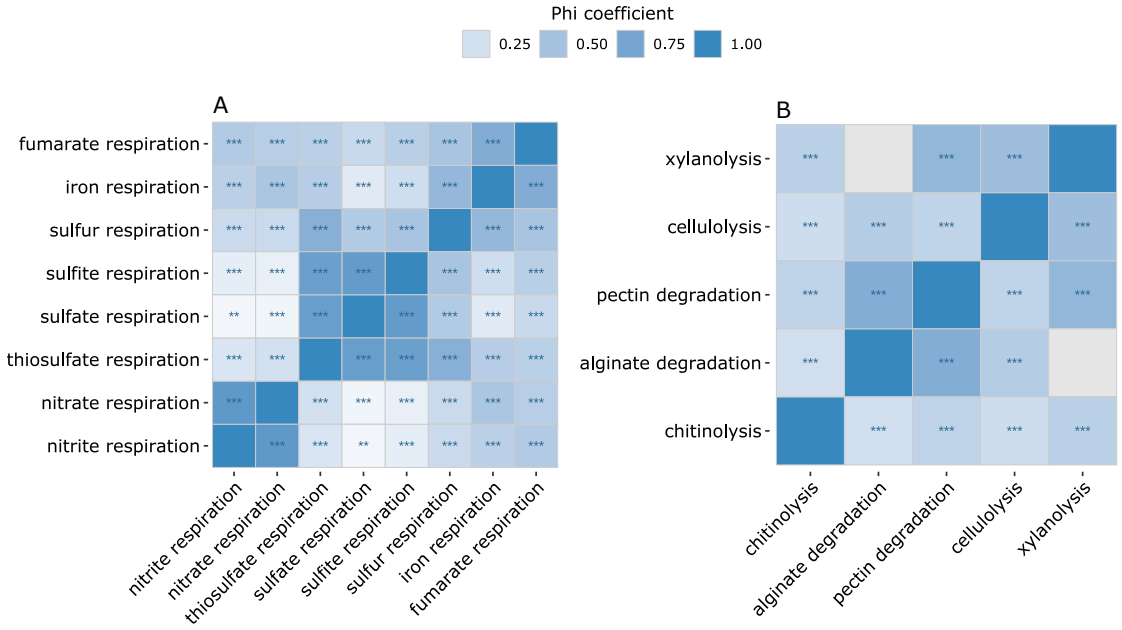


Figure 4. **Functional co-occurrences in FAPROTAX2-db.** Pairwise Phi correlation coefficients [62] of presences/absences between functions, across FAPROTAX2-db species records, for a selection of functions. (A) Functions in FAPROTAX2-db related to respiration, where the name corresponds to the terminal electron acceptor. (B) Functions in FAPROTAX2-db related to the degradation of different complex polysaccharides. For both heatmaps, function pairs were ordered via average hierarchical clustering (UPGMA [69]). Asterisks indicate statistical significance (via permutation tests with 1000 permutations) in cases where $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***).

Table 1. **Phi coefficients of FAPROTAX2-db function pairs with the highest co-occurrences.**

Function 1	**Function 2**	**Overlapping species records**	**Phi**	**p-value**
aerobic nitrite oxidation	nitrification	1661	1	< 0.001
oil bioremediation	hydrocarbon degradation	46	1	< 0.05
hydrocarbon degradation	plastic degradation	47	1	< 0.001
aromatic hydrocarbon degradation	aromatic compound degradation	376	0.994	< 0.001
nitrate ammonification	nitrite ammonification	3186	0.989	< 0.001
aerobic ammonia oxidation	nitrification	1449	0.982	< 0.001
lactic acid fermentation	alcohol fermentation	4221	0.977	< 0.001
nitrate respiration	nitrogen respiration	4143	0.969	< 0.001
dark sulfur oxidation	dark thiosulfate oxidation	542	0.969	< 0.001
dark thiosulfate oxidation	dark oxidation of sulfur compounds	355	0.961	< 0.001
anoxygenic photoautotrophy S oxidizing	anoxygenic photoautotrophy	1892	0.960	< 0.001
reductive acetogenesis	anaerobic chemoheterotrophy	410	0.959	< 0.001

2.3.4 Phylogenetic distribution of FAPROTAX2-db functions.

Among the functions mapped onto a reference phylogeny of FAPROTAX2-db species records (Figure 5), those with the greatest degree of conservatism were oxygenic photoautotrophy and methanogenesis, with τ_D of $\sim$1.10 and 0.57 16S rRNA sequence distance, respectively. These results are not entirely surprising, as oxygenic photoautotrophy is only known in prokaryotes to be present in a single phylum (i.e., Cyanobacteriota), and nearly universally present therein. Similarly, methanogenesis, which is exclusively present in the domain Archaea, was found in 4 relatively deep clades (Figure 5, Figure 3D). All other functions in FAPROTAX2-db had either intermediate phylogenetic depth - such as nitrification, obligate anaerobes, and sulfate respiration, with τ_D of 0.14, 0.12, and 0.12, respectively - or low phylogenetic depth. For example, nitrate reduction and acetate oxidation were the least conserved functions with statistically significant (i.e., $p < 0.05$) phylogenetic signal, with τ_D of $\sim$0.04 and present in 2752 and 1102 clades, respectively (Figure 5, Table 2).

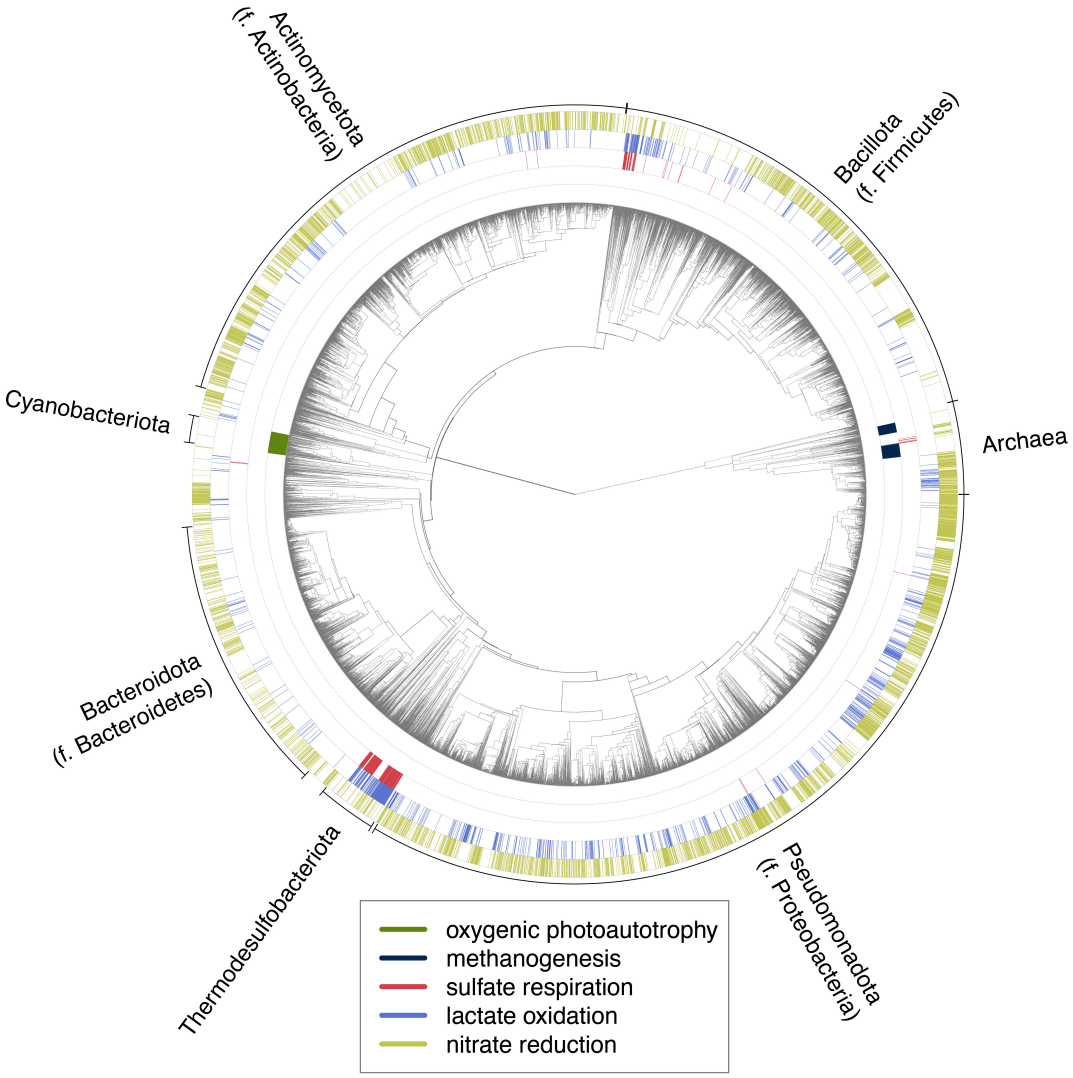


Figure 5. **Phylogenetic distribution of FAPROTAX2-db functions.** Phylogeny of FAPROTAX2-db species records, built via IQ-TREE2 [65] with automatic model selection, using full-length 16S rRNA gene sequences from the SILVA database version 138.2 [8]. The names of major taxonomic groups, including common former names if applicable, are shown along the perimeter. Tips are annotated with colored line segments where they are present for one of the 5 colored functions.

Table 2. **FAPROTAX2-db functions with the highest and lowest significant phylogenetic conservatism.** P-values were obtained using permutation tests from N=1000 permutations.

Function	**Conservatism** (τ_D)	**p-value**
oxygenic photoautotrophy	1.096	< 0.001
methanogenesis	0.571	< 0.001
methanogenesis by CO_2 reduction with H_2	0.329	< 0.001
methanogenesis by disproportionation of methyl groups	0.295	< 0.001
predatory or exoparasitic	0.182	< 0.001
...		
lactate oxidation	0.046	< 0.001
nitrogen fixation	0.045	< 0.001
nitrite denitrification	0.045	0.014
nitrate reduction	0.043	< 0.001
acetate oxidation	0.043	0.001

2.4 Discussion

In this chapter, we develop FAPROTAX2-db, a comprehensive database for 13,574 prokaryotic species across 101 functions. Compared to other previously published databases of prokaryotic organisms that generally focus on physiological or morphological attributes [55, 56, 57], FAPROTAX2-db prioritizes data on metabolic pathways relevant to ecological processes (e.g., carbon, nitrogen, and sulfur cycling) that can help provide insight into the ways in which microbes interact with and cycle nutrients in their environments [70, 71]. Further, compared to FAPROTAX v1 [54], FAPROTAX2-db incorporates these data universally at the species level, for thousands of additional organisms, and with increased representation of organisms across more environment types.

FAPROTAX2-db also represents a substantial improvement upon FAPROTAX v1 in terms of taxonomic resolution by providing functional trait

records at the species level (Figure 2C and D). For some functions, the number of explicit species records has increased by nearly an order of magnitude compared to FAPROTAX v1 (Figure 2C). In contrast to FAPROTAX v1, in which many functions were assigned at higher taxonomic ranks (e.g., genus or family), the updated framework explicitly links functions to individual species. Higher-level functional assignments, as commonly encountered in FAPROTAX v1, implicitly assume functional homogeneity within clades. Consequently, as novel prokaryotic species are described, such assumptions are increasingly at risk of violation. Species-level annotation removes this overgeneralization and prevents functional assignments from becoming obsolete as taxonomic coverage and phenotypic knowledge advances. By explicitly capturing interspecific heterogeneity, FAPROTAX2-db improves annotation accuracy and robustness to novel taxonomic discoveries.

In this chapter, we also demonstrate the dominance of well-documented phyla present in FAPROTAX2-db. Pseudomonadota, Actinomycetota, Bacillota, and Bacteroidota (formerly Proteobacteria, Actinomycetes, Firmicutes, and Bacteroidetes, respectively) make up nearly 90 percent of annotations in the database (Figure 2B). This overrepresentation inevitably exists in other culture-based prokaryotic databases [57, 56, 55] and reflects the history of culturing bias towards organisms in these phyla [67, 68]. Similarly, in addition to taxonomic bias, for the 101 functions in FAPROTAX2-db, we observed substantial variation in terms of data richness (i.e., the number of annotations of presence and absence for a function) (Figure 3).

The difference in data richness between these functions could be explained by multiple factors. First, some functions may truly be less prevalent than other

functions, for example due to physiological, ecological or evolutionary constraints, as evidenced by metagenomic and genomic studies [72, 73]. Second, there is considerable variation in the broadness of function definitions in the database, and therefore some traits have much wider qualifying windows than others. For instance, the presence or absence of chemoheterotrophy is implied simply based on an organism's ability to metabolize any external carbon source, while the presence or absence of energy-yielding arsenite oxidation is a much more narrowly defined condition. Lastly, there may be an underrepresentation of cultured organisms capable of some functions, leading to biases against these functions in the experimental literature. Indeed, culture-dependent techniques are heavily biased towards a small fraction of extant prokaryotic clades [74], due in part to the lack of appropriate laboratory-based growth conditions for many other prokaryotic groups. For example, methanogenic archaea generally require anoxic and highly reduced conditions for proper growth [75]. Recent work has highlighted that these biases may lead to a view of prokaryotic diversity that is unrepresentative of what is found in nature [72]. Future revisions of FAPROTAX2-db are expected to benefit from developments in novel culturing techniques [76, 77, 78], which provide promise for better taxonomic and functional representation of experimentally described species.

FAPROTAX2-db provides a foundation for data-driven discovery using machine learning and statistical modeling approaches. Because the database synthesizes experimentally verified metabolic functions, it can serve as a high-quality, "ground-truth" training dataset for identifying genomic predictors of functional traits. Functional genetic databases such as KEGG [79] and eggNOG [80] provide curated annotations of orthologs. However, recent work exploring

the genetics of microbial communities has highlighted the presence of yet uncharacterized gene products [53]. For instance, a large-scale metagenomic comparative study [81] identified over 100,000 novel metagenome protein families (NMPFs) across a range of environment types, with NMPFs dominating most strongly in soil, freshwater, marine, and plant-associated systems. Even in well-studied systems such as the human microbiome, up to 70 percent of proteins are uncharacterized with unknown function [82, 83, 84]. This necessitates further work to resolve the overwhelming presence of these "dark matter" gene products. FAPROTAX2-db can be used in conjunction with machine learning methods and prokaryotic genomic data to identify novel gene groups associated with metabolic functions. These novel gene groups could in principle then be incorporated into genome-resolved datasets to facilitate functional profiles from metagenomes.

FAPROTAX2-db can also be leveraged in macroevolutionary analyses that explicitly model the relationship between functional traits and lineage diversification. State-dependent speciation-extinction approaches, particularly the Binary State Speciation and Exctinction (BiSSE) Model [85], provide a means of testing whether the presence or absence of a discrete trait influences speciation and extinction rates across phylogenies. While this approach has been used in studies exploring the diversification of plant, animal and fungal lineages [86, 87, 88, 89], their application to microbial systems has been comparatively limited. By providing standardized binary annotations across thousands of prokaryotic taxa, FAPROTAX2-db enables the integration of metabolic functions into diversification analyses. FAPROTAX2-db can serve as an empirical foundation for extending state-dependent diversification models to the prokaryotic tree of life, enabling more rigorous evaluation of how metabolic innovation shapes microbial evolution.

In summary, FAPROTAX2-db is a more comprehensive and extensible database of prokaryotic metabolic functions that can aid in functional predictions. Its species-level resolution and emphasis on experimentally validated traits enable more reliable and robust functional inference. Beyond existing as a tool in the broader FAPROTAX2 bundle, FAPROTAX2-db can be leveraged to support emerging applications in metagenomics, machine learning, and macroevolutionary analysis.

All code for analysis and visualization of data in this chapter is available at `https://github.com/robertporch/faprotax2_database_manuscript`.

BRIDGE

In Chapter II, I developed FAPROTAX2-db: a large, comprehensive metabolic database for thousands of cultured prokaryotic species. I provided an overview of the functional and taxonomic records present in FAPROTAX2-db, demonstrated the methodological improvements upon the FAPROTAX v1 database (including improved taxonomic scope and resolution), and explored biological relationships in the form of functional co-occurrences and phylogenetic distribution of FAPROTAX2-db functions. In Chapter III, I explicitly integrate FAPROTAX2-db as a central tool in a new functional prediction software, FAPROTAX2-mapper. FAPROTAX2-mapper employs the FAPROTAX2-db database, an underlying reference phylogeny, and rigorous hidden-state prediction (HSP) algorithms to predict the functional states of query Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs) detected in microbiome surveys. In Chapter III, I compare the methods of FAPROTAX2-mapper to FAPROTAX v1 using cross-validation, empirical microbiome data, and metagenomic data to demonstrate marked improvements in the FAPROTAX2-mapper framework.

CHAPTER III

FAPROTAX2-MAPPER: A TOOL FOR PHENOTYPIC PREDICTIONS IN 16S RRNA MICROBIOME SURVEYS

This chapter is coauthored with Sophia Lambert and Stilianos Louca. Sophia Lambert assisted with data analysis and manuscript writing. Stilianos Louca assisted with development of FAPROTAX2, analyses of data, and supervised the project. I built the underlying pipeline preparing FAPROTAX2, implemented all downstream analyses of data, and wrote most of the manuscript. The manuscript presenting this work is currently in submission at *Nature Methods*. This project was supported by US National Science Foundation grant 2135169.

3.1 Introduction

The sequencing of marker genes, most notably the 16S ribosomal RNA gene, remains one of the most popular methods for characterizing microbial communities. However, because this method provides only taxonomic and not functional information a priori, it poses a significant limitation for many ecological studies that seek to link community composition, experimental treatments, and environmental conditions to function. Indeed, taxonomic data alone do not always provide direct insight into community functional potential, as many prokaryotic taxa share similar functional capabilities across environments (i.e., functional redundancy) [90]. While metagenomic sequencing has improved our ability to resolve microbial functional potential, it is associated with significant financial, computational, and practical constraints, thereby sustaining the use of 16S rRNA amplicon sequencing. Additionally, genes are often difficult to map to specific functions because (1) the functions of many prokaryotic genes are poorly understood [52, 51, 53], (2) some genes can be pleiotropic (i.e., involved in multiple

alternative processes) [91, 92], and (3) many phenotypes are polygenic [93], and therefore difficult to infer from gene-centric profiles alone. This limitation applies to both gene-centric metagenomics, and to gene contents estimated indirectly with tools such as PICRUSt and Tax4Fun [94, 95].

FAPROTAX (Functional Annotation of PROkaryotic TAXa) [54], first released in 2016, is a tool for predicting metabolic functions of bacteria and archaea (henceforth "prokaryotes" for brevity) in 16S rRNA-sequenced microbial communities. Compared to other tools that predict gene content [79], FAPROTAX (henceforth "FAPROTAX v1") uses phenotypic data published in the experimental literature to predict the presence of ecologically relevant metabolic functions such as nitrogen and sulfur cycling, chemolithotrophic oxidation, methanogenesis, and photosynthesis. While FAPROTAX v1 has been used in thousands of microbiome studies worldwide, it has several limitations. First, FAPROTAX v1 exclusively uses taxonomic identity as a basis for its predictions, and thus the impact of evolutionary history between prokaryotic organisms on functional trait states is not fully captured. Second, to minimize the rate of false positive annotations, FAPROTAX v1 uses a highly conservative non-adjustable approach for its functional predictions. For instance, a query operational taxonomic unit (OTU) from a microbiome survey is only predicted to have a given function if all cultured descendants of a parent taxon have been experimentally confirmed and documented to be capable of the function. Third, the presence of functions in FAPROTAX v1 is often hard-coded in the form of higher-taxon records (e.g., genus or family-level records), rendering the incorporation of knowledge from newly cultured species more difficult. Lastly, for many functions, the underlying phenotypic data in FAPROTAX v1 are biased toward prokaryotic organisms from aquatic

environments, and species from other environments, such as soil, are therefore underrepresented.

In this chapter, we present FAPROTAX2-mapper, a completely rewritten software for performing functional annotations of known and novel prokaryotes. While FAPROTAX2-mapper could in principle be used in conjunction with any underlying phenotypic database, it is optimized for - and distributed with - the FAPROTAX2-db phenotypic database (described in Chapter II [96]), as part of a broader FAPROTAX2 bundle. In addition to using a more comprehensive underlying database, FAPROTAX2-mapper significantly resolves the aforementioned limitations of FAPROTAX v1 by (1) introducing a sequence-based placement approach for adding query OTUs or amplicon sequence variants (ASVs) onto a reference phylogeny (in addition to the older approach of matching OTUs or ASVs based on taxonomy), (2) employing multiple rigorous phylogenetic prediction algorithms to infer functional trait states of query OTUs, (3) providing a greater degree of user-end flexibility for phenotypic predictions, for example by allowing probability thresholds to be adjusted for predicting functional trait states, and (4) outputting annotation probabilities using stochastic models, thus permitting a more nuanced evaluation of results by users. In this chapter, we describe in detail the improvements of FAPROTAX2-mapper, and demonstrate its utility in conjunction with FAPROTAX2-db, henceforth jointly referred to as FAPROTAX2. We specifically analyze 32 microbiome samples across diverse environments using 16S rRNA sequencing and FAPROTAX2, as well as using metagenomics for comparison. We also assess the predictive accuracy of FAPROTAX2 using cross-validation, and compare functional predictions between FAPROTAX2 and FAPROTAX v1.

3.2 Methods

3.2.1 Preparation of the underlying phenotypic data and reference phylogeny.

For functional predictions, FAPROTAX2-mapper uses phenotypic data on cultured prokaryotic organisms published in the experimental literature. Binary presence/absence data are available in FAPROTAX2 for a total of 101 functional traits, across 13,574 cultured species, with a primary focus on dissimilatory metabolic functions, including the ability to respire using various electron acceptors (e.g., nitrate, nitrite, elemental sulfur, thiosulfate, fumarate, manganese), degrade environmentally relevant compounds (e.g., xylan, chitin, lignin, and pectin), and the capacity to perform anaerobic metabolism (i.e., as either obligate or facultative anaerobes). These data were collected at the species level, while generally favoring type strains. Thus, for each species and each function, a state of "presence", "absence" or "unknown" are recorded. A detailed description of this database, which is released as a standalone resource, is given in Chapter II.

Species were subsequently matched to taxonomic and sequence records in the SILVA v138.2 16S SSU database [8]. The full-length SILVA 16S rRNA sequences for each matched species were used to construct a phylogenetic tree via IQ-TREE2 [65] with automatic model selection. This tree serves as the reference phylogeny on which user-provided query OTUs are placed prior to functional prediction in the FAPROTAX2-mapper pipeline. User-provided OTUs are either matched to SILVA taxonomies or, under sequence-based placement, are aligned against SILVA 16S rRNA sequences prior to phylogenetic placement.

3.2.2 The FAPROTAX2-mapper pipeline.

FAPROTAX2-mapper uses the FAPROTAX2 database (Chapter II, [96]) and reference phylogeny to

functionally annotate a user-provided set of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs). FAPROTAX2-mapper can optionally predict the relative abundances of functions in one or more microbiome samples, given user-provided OTU abundances for each sample as an input (Figure 6A). FAPROTAX2-mapper places query OTUs onto its reference phylogeny using either taxonomy- or sequence-based placement approaches. Under taxonomy-based placement, the user provides the taxonomic identities of query OTUs, and each OTU is then placed onto the reference tree by matching the OTU's taxonomy with the reference taxonomies at the lowest possible level. Under sequence-based placement, the user provides representative partial or full-length 16S rRNA gene sequences for the query OTUs, which are then placed on the reference tree as follows. OTUs are aligned against FAPROTAX2 references via MAFFT [97], and subsequently placed onto the FAPROTAX2 reference phylogeny with the third-party phylogenetic placement program EPA-ng [98], or alternatively APPLES [99] (Figure 6B). Users can also optionally provide FAPROTAX2-mapper with their own alignment or phylogenetic placement files, for example if there are alternative third-party tools that are preferred over what is incorporated in FAPROTAX2.

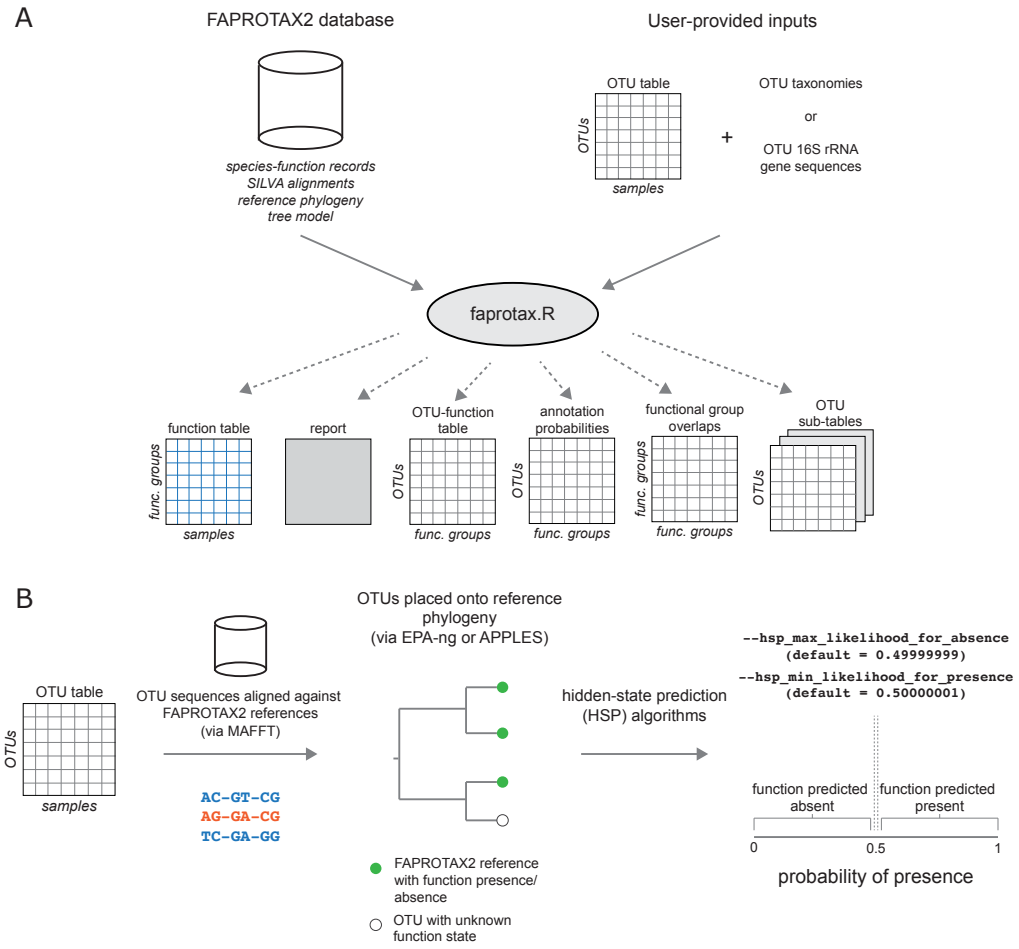


Figure 6. **Functionality of the FAPROTAX2-mapper pipeline.** (A) The faprotax.R script uses the underlying database and user-provided inputs (e.g., OTU tables and either corresponding taxonomies, or corresponding 16S rRNA gene sequences) to generate functional predictions of OTUs. The primary output table lists the abundance of functions predicted by FAPROTAX2-mapper, separately for each microbiome sample. Other optional output tables are illustrated in panel (A). (B) Schematic depicting the general workflow of the FAPROTAX2-mapper pipeline. From left to right: user-provided OTUs are aligned against FAPROTAX2 reference sequences via MAFFT [97], and then placed onto the FAPROTAX2 reference phylogeny via EPA-ng [98] or APPLES [99]. Hidden-state prediction (HSP) algorithms [100] are employed for each function to generate, for each OTU, the probability of presence for each function. Depending on annotation probability, presence or absence is assigned for a function. Users can adjust these thresholds using the aforementioned command-line parameters depending on their specific preferences.

Upon placement of query OTUs onto the FAPROTAX2 reference phylogeny, a new expanded tree is obtained that includes the OTUs and the reference records as tips, with the latter being accompanied by information on the presence or absence of each function. The states of functional traits (presence or absence), or the probabilities thereof at each OTU are predicted using hidden state prediction (HSP) algorithms [100]. Generally speaking, HSP algorithms estimate the state of a trait (in this case, presence or absence of a function) at uncharacterized tips by leveraging the phylogenetic distribution of experimentally validated functions across the reference tree. Under this framework, function states for query OTUs are inferred based on evolutionary relationships to organisms with known phenotypes. For each function, the user may choose either to completely omit any reference records with unknown function state from the HSP, or to set their function state to "absence" or "presence"; in the absence of any user choice, FAPROTAX2-mapper applies default choices according to our best judgment.

FAPROTAX2-mapper can use five alternative HSP algorithms, which can be grouped into two broad categories: model-based probabilistic algorithms, and heuristic, non-probabilistic algorithms. The probabilistic algorithms mainly differ in the assumptions that they make regarding trait evolution and/or the resulting distribution of states across tips. The probabilistic methods include Markov chain HSP, binomial HSP, and empirical probabilities HSP. These approaches estimate the likelihood of function presence based on statistical models of trait evolution or observed trait frequencies among related taxa. The non-probabilistic algorithms include nearest neighbor and maximum parsimony, which infer function states deterministically based on phylogenetic proximity or minimal number of evolutionary changes required, respectively. Internally, FAPROTAX2-

mapper implements these HSP algorithms using the *castor* R package [101], functions `hsp_mk_model`, `hsp_binomial`, `hsp_empirical_probabilities`, `hsp_nearest_neighbor` and `hsp_max_parsimony`. Mathematical details of each algorithm are provided in the official documentation of *castor*.

While users can choose which algorithms they want to use for each function via an optional configuration file, by default FAPROTAX2-mapper will use an optimal algorithm for each function, which we systematically determined via cross-validation as follows. For each function and each cross-validation round, we randomly selected 20% of reference records with known presence/absence to serve as a test set, while the remaining records served as a training set. For the majority of functions, the test set was restricted to species with empirically validated presence and absence annotations. However, for functions where absence cannot be reliably confirmed experimentally (e.g., human pathogenicity), the test set was allowed to include species with unknown (i.e., NA) function states, which were conservatively treated as absences to enable estimation of true negative rates. We evaluated all HSP algorithms by predicting function states for species in the test set across 20 independent cross-validation rounds. For each function and round, we calculated performance metrics including balanced accuracy, true positive rate, and true negative rate. The optimal HSP algorithm for a given function was defined as the method yielding the highest mean balanced accuracy across rounds. This selected HSP algorithm was subsequently set as the default prediction method for that function in FAPROTAX2-mapper. A comprehensive summary of cross-validation performance across all functions is provided in Figure A.1, and across all HSP algorithms in Figure A.2. HSP performance is inherently influenced by both

the prevalence and distribution of all functions across the FAPROTAX2 reference phylogeny (Figure 7).

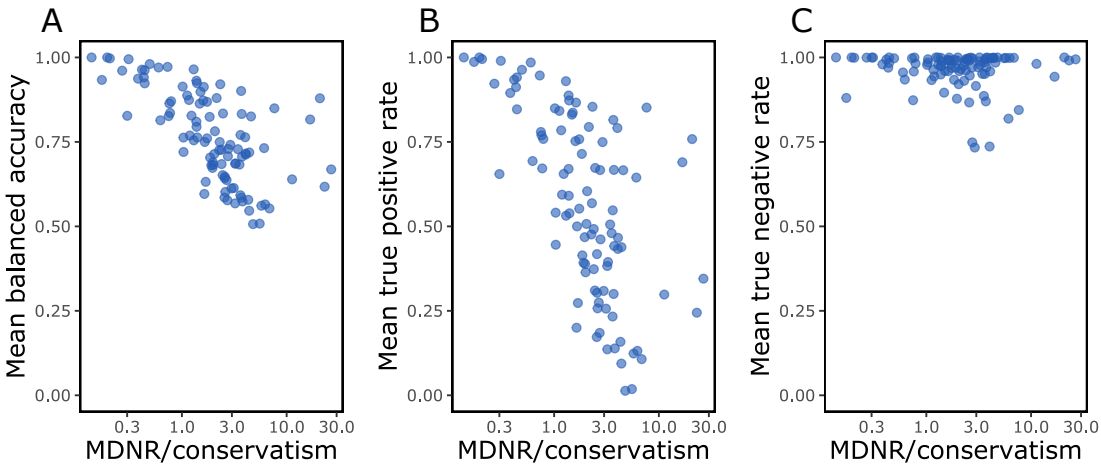


Figure 7. **Summary of internal cross-validation results for FAPROTAX2-mapper.** Scatterplots depicting the predictive performance for each function present in FAPROTAX2, using the optimal HSP algorithm (i.e., the HSP algorithm that maximizes balanced accuracy for each function) identified via internal cross-validation. Horizontal refer to the ratio between the mean distance to nearest FAPROTAX2 reference (MDNR: that is, the phylogenetic distance - in nucleotide substitutions per site - from a query/test species to its nearest reference species, averaged across all species in the test set for each function) and phylogenetic conservatism, given by the consenTRAIT metric [66], which computes the mean phylogenetic depth of all clades for which at least 90% of descending tips exhibit the function. Each point corresponds to a single FAPROTAX2 function, with accuracy metrics averaged across 20 rounds of cross-validation. The balanced accuracy of a function (A) is given by averaging the true positive rates (B) and true negative rates (C). Refer to Figure A.2 for a comparative overview of predictive performance of all HSP algorithms in FAPROTAX2.

The resulting default configuration file generated by this procedure specifies, for each function, not only the optimal HSP algorithm, but also additional parameters governing functional inference. These include the treatment of unknown reference states (`"unknown_ref_states"`), which determines how unknown function states in reference records should be handled (e.g., omitted, or

treated as absence), as well as likelihood thresholds for assigning discrete function states to query OTUs. Specifically, "`min_likelihood_for_presence`" defines the minimum predicted likelihood between 0 and 1 required to infer presence of a function, whereas "`max_likelihood_for_absence`" defines the maximum likelihood threshold (between 0 and 1) below which absence is inferred (Figure 6B). Annotations with values falling between these thresholds are omitted from prediction, depending on user settings. By default, FAPROTAX2-mapper sets the minimum likelihood for presence at 0.5, and a maximum likelihood for absence at 0.5.

The above cross-validation and configuration system are fully implemented within the FAPROTAX2-mapper pipeline, using the "`-mode`" command-line argument, and can thus be fully reproduced and explored by end users. In addition to balanced accuracy, users may specify alternative optimization criteria, including true positive rate, true negative rate, or Matthews correlation coefficient (MCC) [102], depending on specific preferences. The software then generates a corresponding configuration file that lists, for each function, the HSP algorithm that maximizes the selected performance metric. All configuration parameters are user-editable, enabling flexible tuning depending on scientific objectives, for example to prioritize minimizing false positives, minimize false negatives, or achieve a desired balance between the two.

Collection of microbiome samples. For samples specifically collected for this study (Lak01, Sea01, Sed02, Sed03, Sed04), collection took place on dates provided in Table A.1. Run accessions in Table A.1 provide specific latitudes and longitudes of sample collection locations, as well as additional information on how samples were sequenced. The freshwater and marine water microbiome

samples (Lak01, Sea01, Table A.1) collected for this study involved the filtering of 100 mL of freshwater and marine water, respectively, in the field, and subsequent extractions using materials and instructions from the Qiagen® DNeasy PowerWater Kit. The freshwater and marine sediment samples (Sed02-Sed04, Table A.1) were extracted using the Qiagen® DNeasy PowerBiofilm Kit. The unpublished hot spring microbiome samples (Hot01-Hot05, Table A.1) were collected as described in [103]. The unpublished hypersaline lagoon microbiome samples (Sed05-Sed07) were collected as described in [104]. The unpublished surface soil microbiome sample (Soi04) was collected as described in [105].

Analysis of 16S rRNA gene amplicons. For samples specifically collected for this study, amplification and sequencing of the 16S rRNA gene was performed by the IMR. Specifically, the V4-V5 region of the gene was PCR-amplified using the Phusion Plus polymerase and "universal" bacterial + archaeal primers 515FB = GTGYCAGCMGCCGCGGTAA, 926R = CCGYCAATTYMTTTRAGTTT [106, 107], and sequenced on an Illumina MiSeq (2×300 bp paired ends). For all other samples raw 16S rRNA amplicon sequences covering the same V4-V5 region were obtained from the original studies. In total, 3,625,246 raw amplicon sequence pairs were obtained across all 32 samples.

Reads were quality-filtered and ASVs were inferred and chimera-filtered using the R package `dada2` v1.34.0 [12]. Reads were quality-filtered using the `dada2` function `filterAndTrim`, with options "`truncLen = (260,200), maxEE=(1,2), truncQ=(0,0), trimLeft=(6,6), minLen = (100,100), maxLen = (100000,100000)`", retaining in total 2,797,806 read pairs. Error model calibration and ASV inference was performed separately for forward and reverse reads, forward and reverse ASVs were subsequently merged, and the merged

ASVs were chimera-filtered using `dada2` as described previously [105]. ASVs were taxonomically classified based on a comparison to the SILVA database v138.2 [8], using a consensus approach [108]. Non-prokaryotic ASVs, including mitochondria and chloroplasts, were omitted from all subsequent analyses. ASVs uncovered from this 16S analysis were then used as inputs for downstream functional prediction in FAPROTAX2 (Figure 8).

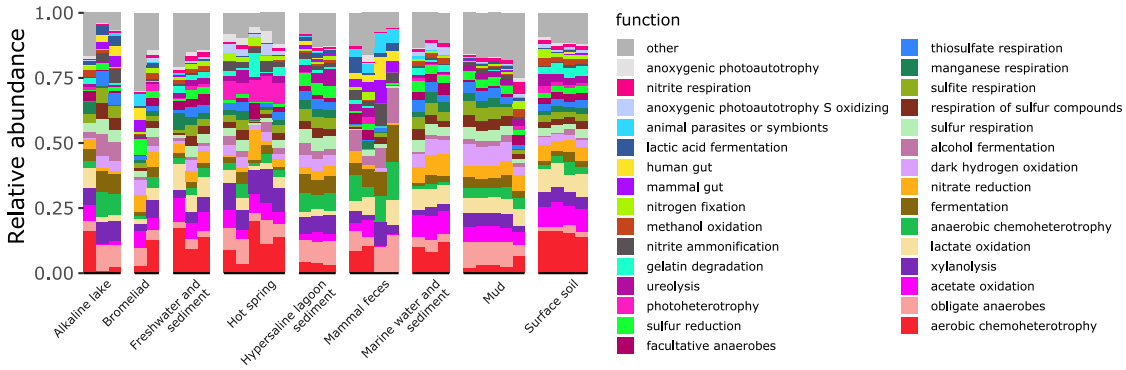


Figure 8. **Functional composition of microbiome samples predicted by FAPROTAX2-mapper.** Example of functional data generated by FAPROTAX2-mapper. For 32 microbiome samples across 9 environment types (Table A.1), functional predictions were made using FAPROTAX2-mapper (under sequence-based placement), and predicted functions are plotted such that the height of each colored bar corresponds to its relative abundance in each microbiome sample. The 30 most abundant functions overall are listed for every microbiome sample. All other functions predicted by FAPROTAX2-mapper for these samples are grouped within "other."

Analysis of metagenomes. For samples specifically collected for this study, shotgun metagenomic sequencing was performed by the Integrated Microbiome Resource in Dalhousie (IMR). Metagenomic libraries were prepared using the Illumina Nextera Flex kit and sequenced using an Illumina NextSeq2000 (2×150 bp paired ends). For all other samples, raw Illumina paired shotgun

metagenomic sequences were obtained from the original studies. In total, 615,846,190 raw metagenomic sequence pairs were obtained across all 32 samples.

Metagenomic sequences were processed as follows. Unless mentioned otherwise, software tools were used with default parameters. Adapters were trimmed from reads using the tool `skewer` v0.2.2 [109] with option "`-mode pe`". Reads were then quality-filtered using `vsearch` v2.28.1 [110] with options "`-fastq_ascii 33 -fastq_maxee 1 -fastq_truncee 1 -fastq_qmax 64 -fastq_maxee_rate 0.005 -fastq_stripleft 0 -fastq_trunclen_keep 10000`", keeping 468,368,059 high-quality read pairs across all samples. Sequences from all samples were pooled and coassembled into longer contiguous sequences (contigs) using `megahit` v1.2.9 [111] with option "`-min-contig-len 200`". Gene-centric functional profiles were computed as described previously [108, 112], and hence here we only provide a brief summary. Protein-coding genes were detected in contigs using `prodigal` v2.6.3, their relative abundances estimated in each sample based on reads mapped with `Bowtie2` v2.5.1 [113], and functionally annotated in terms of KEGG gene orthologs (KOs) using the KOfam HMM database r111 [114] and hmmsearch v3.3 [115]. KOs were aggregated into higher-level functional groups, roughly corresponding to FAPROTAX2 functions, as described previously [108, 112].

3.2.3 Comparing accuracy between FAPROTAX2-mapper and FAPROTAX v1. To assess the changes in predictive accuracy in FAPROTAX2 compared to FAPROTAX v1 [54], we performed cross-validation among functions shared between the two versions as follows. For each function, test set species were randomly chosen among FAPROTAX2 records, while ensuring that these species were not covered within the FAPROTAX v1 database either, neither

directly as species-level records, nor indirectly by higher taxon records according to taxonomies known at the time FAPROTAX v1 was constructed. We further restricted our analysis to functions with at least 100 verified presence records and 100 verified absence records, to ensure sufficient sample sizes for both positive and negative records within each cross-validation round. For each function, we implemented a stratified random sampling approach to construct test sets. Specifically, 5% of species with confirmed presence and 5% of taxa with confirmed absence for a given function were randomly selected and withheld as the test set. The remaining taxa were used as the reference database for prediction.

Functional predictions for species in the test sets were then generated using three alternative approaches: (1) FAPROTAX v1 [54]; (2) FAPROTAX2 using taxonomy-based placement; and (3) FAPROTAX2 using sequence-based placement. For each function and method, predictive performance was quantified by calculating the true positive rates and true negative rates. From these values, balanced accuracy was computed for each cross-validation replicate. This procedure was repeated across 10 independent rounds of cross-validation, and mean balanced accuracy was reported for each function and method.

3.3 Results

3.3.1 Comparison of functional predictions between FAPROTAX v1 and FAPROTAX2.

In order to highlight differences between FAPROTAX v1 and FAPROTAX2 with empirical microbiome data, we compared functional predictions generated by FAPROTAX v1 and FAPROTAX2 across 32 microbiome samples belonging to 9 distinct environment types (Table A.1) by quantifying the number of ASVs assigned to each function shared between both FAPROTAX versions. Across microbiome sample types, the vast majority

of functions exhibited a greater number of ASVs predicted as present under FAPROTAX2 relative to FAPROTAX v1, as evidenced by the predominance of points above the 1:1 reference line in all scatterplots (Figure 9). In many cases, some functions were predicted to be present in one or more orders of magnitude more ASVs under FAPROTAX2 compared to FAPROTAX v1. This pattern was consistent across environments, indicating that FAPROTAX2 systematically increases the number of taxa assigned to functional groups. Importantly, this expansion in predicted functional potential is not limited to specific environments or functions, but rather represents an increase in annotation coverage across the FAPROTAX2 reference database and implementation of less conservative prediction algorithms in FAPROTAX2-mapper.

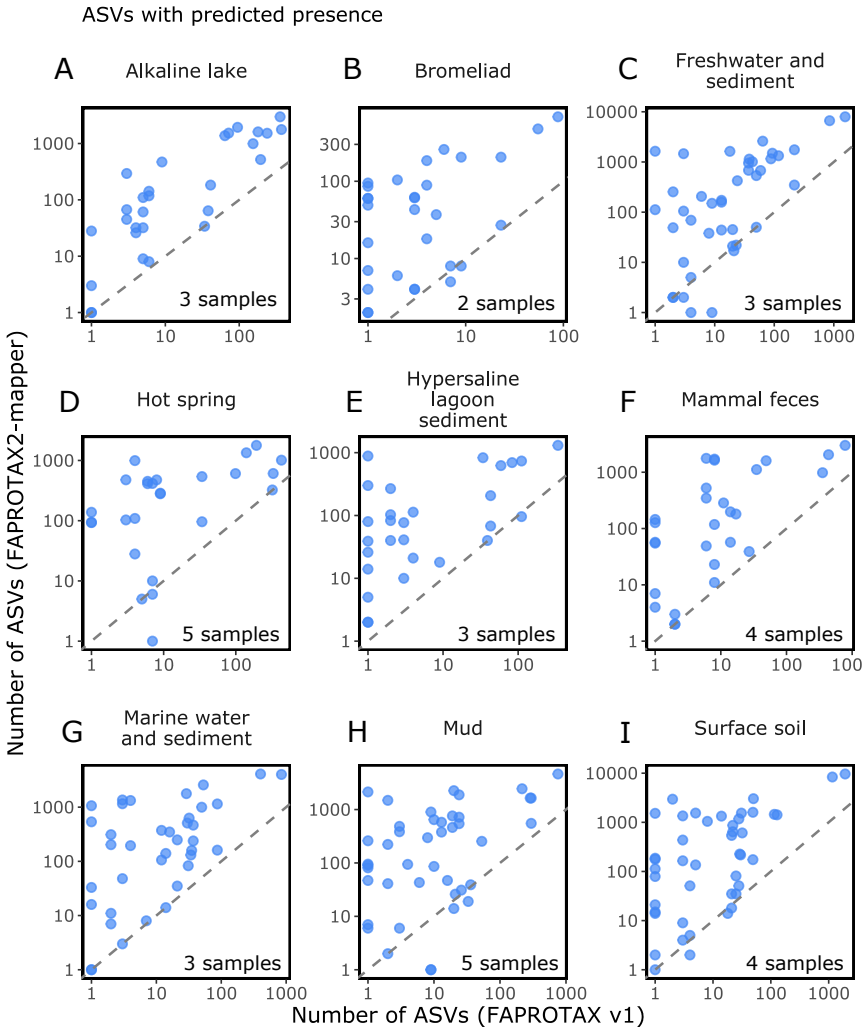


Figure 9. **Number of ASVs with predicted presence for each function by FAPROTAX2-mapper vs. FAPROTAX v1.** Scatterplots depicting the number of ASVs (on a log10 scale) that were predicted present for each function by FAPROTAX v1 (horizontal axes) vs. FAPROTAX2-mapper (vertical axes), using sequence-based placement. Each point in the plots corresponds to the sum of all distinct ASVs predicted to exhibit a function by FAPROTAX v1 and FAPROTAX2 within each microbiome sample type (see Table A.1 for details on microbiome samples). The dashed diagonal in each plot represents hypothetical scenarios of equal numbers of predicted ASVs, and is shown for comparison purposes.

Next, we evaluated the predictive accuracy of FAPROTAX2 compared to FAPROTAX v1 using cross-validation. Across all evaluated functions, the mean balanced accuracy was higher among FAPROTAX2-mapper methods than FAPROTAX v1 (Figure 10). Mean balanced accuracy was the lowest for FAPROTAX v1 at 0.61, whereas FAPROTAX2 under taxonomy-based placement and sequence-based placement yielded mean balanced accuracies of 0.77 and 0.81, respectively (Figure 10A, Table 3). This same result was observed for mean true positive rates, with both FAPROTAX2-mapper approaches substantially outperforming FAPROTAX v1 (Figure 10B, Table 3). In contrast, all three methods exhibited relatively high true negative rates, although FAPROTAX v1 achieved the highest values (mean true negative rate ≈ 0.99), followed by FAPROTAX2 with taxonomy-based placement (≈ 0.98) and FAPROTAX2 with sequence-based placement (≈ 0.97) (Figure 10C, Table 3).

Despite a slightly lower observed mean true negative rate, FAPROTAX2 exhibits a marked increase in true positive rate, resulting in a considerably higher overall balanced accuracy. This pattern reflects a well-established tradeoff in binary classification, whereby approaches that strongly prioritize minimizing false positives (i.e., Type I errors) typically do so at the expense of reduced true positives (i.e., increased Type II errors) [116, 117]. In this context, FAPROTAX v1 can be interpreted as a highly conservative method that favors true negatives, but at the cost of failing to recover a large proportion of true positives. FAPROTAX2 shifts this balance toward a more optimal tradeoff between sensitivity and specificity, yielding improved overall predictive performance.

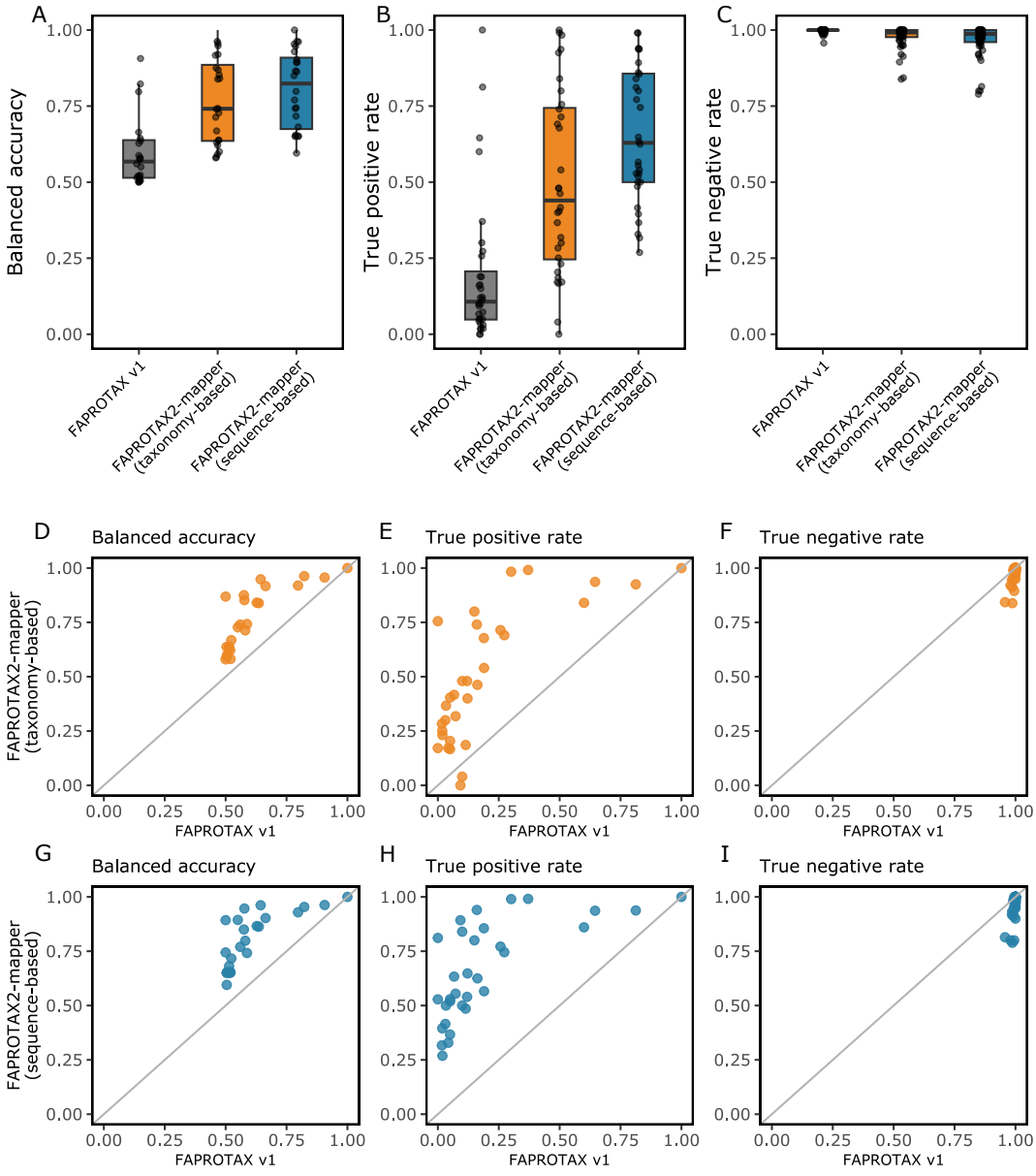


Figure 10. **Comparison of predictive accuracy between FAPROTAX v1 and FAPROTAX2-mapper.** (A,B,C) Cross-validation results between FAPROTAX v1, FAPROTAX2 using taxonomy-based placement, and FAPROTAX2 using sequence-based placement. Each point represents the average score (i.e., balanced accuracy, true positive rate, and true negative rate) across 10 rounds of cross-validation for one function. (D,E,F) Average scores from cross-validation between FAPROTAX v1 and FAPROTAX2 using taxonomy-based placement. (G,H,I) Average scores from cross-validation between FAPROTAX v1 and FAPROTAX2 using sequence-based placement.

Table 3. **Summary of predictive accuracy of FAPROTAX2-mapper vs. FAPROTAX v1.** Scores are averaged across 10 rounds of cross-validation, as described in the Methods. Values presented in this table reflect the data visualized in Figure 10.

FAPROTAX version	**Accuracy metric**	**Mean score**
FAPROTAX v1	Balanced accuracy	0.61
FAPROTAX2-mapper (taxonomy-based)	Balanced accuracy	0.77
FAPROTAX2-mapper (sequence-based)	Balanced accuracy	0.81
FAPROTAX v1	True positive rate	0.19
FAPROTAX2-mapper (taxonomy-based)	True positive rate	0.50
FAPROTAX2-mapper (sequence-based)	True positive rate	0.66
FAPROTAX v1	True negative rate	0.99
FAPROTAX2-mapper (taxonomy-based)	True negative rate	0.98
FAPROTAX2-mapper (sequence-based)	True negative rate	0.97

The sequence-based implementation of FAPROTAX2-mapper achieved the highest average balanced accuracy and average true positive rate across functions. However, the taxonomy-based approach also represents a substantial improvement over FAPROTAX v1, indicating that the new database and HSP algorithms alone provide meaningful gains in predictive performance. This distinction may be particularly relevant for users seeking to minimize computational demands or installation of third-party dependencies, as taxonomy-based placement is considerably easier to set up in terms of dependencies, and executes much faster than sequence-based placement. Finally, we emphasize that FAPROTAX2 provides users with the ability to define specific probability thresholds for assigning functional presence, using the configuration file available in the software package. This flexibility allows users to tune predictions according to their preferred tolerance for false positives and false negatives; for example, increasing

the threshold can further improve true negative rates when more conservative predictions are desired for a subset of functions.

3.3.2 FAPROTAX2 predictions versus metagenomes. To evaluate the extent to which phylogenetically inferred functions via FAPROTAX2 correspond to gene-based functional profiles, we compared predicted abundances of FAPROTAX2 functions to corresponding KEGG [79] functional groups across the 32 aforementioned microbiome samples (Table A.1). We limited this analysis to FAPROTAX2 functions for which an unambiguous synonymous mapping to KEGG functions existed (Table A.2). Among these FAPROTAX2-KEGG function pairs, we observed substantial variability in terms of strength of associations. Certain functions, such as methanogenesis and dark thiosulfate oxidation, exhibited strong and statistically significant positive correlations across samples (Figure A.3), indicating broad agreement between phenotypic inference and metagenomic gene content for these functions. In contrast, other functions, including xylanolysis and cellulolysis, exhibited very weak or non-significant correlations.

Overall, the strength of correlation varied among functions, with some showing consistent agreement between phenotype-based predictions and gene-based functional estimates, and others showing discordance. These observations provide a descriptive overview of the relationship between phylogenetically inferred functional potential and metagenomic gene content across microbiome samples from diverse environments.

3.4 Discussion

3.4.1 Advances in FAPROTAX2. FAPROTAX2 represents a substantial methodological improvement upon FAPROTAX v1 [54] in predicting prokaryotic functional traits in microbiome surveys. Whereas FAPROTAX v1

employed a highly conservative framework, in which a query OTU is annotated with a function only if all members of its parent taxa in the underlying database were verified to exhibit that function, FAPROTAX2-mapper leverages rigorous phylogenetic inference via hidden-state prediction (HSP) algorithms [100] to estimate the probability of each OTU exhibiting each function. This approach reduces the overly conservative bias of FAPROTAX v1, leading to higher detection rates while also greatly improving predictive accuracy. Consequently, FAPROTAX2 has an improved balance between sensitivity and specificity, as evidenced by higher true positive rates and overall balanced accuracy across all evaluated functions (Figure 10). Additionally, beyond the core phylogenetic methodology, FAPROTAX2 provides users with an expanded suite of tunable parameters for functional predictions. Users can select among multiple HSP algorithms, define likelihood thresholds for assigning presence or absence, and specify how unknown reference states in the underlying database should be treated, separately for each function. This flexibility allows users to tailor predictions to their unique research objectives, whether prioritizing conservative assignments or maximizing functional coverage, and represents a significant advancement over the rigid design of FAPROTAX v1.

The methodological improvements of FAPROTAX2-mapper have clear consequences for microbiome research. Across diverse environmental samples, FAPROTAX2 systematically predicts a greater number of ASVs for most functional traits compared to FAPROTAX v1 (Figure 9). Importantly, this increase is not accompanied by a proportional rise in false positives (Figure 10). Rather, it reflects the software's enhanced ability to detect true functional potential. These results

indicate that FAPROTAX2 provides a more reliable representation of the metabolic capacities of microbial communities.

3.4.2 FAPROTAX2 functions versus KEGG orthologs. The comparison between FAPROTAX2-mapper predictions and KEGG-derived functional abundances highlights both the potential and limitations of phenotype-based inference. While certain functions, such as methanogenesis and dark thiosulfate oxidation, showed strong correspondence between phylogenetically inferred functions and metagenomic gene content, many other functions, including xylanolysis and cellulolysis, exhibited weak or inconsistent correlations (Figure A.3). This heterogeneity likely reflects both methodological and biological factors. From a methodological perspective, KEGG functional abundances in this study are derived from genes identified in metagenome-assembled genomes (MAGs), and therefore represent the aggregate gene content present within a microbial community. In contrast, FAPROTAX2-mapper predictions are based on the phylogenetic placement of OTUs onto a curated reference tree, with functions inferred from closely related, functionally annotated cultured taxa. As such, the two approaches interrogate distinct yet important representations of functional potential: one grounded in observed gene content, and the other in empirically validated functional annotations based on cultured prokaryotic taxa. Discrepancies between these approaches are therefore unsurprising, and even expected, particularly in cases where taxonomic resolution is limited, or where functional traits are unevenly distributed within clades.

Beyond methodological differences, our results highlight fundamental limitations in mapping genes to higher-level functional traits. The presence of genes associated with a given KEGG function does not necessarily imply that the

corresponding phenotype is expressed in situ, nor that it contributes meaningfully to ecological processes. Gene expression is regulated by environmental context, and many metabolic capabilities are facultative rather than constitutive [118]. Conversely, the presence of a phenotype, especially one defined at a coarse ecological or metabolic level, may not be captured by currently annotated KEGG pathways. Many microbial functions likely arise from combinations of genes or yet-uncharacterized genetic mechanisms that are not fully represented in existing databases [83, 119]. These factors likely contribute to the weak or inconsistent correspondence observed for several functions.

The observed discordance between phenotype-based and gene-based functional estimates should not be interpreted solely as a limitation of either approach, but rather as an opportunity to refine our understanding of gene-phenotype relationships in microbial systems. The curated phenotypic associations within the FAPROTAX2 database provide a valuable framework for linking taxa to ecological functions, independent of current gene annotations. As such, this resource could serve as a starting point for identifying novel genetic determinants or proxies of microbial functions. For instance, genome-resolved datasets corresponding to taxa represented in FAPROTAX2 could be leveraged in conjunction with machine learning approaches to identify gene families or genomic signatures that are associated with specific functions. Such analyses could help uncover previously uncharacterized genes or pathways underlying key microbial functions, ultimately improving both gene-centric and phenotype-centric methods for functional predictions. These results emphasize that while strong concordance between gene- and phenotype-based approaches can be observed for certain well-characterized functions, substantial gaps remain in our ability to directly map genes

to ecological functions. Addressing these gaps will require integrative approaches that combine phylogenetic, genomic, and ecological data to more fully resolve the genetic basis of microbial functional diversity.

3.4.3 Practical considerations and conclusions. FAPROTAX2 represents a substantial advancement over FAPROTAX v1 in both methodology and database scope. Rather than assigning functions strictly based on taxonomic affiliation, as was done in FAPROTAX v1, FAPROTAX2 considers the phylogenetic relationship between user-provided OTUs and species in the underlying database with rigorous HSP algorithms. Cross-validation analyses demonstrate marked improvements in predictive performance, with consistently higher true positive rates and balanced accuracy across functions. These gains reflect both the expanded reference database and the incorporation of improved HSP algorithms, enabling more sensitive and reliable inference of microbial metabolic functions.

The performance of FAPROTAX2 should be interpreted in the context of several important considerations. First, the reference database and phylogeny are intentionally restricted to cultured prokaryotic taxa. This design prioritizes empirically validated functions, ensuring that predictions are based on experimental evidence. Predictions for OTUs from poorly characterized or deeply branching lineages may therefore be associated with increased uncertainty. However, this uncertainty can be explicitly captured by probabilistic HSP algorithms in FAPROTAX2-mapper. Users can also adjust function-specific probability thresholds for predictions; for example, the minimum likelihood required to assign presence can be increased in order to reduce false positives. In addition, FAPROTAX2 provides an optional constraint on phylogenetic extrapolation through the `-hsp_max_phylodistance` parameter, which sets a maximum allowable

phylogenetic distance between a query OTU and its nearest FAPROTAX2 reference with a known state for a given function. When this threshold is exceeded, no prediction is made, allowing users to restrict inferences to closely-related organisms in the reference phylogeny. This filter can be specified as an absolute distance or relative to each function's phylogenetic conservatism, and can be set per-function via the configuration file. As efforts to culture and characterize novel microbial lineages continue to expand [76, 77, 78], the breadth of the reference database is expected to improve.

Second, FAPROTAX2 operates at the species level, reflecting the resolution at which functions can be most practically documented from the experimental literature. Although this enables standardized annotation across thousands of taxa, it does not capture functional heterogeneity among strains within a species. Such variation is well documented in microbial systems, where closely related strains can differ substantially in gene content [120] and metabolic capabilities [66]. FAPROTAX2 predictions are thus estimates of species-level functional potential. Importantly, the phylogenetic placement framework partially mitigates this limitation by allowing predictions to be informed by explicit tree-based relationships rather than strict taxonomic affiliation, but functions whose presence varies intraspecifically remain inherently more difficult to predict.

Lastly, all predictions generated by FAPROTAX2 reflect the potential to exhibit a function rather than realized functional activity. The predicted presence of a function indicates that the underlying metabolic capability likely exists in an OTU, and therefore in its corresponding microbial community, but does not guarantee expression *in situ*. Indeed, this distinction is fundamental to all functional inference methods from amplicon sequencing. FAPROTAX2

therefore serves as a powerful tool for functional prediction that can be applied in downstream analyses and hypothesis-driven studies. The probabilistic and phylogenetically informed predictions made by FAPROTAX2 provide a basis for comparing functional profiles across microbiome samples and conditions, and can be integrated with complementary approaches including metagenomics, metatranscriptomics, metabolic models, or cultivation-based assays.

Overall, FAPROTAX2 is a flexible tool for functional inference from 16S-based microbiome surveys. By integrating an empirically curated, species-level database with rigorous phylogenetic placement and alternative HSP algorithms, FAPROTAX2 moves beyond strict taxonomic annotation toward quantitative estimates of functional potential in microbial communities. The probabilistic outputs of FAPROTAX2-mapper, combined with user-configurable thresholds and HSP selection, allow researchers to tailor predictions to specific analytical goals, balancing sensitivity and specificity as preferred. FAPROTAX2 provides a foundation for interpretable links between microbial community composition and function, enabling evaluation of biological hypotheses across diverse environmental and host-associated systems.

All code for analysis and visualization of data in this chapter is available at `https://github.com/robertporch/faprotax2_software_manuscript`.

CHAPTER IV

CONCLUSION

Understanding both the taxonomic and functional composition of microbial communities is crucial for answering many important ecological questions. However, the popular approach of 16S rRNA amplicon sequencing is limited to only taxonomic information a priori, posing a significant limitation. Therefore, there is a continued demand for robust, scalable approaches that can bridge microbial taxonomic identity and function within marker-gene-based methods.

In my dissertation, I addressed this gap by developing integrated resources and methods to improve functional inference from 16S rRNA microbiome surveys. Central to this effort was the construction of a large, curated phenotypic database of cultured prokaryotic species, capturing experimentally validated metabolic functions. This database serves as a foundation for linking taxonomy to function, and enables the development of predictive approaches based on empirical observations.

In Chapter II, I developed FAPROTAX2-db, a comprehensive database encompassing 13,574 bacterial and archaeal species across 101 functions. Analyses of the collated data in FAPROTAX2-db revealed important patterns in the distribution of functional traits, including biases in taxonomic representation, structured relationships among functions, and variability in phylogenetic conservatism. These results highlight both the strengths and current limitations of experimentally derived phenotypic data, and demonstrate the broader utility of FAPROTAX2-db for ecological, evolutionary, and computational applications. For example, the database can be used as training data to incorporate machine learning models to identify novel gene groups as proxies for metabolic functions.

Indeed, similar recent work has been done to try to characterize the many genes of unknown function in microbial communities [121, 122], and we expect future efforts to benefit from the data available in FAPROTAX2-db. Additionally, the data in FAPROTAX2-db can be leveraged to better understand the role of function gain/loss in prokaryotic diversification, using Binary State Speciation and Extinction (BiSSE) Models [85].

In Chapter III, I developed FAPROTAX2-mapper, a functional prediction software that uses FAPROTAX2-db to infer metabolic functions from 16S rRNA microbiome surveys. FAPROTAX2-mapper integrates phylogenetic placement with hidden-state prediction (HSP) algorithms, which represent marked improvements over the original, rigid methods of FAPROTAX v1. This results in more flexible and accurate functional predictions, which I demonstrated through cross-validation analyses and application to real microbiome data, where I observed broader and more sensitive detection of functional potential across environment types. Overall, FAPROTAX2-mapper methods have considerably higher predictive performance, both in terms of balanced accuracy and true positive rates, across all functions. Additionally, analyses of real microbiome data revealed that FAPROTAX2-mapper yields substantially higher detection of functions in amplicon sequence variants (ASVs) than in FAPROTAX v1. These results together indicate that FAPROTAX2-mapper is a more rigorous, flexible approach for functional predictions. Importantly, in this chapter I identified significant variation in concordance between FAPROTAX2 functions and KEGG orthologs from microbiome samples. In other words, the associations between FAPROTAX2-KEGG function pairs were highly variable, with some showing strong concordance (e.g., methanogenesis) and others exhibiting very weak correlations

(e.g., xylanolysis). These results highlight an outstanding issue in microbial ecology: an incomplete and uneven mapping between genes and phenotypes across microbial lineages. Expanding our knowledge of the genetic basis underlying microbial phenotypes will therefore be critical for improving these associations, and further motivates future work to identify novel gene profiles associated with the metabolic functions defined in FAPROTAX2-db (Chapter II).

Amplicon sequencing of the 16S rRNA gene is expected to remain one of the most popular methods for surveying microbial communities and answering key questions in microbial ecology. The methods I developed for my dissertation contribute both a structured synthesis of experimentally validated functional traits and a novel approach for improving the accuracy of functional inference in microbiome research. By enabling more reliable linkage between taxonomic composition and metabolic potential, these tools expand the interpretive power of marker-gene microbiome surveys. Ultimately, the advances described in my dissertation support a more comprehensive understanding of how microbial communities contribute to ecological processes across environments.

APPENDIX

APPENDIX

A.1 Additional Figures for Chapter III

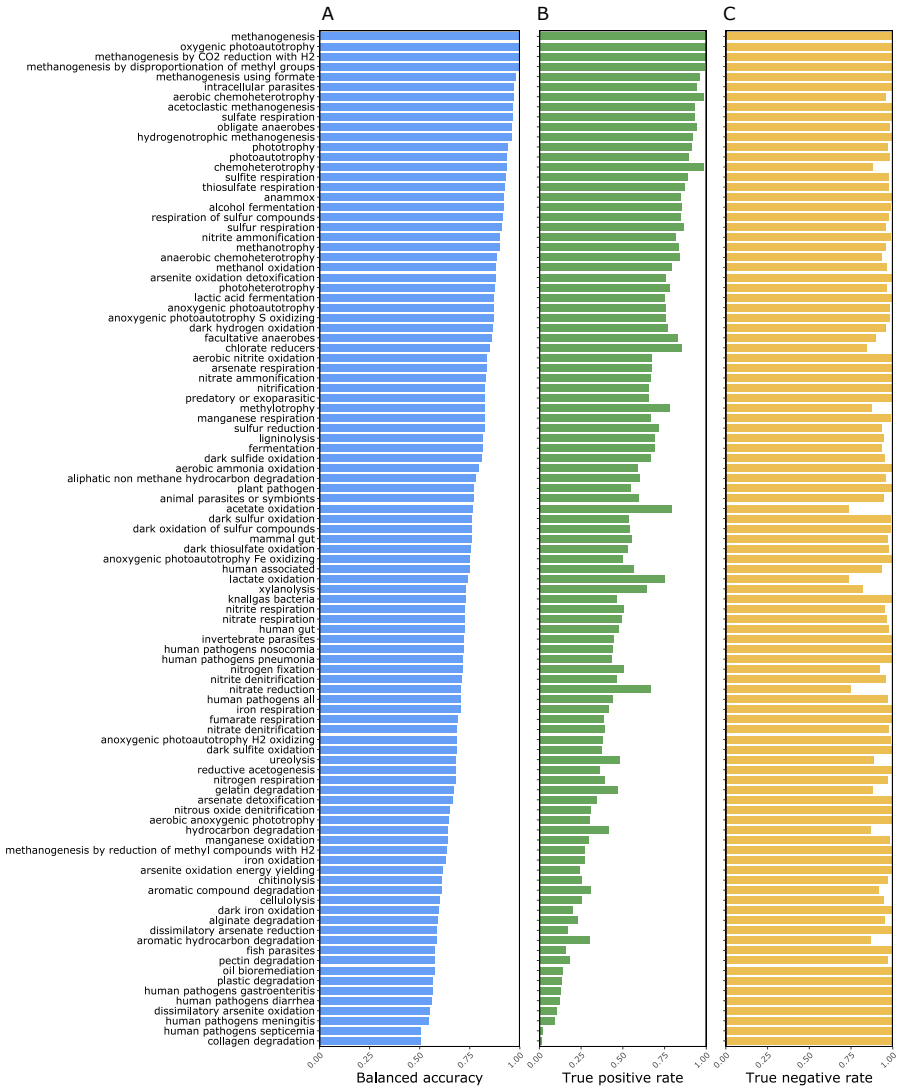


Figure A.1. **Accuracy scores from internal cross-validation of FAPROTAX2.** Bar plots of prediction accuracy metrics for each function based on internal cross-validation. For each function, the HSP algorithm yielding the highest balanced accuracy was selected as the default choice in FAPROTAX2. Corresponding accuracy metrics are shown for balanced accuracy (A), true positive rate (B), and true negative rate (C).

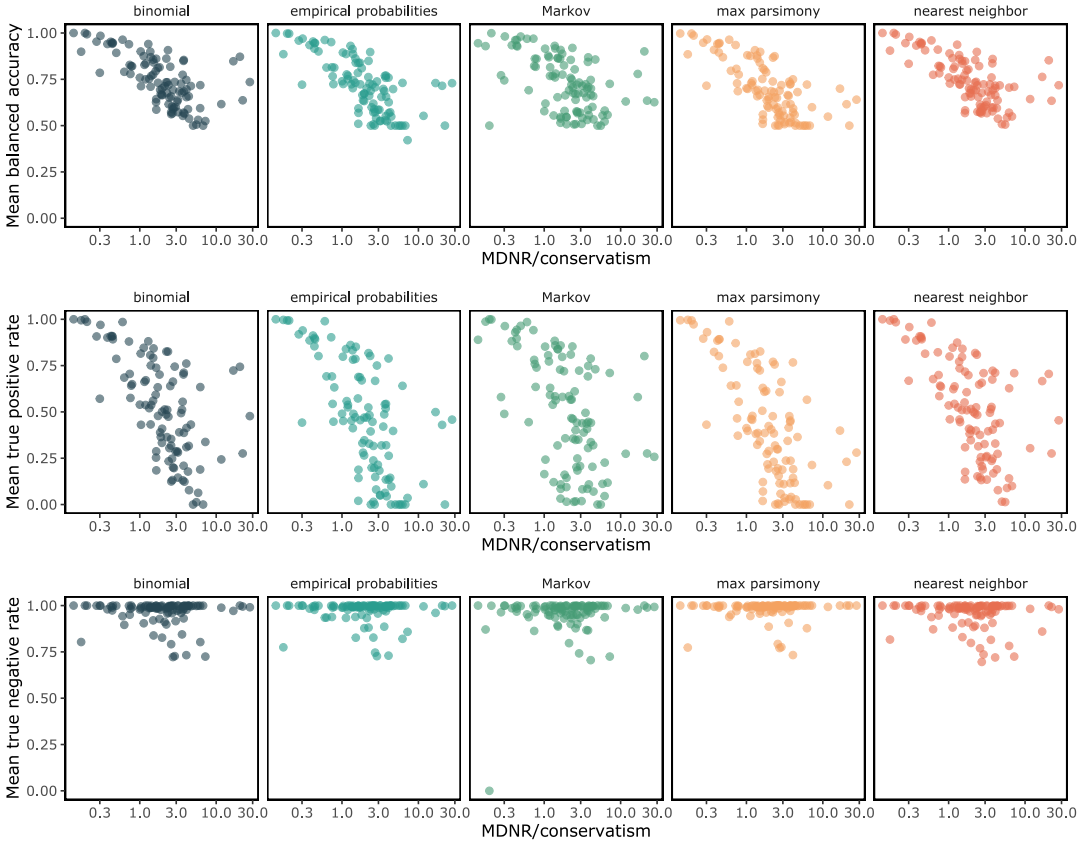


Figure A.2. **Comparative predictive performance of HSP algorithms in FAPROTAX2-mapper.** Scatterplots depicting the predictive performance for each function present in FAPROTAX2 across all candidate HSP algorithms. Horizontal axes refer to the ratio between the mean distance to nearest FAPROTAX2 reference (MDNR: that is, the phylogenetic distance - in nucleotide substitutions per site - from a query/test species to its nearest reference species, averaged across all species in the test set for each function) and phylogenetic conservatism, given by the consenTRAIT metric [66], which computes the mean phylogenetic depth of all clades for which at least 90% of descending tips exhibit the function. Each point corresponds to a single FAPROTAX2 function, with accuracy metrics averaged across 20 rounds of cross-validation. The balanced accuracy (top row) of a function is given by averaging the true positive rates (middle row) and true negative rates (bottom row).

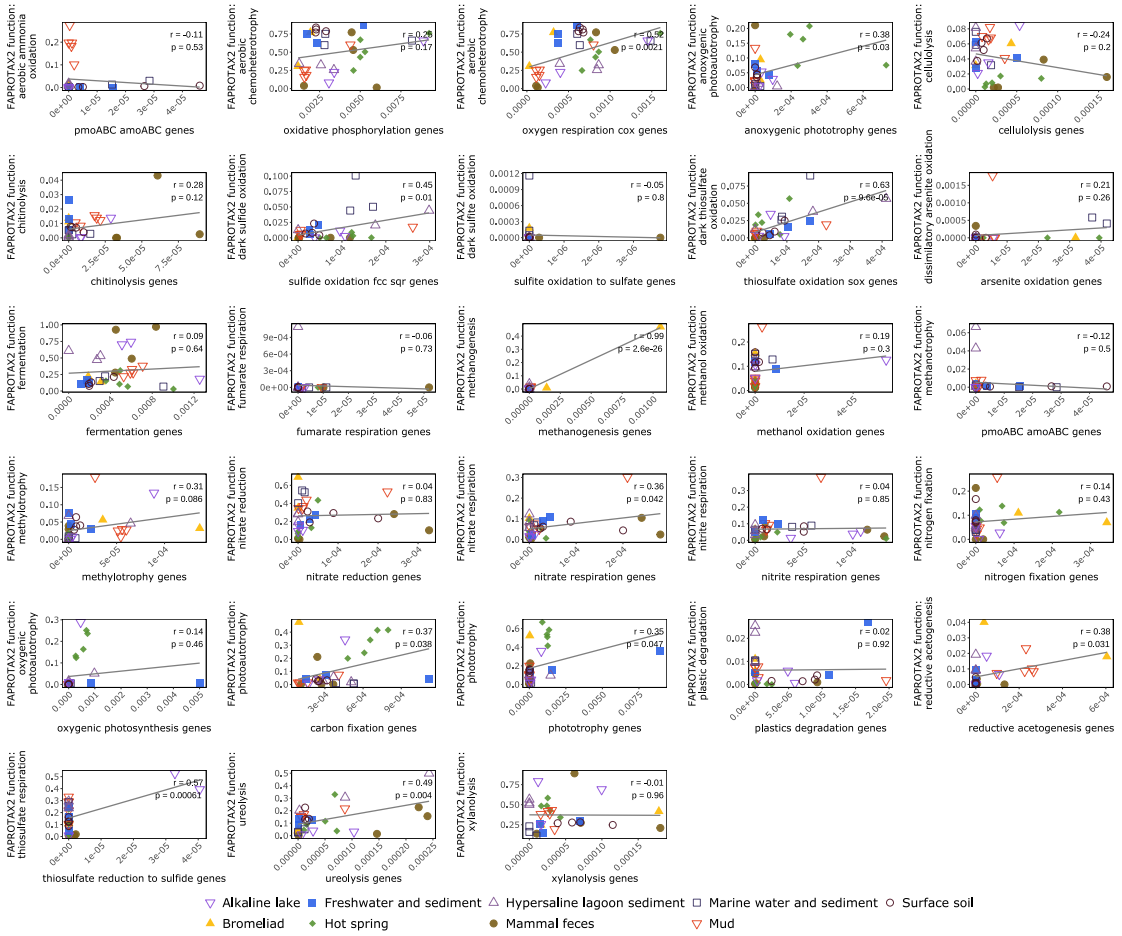


Figure A.3. **Associations between abundances of KEGG functional groups and predicted FAPROTAX2 functions.** Scatterplots depicting abundances of KEGG functional groups (horizontal axes) and abundances of FAPROTAX2-mapper functions (vertical axes) detected in microbiome surveys, for functions in FAPROTAX2-mapper with unambiguous synonymous mappings to KEGG functions (Table A.2). Points in each scatterplot are colored and shaped based on microbiome environment type, as described in Table A.1.

A.2 Additional Tables for Chapter III

Table A.1. Description of microbiome samples used to demonstrate the utility of FAPROTAX2-mapper and compare functional predictions made by FAPROTAX2-mapper compared to FAPROTAX v1.

Sample	**Sample type**	**Run accession**	**Collection date**	**Project reference**
Brm01	Bromeliad	SRR5248631	2015.01.10	DOI:10.1038/s41559-016-0015
Brm02	Bromeliad	SRR3498566	2015.01.10	DOI:10.1038/s41559-016-0015
Fec01	Mammal feces	SRR31749522	2022.11.16	DOI:10.1111/1755-0998.70020
Fec02	Mammal feces	SRR31749556	2022.11.16	DOI:10.1111/1755-0998.70020
Fec03	Mammal feces	SRR31749545	2022.11.16	DOI:10.1111/1755-0998.70020
Fec04	Mammal feces	SRR35428645	2024.06.15	Unpublished
Hot01	Hot springs	SRR32621835	2021.09.11	Unpublished
Hot02	Hot springs	SRR32621823	2021.09.12	Unpublished
Hot03	Hot springs	SRR32621797	2021.09.13	Unpublished
Hot04	Hot springs	SRR32621785	2021.09.14	Unpublished
Hot05	Hot springs	SRR32621745	2021.08.30	Unpublished
Lak01	Freshwater and sediment	SRR32997310	2024.03.20	This study
Mud01	Mud	SRR10091015	2018.01.21	DOI:10.1111/1758-2229.12866
Mud02	Mud	SRR10091014	2018.01.21	DOI:10.1111/1758-2229.12866
Mud03	Mud	SRR10091140	2016.12.24	DOI:10.1111/1758-2229.12866
Mud04	Mud	SRR10091047	2018.01.20	DOI:10.1111/1758-2229.12866
Mud05	Mud	SRR10091045	2018.01.20	DOI:10.1111/1758-2229.12866
Sea01	Marine water and sediment	SRR32997308	2023.11.12	This study

Table A.1. Table continued from previous page.

Sample	**Sample type**	**Run accession**	**Collection date**	**Project reference**
Sed01	Marine water and sediment	SRR29138649	2023.10.08	DOI:10.1038/s42003-024-07384-y
Sed02	Freshwater and sediment	SRR32997309	2024.03.20	This study
Sed03	Marine water and sediment	SRR32997307	2023.11.12	This study
Sed04	Freshwater and sediment	SRR32997306	2024.03.20	This study
Sed05	Hypersaline lagoon sediment	SRR6266499	2015.01.12	Unpublished
Sed06	Hypersaline lagoon sediment	SRR6266503	2015.01.12	Unpublished
Sed07	Hypersaline lagoon sediment	SRR6266505	2015.01.12	Unpublished
Sod01	Alkaline lake	SRR25252483	2022.08.20	DOI:10.1111/1462-2920.16578
Sod02	Alkaline lake	SRR25252480	2022.08.20	DOI:10.1111/1462-2920.16578
Sod03	Alkaline lake	SRR25252479	2022.08.21	DOI:10.1111/1462-2920.16578
Soi01	Surface soil	SRR31749538	2022.11.09	DOI:10.1111/1755-0998.70020
Soi02	Surface soil	SRR31749527	2022.11.09	DOI:10.1111/1755-0998.70020
Soi03	Surface soil	SRR31749512	2022.11.09	DOI:10.1111/1755-0998.70020
Soi04	Surface soil	SRR35428628	2024.06.06	Unpublished

Table A.2. List of FAPROTAX2 functions mapped to KEGG functional groups and KEGG orthologous groups (KOGs).

FAPROTAX2 function	**KEGG functional group**	**KOGs**
aerobic ammonia oxidation	pmoABC/amoABC	K10944, K10945, K10946
aerobic chemoheterotrophy	oxidative phosphorylation	KEGG pathway ko00190
aerobic chemoheterotrophy	oxygen respiration (*cox*)	K02277, K02276, K02274, K15408, K02275, K02262, K02256, K02261, K02263, K02264, K02265, K02266, K02267, K02268, K02269, K02270, K02271, K02272, K02273, K02258, K02259, K02260, K00404, K00405, K15862, K00407, K00406, K02259, K02259, K00406, K00407, K00405, K00404, K00425, K00426
anoxygenic photoautotrophy	anoxygenic phototrophy	K08944, K08940, K08941, K08942, K08943, K08946, K08945, K08947, K08948, K08949, K08950, K08951, K08952, K08953, K08954, K08926, K08927, K08928, K08929, K13991, K13992, K13994, K08939, K08930
cellulolysis	cellulolysis	K19356, K19357, K01179, K20542, K01225, K19668
chitinolysis	chitinolysis	K01183, K13381, K01452, K17523, K20547
dark sulfide oxidation	sulfide oxidation fcc sqr	K17229, K17230, K17218
dark sulfite oxidation	sulfite oxidation to sulfate	K00386, K00387, K05301
dark thiosulfate oxidation	thiosulfate oxidation (*sox*)	K17222, K17223, K17224, K17225, K17226, K17227

Table A.2. Table continued from previous page.

FAPROTAX2 function	**KEGG functional group**	**KOGs**
dissimilatory arsenite oxidation	arsenite oxidation	K08355, K08356
fermentation	fermentation	K01568, K13951, K00114, K00002, K04022, K00128, K00129, K00016, K00102, K00656, K00825, K00004, K00929, K00248, K00239, K00240, K00241, K00242
fumarate respiration	fumarate respiration	K00244, K00245
methanogenesis	methanogenesis	K00399, K00401, K00402
methanol oxidation	methanol oxidation	K14028, K14029, K17066, K16254, K16255, K16256, K16257, K16258, K16259, K16260, K00093, K17066
methanotrophy	pmoABC/amoABC	K10944, K10945, K10946
methylotrophy	methylotrophy	K14028, K14029, K17066, K16254, K16255, K16256, K16257, K16258, K16259, K16260, K00093, K17066, K16157, K16158, K16159, K16160, K16161, K16162, K10713, K14028, K14029, K17066
nitrate reduction	nitrate reduction	K02567, K02568, K00370, K00371, K00372, K00373, K00374, K00367, K10534, K00360
nitrate respiration	nitrate respiration	K02567, K02568, K00370, K00371, K00374, K00373
nitrite respiration	nitrite respiration	K00368, K15864, K00362, K00363, K03385, K15876
nitrogen fixation	nitrogen fixation	K02588, K02586, K02591, K00531

Table A.2. Table continued from previous page.

FAPROTAX2 function	**KEGG functional group**	**KOGs**
oxygenic photoautotrophy	oxygenic photosynthesis	K02703, K02704, K02705, K02706, K02707, K02708, K02709, K02710, K02711, K02712, K02713, K02714, K02716, K02717, K02718, K02719, K02720, K02721, K02722, K02723, K02724, K03541, K03542, K08901, K08902, K08903, K08904
photoautotrophy	carbon fixation	K02703, K02704, K02705, K02706, K02707, K02708, K02709,K 02710,K 02711, K02712, K02713, K02714, K02716, K02717, K02718, K02719, K02720, K02721, K02722, K02723, K02724, K03541, K03542, K08901, K08902, K08903, K08904, K02689, K02690, K02691, K02692, K02693, K02694, K08905, K02695, K02702, K14332, K02701, K02700, K02699, K02698, K02697, K02696, K02638, K02639, K02641, K08906
phototrophy	phototrophy	KOGs from **anoxygenic phototrophy** and **oxygenic photosynthesis**

Table A.2. Table continued from previous page.

FAPROTAX2 function	**KEGG functional group**	**KOGs**
plastic degradation	plastics degradation	K18068, K18069, K18067, K18251, K18254, K18255, K18074, K18075, K18077
reductive acetogenesis	reductive acetogenesis	K00192, K00193, K00194, K00195, K00197
thiosulfate respiration	thiosulfate reduction to sulfide	K08352, K08353, K08354
ureolysis	ureolysis	K01941
xylanolysis	xylanolysis	K15924, K13465, K01198, K15920, K01181, K05972

Bibliography

[1] Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, 68(4):669–685, 2004.

[2] Helen L Steele and Wolfgang R Streit. Metagenomics: advances in ecology and biotechnology. *FEMS microbiology letters*, 247(2):105–111, 2005.

[3] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, et al. A new view of the tree of life. *Nature microbiology*, 1(5):1–6, 2016.

[4] Krista M Ruppert, Richard J Kline, and Md Saydur Rahman. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17:e00547, 2019.

[5] M.S. Rappé and S.J. Giovannoni. The uncultured microbial majority. *Annual Review of Microbiology*, 57(1):369–394, 2003.

[6] Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, 2014.

[7] Carl R Woese, George E Fox, Lawrence Zablen, Tsuneko Uchida, Linda Bonen, Kenneth Pechman, Bobby J Lewis, and David Stahl. Conservation of primary structure in 16S ribosomal RNA. *Nature*, 254(5495):83–86, 1975.

[8] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.

[9] Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, 72(7):5069–5072, 2006.

[10] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639–2643, 2017.

[11] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban De Vargas, and Micah Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420, 2015.

[12] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7):581–583, 2016.

[13] Robert C Edgar. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, page 081257, 2016.

[14] Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, Luke R Thompson, Embriette R Hyde, Antonio Gonzalez, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2):10–1128, 2017.

[15] Sydney I Glassman and Jennifer BH Martiny. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *MSphere*, 3(4):10–1128, 2018.

[16] THERESA B Britschgi and Stephen J Giovannoni. Phylogenetic analysis of a natural marine bacterioplankton population by rrna gene cloning and sequencing. *Applied and Environmental Microbiology*, 57(6):1707–1713, 1991.

[17] Edward F DeLong. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences*, 89(12):5685–5689, 1992.

[18] Anna-Louise Reysenbach, Gene S Wickham, and Norman R Pace. Phylogenetic analysis of the hyperthermophilic pink filament community in octopus spring, yellowstone national park. *Applied and environmental microbiology*, 60(6):2113–2119, 1994.

[19] James Borneman, Paul W Skroch, Katherine M O'Sullivan, James A Palus, Norma G Rumjanek, Jennifer L Jansen, James Nienhuis, and Eric W Triplett. Molecular microbial diversity of an agricultural soil in wisconsin. *Applied and Environmental Microbiology*, 62(6):1935–1943, 1996.

[20] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson,

William Nelson, et al. Environmental genome shotgun sequencing of the sargasso sea. *science*, 304(5667):66–74, 2004.

[21] Christopher A Francis, Kathryn J Roberts, J Michael Beman, Alyson E Santoro, and Brian B Oakley. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences*, 102(41):14683–14688, 2005.

[22] Christopher T Brown, Laura A Hug, Brian C Thomas, Itai Sharon, Cindy J Castelle, Andrea Singh, Michael J Wilkins, Kelly C Wrighton, Kenneth H Williams, and Jillian F Banfield. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523(7559):208–211, 2015.

[23] Benjamin Hillmann, Gabriel A Al-Ghalith, Robin R Shields-Cutler, Qiyun Zhu, Daryl M Gohl, Kenneth B Beckman, Rob Knight, and Dan Knights. Evaluating the information content of shallow shotgun metagenomics. *Msystems*, 3(6):10–1128, 2018.

[24] Rahat Zaheer, Noelle Noyes, Rodrigo Ortega Polo, Shaun R Cook, Eric Marinier, Gary Van Domselaar, Keith E Belk, Paul S Morley, and Tim A McAllister. Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific reports*, 8(1):5890, 2018.

[25] H Soon Gweon, Liam P Shaw, Jeremy Swann, Nicola De Maio, Manal AbuOun, Rene Niehus, Alasdair TM Hubbard, Mike J Bowes, Mark J Bailey, Tim EA Peto, et al. The impact of sequencing depth on the inferred taxonomic composition and amr gene content of metagenomic samples. *Environmental Microbiome*, 14(1):7, 2019.

[26] Joana Pereira-Marques, Anne Hout, Rui M Ferreira, Michiel Weber, Ines Pinto-Ribeiro, Leen-Jan Van Doorn, Cornelis Willem Knetsch, and Ceu Figueiredo. Impact of host dna and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in microbiology*, 10:1277, 2019.

[27] Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S McGee, and David L Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977, 2016.

[28] Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, 37(8):852–857, 2019.

[29] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[30] Suzanne M Prober, Jonathan W Leff, Scott T Bates, Elizabeth T Borer, Jennifer Firn, W Stanley Harpole, Eric M Lind, Eric W Seabloom, Peter B Adler, Jonathan D Bakker, et al. Plant diversity predicts beta but not

alpha diversity of soil microbes across grasslands worldwide. *Ecology letters*, 18(1):85–95, 2015.

[31] Emily E Curd, Jennifer BH Martiny, Huiying Li, and Thomas B Smith. Bacterial diversity is positively correlated with soil heterogeneity. *Ecosphere*, 9(1):e02079, 2018.

[32] Kyle M Meyer, Robert Porch, Isabella E Muscettola, Ana Luisa S Vasconcelos, Julia K Sherman, C Jessica E Metcalf, Steven E Lindow, and Britt Koskella. Plant neighborhood shapes diversity and reduces interspecific variation of the phyllosphere microbiome. *The ISME journal*, 16(5):1376–1387, 2022.

[33] Matthias Noll, Diethart Matthies, Peter Frenzel, Manigee Derakshani, and Werner Liesack. Succession of bacterial community structure and diversity in a paddy soil oxygen gradient. *Environmental Microbiology*, 7(3):382–395, 2005.

[34] Daniel PR Herlemann, Matthias Labrenz, Klaus Jürgens, Stefan Bertilsson, Joanna J Waniek, and Anders F Andersson. Transitions in bacterial communities along the 2000 km salinity gradient of the baltic sea. *The ISME journal*, 5(10):1571–1579, 2011.

[35] Márton Mucsi, Andrea K Borsodi, Melinda Megyes, and Tibor Szili-Kovács. Response of the metabolic activity and taxonomic composition of bacterial communities to mosaically varying soil salinity and alkalinity. *Scientific Reports*, 14(1):7460, 2024.

[36] MS Girvan, CD Campbell, K Killham, James Ivor Prosser, and Lesley Anne Glover. Bacterial diversity promotes community stability and functional resilience after perturbation. *Environmental microbiology*, 7(3):301–313, 2005.

[37] Melanie Y Sun, Katherine A Dafforn, Emma L Johnston, and Mark V Brown. Core sediment bacteria drive community response to anthropogenic contamination over multiple environmental gradients. *Environmental microbiology*, 15(9):2517–2531, 2013.

[38] Luke R Thompson, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, Anupriya Tripathi, Sean M Gibbons, Gail Ackermann, et al. A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017.

[39] William B. Whitman, Paul DeVos, Svetlana Dedysh, Brian Hedlund, Peter Kämpfer, Fred Rainey, and Martha E. Trujillo, editors. *Bergey's Manual of Systematics of Archaea and Bacteria*. John Wiley & Sons, Inc., Hoboken, NJ, 2015. Updated continuously online.

[40] P.G. Falkowski, T. Fenchel, and E.F. Delong. The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879):1034–1039, May 2008.

[41] Scott D Bridgham, Hinsby Cadillo-Quiroz, Jason K Keller, and Qianlai Zhuang. Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. *Global change biology*, 19(5):1325–1346, 2013.

[42] Michael WI Schmidt, Margaret S Torn, Samuel Abiven, Thorsten Dittmar, Georg Guggenberger, Ivan A Janssens, Markus Kleber, Ingrid Kögel-Knabner,

Johannes Lehmann, David AC Manning, et al. Persistence of soil organic matter as an ecosystem property. *Nature*, 478(7367):49–56, 2011.

[43] Philippe D Tortell, Maria T Maldonado, Julie Granger, and Neil M Price. Marine bacteria and biogeochemical cycling of iron in the oceans. *FEMS Microbiology Ecology*, 29(1):1–11, 1999.

[44] Melanie A Spero and Dianne K Newman. Chlorate specifically targets oxidant-starved, antibiotic-tolerant populations of Pseudomonas aeruginosa biofilms. *MBio*, 9(5):10–1128, 2018.

[45] Hussein E Al-Hazmi, Gamal K Hassan, Mojtaba Maktabifard, Dominika Grubba, Joanna Majtacz, and Jacek Mąkinia. Integrating conventional nitrogen removal with anammox in wastewater treatment systems: Microbial metabolism, sustainability and challenges. *Environmental research*, 215:114432, 2022.

[46] J. Ru, Y. Huo, and Y. Yang. Microbial degradation and valorization of plastic wastes. *Frontiers in Microbiology*, 11(442), 2020.

[47] C. Rinke, P. Schwientek, A. Sczyrba, N.N. Ivanova, I.J. Anderson, J. Cheng, A. Darling, S. Malfatti, B.K. Swan, E.A. Gies, J.A. Dodsworth, B.P. Hedlund, G. Tsiamis, S.M. Sievert, W. Liu, J.A. Eisen, S.J. Hallam, N.C. Kyrpides, R. Stepanauskas, E.M. Rubin, P. Hugenholtz, and T. Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499:431–437, July 2013.

[48] V.H.T. Pham and J. Kim. Cultivation of unculturable soil bacteria. *Trends in biotechnology*, 30(9):475–484, 2012.

[49] María A Sánchez-Romero and Josep Casadesús. The bacterial epigenome. *Nature Reviews Microbiology*, 18(1):7–20, 2020.

[50] Wiep Klaas Smits, Oscar P Kuipers, and Jan-Willem Veening. Phenotypic variation in bacteria: the role of feedback regulation. *Nature Reviews Microbiology*, 4(4):259–271, 2006.

[51] A.M. Schnoes, S.D. Brown, I. Dodevski, and P.C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12):e1000605, 2009.

[52] B. Lobb, B.J. Tremblay, G. Moreno-Hagelsieb, and A.C. Doxey. An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*, 6(3):e000341, 2020.

[53] P. Escudeiro, C.S. Henry, and R.P.M. Dias. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Current Research in Microbial Sciences*, 3:100159, 2022.

[54] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016.

[55] A. Barberán, H.C. Velazquez, S. Jones, and N. Fierer. Hiding in plain sight: mining bacterial species records for phenotypic trait information. *mSphere*, 2(4), July/August 2017.

[56] J.S. Madin, D.A. Nielsen, M. Brbic, R. Corkrey, D. Danko, K. Edwards, M.K.M. Engqvist, N. Fierer, J.L. Geoghegan, M. Gillings, N.C. Kyrpides, E. Litchman, C.E. Mason, L. Moore, S.L. Nielsen, I.T. Paulsen, N.D. Price,

T.B.K. Reddy, M.A. Richards, E.P.C. Rocha, T.M. Schmidt, H. Shaaban, M. Shukla, F. Supek, S.G. Tetu, S. Vieira-Silva, A.R. Wattam, D.A. Westfall, and M. Westoby. A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data*, 7(170), 2020.

[57] I. Schober, J. Koblitz, J.S. Carbasse, C. Ebeling, M.L. Schmidt, A. Podstawka, R. Gupta, V. Ilangovan, J. Chamanara, J. Overmann, and L.C. Reimer. BacDive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Research*, 53(D1):D748–D756, 2025.

[58] Anthony D Bertagnolli and Frank J Stewart. Microbial niches in marine oxygen minimum zones. *Nature Reviews Microbiology*, 16(12):723–729, 2018.

[59] Jody J Wright, Kishori M Konwar, and Steven J Hallam. Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology*, 10(6):381–394, 2012.

[60] Rachel C Beaver and Josh D Neufeld. Microbial ecology of the deep terrestrial subsurface. *The ISME Journal*, 18(1):wrae091, 2024.

[61] Victor Gambarini, Olga Pantos, Joanne M Kingsbury, Louise Weaver, Kim M Handley, and Gavin Lear. PlasticDB: a database of microorganisms and proteins linked to plastic biodegradation. *Database*, 2022:baac008, 2022.

[62] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.

[63] Fortunato Pesarin and Luigi Salmaso. The permutation testing approach: a review. *Statistica*, 70(4):481–509, 2010.

[64] Phillip Good. *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005.

[65] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

[66] A.C. Martiny, K. Treseder, and G. Pusch. Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal*, 7:830–838, April 2013.

[67] Om Prakash, Mrinalini Parmar, Manali Vaijanapurkar, Vinay Rale, and Yogesh S Shouche. Recent trend, biases and limitations of cultivation-based diversity studies of microbes. *FEMS microbiology letters*, 368(17):fnab118, 2021.

[68] Philip Hugenholtz, Brett M Goebel, and Norman R Pace. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, 180(18):4765–4774, 1998.

[69] Robert R Sokal, Charles D Michener, et al. A statistical method for evaluating systematic relationships. 1958.

[70] Petr Baldrian, Miroslav Kolařík, Martina Štursová, Jan Kopeckỳ, Vendula Valášková, Tomáš Větrovskỳ, Lucia Žifčáková, Jaroslav Šnajdr, Jakub Rídl, Čestmír Vlček, et al. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME journal*, 6(2):248–258, 2012.

[71] Caitlin M Singleton, Carmody K McCalley, Ben J Woodcroft, Joel A Boyd, Paul N Evans, Suzanne B Hodgkins, Jeffrey P Chanton, Steve Frolking, Patrick M Crill, Scott R Saleska, et al. Methanotrophy across a natural permafrost thaw environment. *The ISME journal*, 12(10):2544–2558, 2018.

[72] Sage Albright and Stilianos Louca. Trait biases in microbial reference genomes. *Scientific Data*, 10(1):84, 2023.

[73] Arthur Escalas, Fabiana S Paula, François Guilhaumon, Mengting Yuan, Yunfeng Yang, Linwei Wu, Feifei Liu, Jiaje Feng, Yuguang Zhang, and Jizhong Zhou. Macroecological distributions of gene variants highlight the functional organization of soil microbial systems. *The ISME Journal*, 16(3):726–737, 2022.

[74] Jörg Overmann, Birte Abt, and Johannes Sikorski. Present and future of culturing bacteria. *Annual review of microbiology*, 71(1):711–730, 2017.

[75] S Khelaifia, J-C Lagier, VD Nkamga, E Guilhot, Michel Drancourt, and Didier Raoult. Aerobic culture of methanogenic archaea without an external source of hydrogen. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(6):985–991, 2016.

[76] Tariq Ahmad, Sidra Erum Ishaq, Lewen Liang, Ruize Xie, Yinzhao Wang, and Fengping Wang. A diffusion-based integrative approach for culturing previously uncultured bacteria from marine sediments. *Marine Life Science & Technology*, 7(3):466–477, 2025.

[77] Dhiraj Kumar Chaudhary, Altankhuu Khulan, and Jaisoo Kim. Development of a novel cultivation technique for uncultured soil bacteria. *Scientific reports*, 9(1):6666, 2019.

[78] MV Sizova, T Hohmann, A Hazen, BJ Paster, SR Halem, CM Murphy, NS Panikov, and SS3255620 Epstein. New approaches for isolation of previously uncultivated oral bacteria. *Applied and environmental microbiology*, 78(1):194–203, 2012.

[79] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[80] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1):D309–D314, 2019.

[81] Georgios A Pavlopoulos, Fotis A Baltoumas, Sirui Liu, Oguz Selvitopi, Antonio Pedro Camargo, Stephen Nayfach, Ariful Azad, Simon Roux, Lee Call, Natalia N Ivanova, et al. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983):594–602, 2023.

[82] Regina Joice, Koji Yasuda, Afrah Shafquat, Xochitl C Morgan, and Curtis Huttenhower. Determining microbial products and identifying molecular targets in the human microbiome. *Cell metabolism*, 20(5):731–741, 2014.

[83] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S Pollard, Ekaterina

Sakharova, Donovan H Parks, Philip Hugenholtz, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology*, 39(1):105–114, 2021.

[84] Yancong Zhang, Amrisha Bhosle, Sena Bae, Lauren J McIver, Gleb Pishchany, Emma K Accorsi, Kelsey N Thompson, Cesar Arze, Ya Wang, Ayshwarya Subramanian, et al. Discovery of bioactive microbial gene products in inflammatory bowel disease. *Nature*, 606(7915):754–760, 2022.

[85] Wayne P Maddison, Peter E Midford, and Sarah P Otto. Estimating a binary character's effect on speciation and extinction. *Systematic biology*, 56(5):701–710, 2007.

[86] Hervé Sauquet, Maria Von Balthazar, Susana Magallón, James A Doyle, Peter K Endress, Emily J Bailes, Erica Barroso de Morais, Kester Bull-Hereñu, Laetitia Carrive, Marion Chartier, et al. The ancestral flower of angiosperms and its early diversification. *Nature communications*, 8(1):16047, 2017.

[87] Torda Varga, Krisztina Krizsán, Csenge Földi, Bálint Dima, Marisol Sánchez-García, Santiago Sánchez-Ramírez, Gergely J Szöllősi, János G Szarkándi, Viktor Papp, László Albert, et al. Megaphylogeny resolves global patterns of mushroom evolution. *Nature ecology & evolution*, 3(4):668–678, 2019.

[88] Michael G Harvey, Gustavo A Bravo, Santiago Claramunt, Andrés M Cuervo, Graham E Derryberry, Jaqueline Battilana, Glenn F Seeholzer, Jessica Shearer McKay, Brian C O'Meara, Brant C Faircloth, et al. The evolution of a tropical biodiversity hotspot. *Science*, 370(6522):1343–1348, 2020.

[89] Bonnie B Blaimer, Bernardo F Santos, Astrid Cruaud, Michael W Gates, Robert R Kula, István Mikó, Jean-Yves Rasplus, David R Smith, Elijah J Talamas, Seán G Brady, et al. Key innovations and the diversification of Hymenoptera. *Nature communications*, 14(1):1212, 2023.

[90] Stilianos Louca, Martin F Polz, Florent Mazel, Michaeline BN Albright, Julie A Huber, Mary I O'Connor, Martin Ackermann, Aria S Hahn, Diane S Srivastava, Sean A Crowe, et al. Function and functional redundancy in microbial systems. *Nature ecology & evolution*, 2(6):936–943, 2018.

[91] Olga Khersonsky, Cintia Roodveldt, and Dan S Tawfik. Enzyme promiscuity: evolutionary and mechanistic aspects. *Current opinion in chemical biology*, 10(5):498–508, 2006.

[92] Margaret E Glasner, Dat P Truong, and Benjamin C Morse. How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation. *The FEBS journal*, 287(7):1323–1342, 2020.

[93] Timothy D Read and Ruth C Massey. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome medicine*, 6(11):109, 2014.

[94] Gavin M Douglas, Vincent J Maffei, Jesse R Zaneveld, Svetlana N Yurgel, James R Brown, Christopher M Taylor, Curtis Huttenhower, and Morgan GI Langille. PICRUSt2 for prediction of metagenome functions. *Nature biotechnology*, 38(6):685–688, 2020.

[95] Franziska Wemheuer, Jessica A Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer. Tax4Fun2: prediction of

habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental microbiome*, 15(1):11, 2020.

[96] R. Porch, T. J. Taylor, A. Nassib, L. Versteeg, A. C. Ellis, L. Keane, M. Roy-Hart, E. S. Grunkemeier, P. C. Stephens, A. M. Ryan, S. Williams, W. Qi, O. Loreto, L. Grover, J. Peairson, E. Simpson, J. S. Schnider, A. M. Smith, B. Pederson, J. A. Hoskins, L. Bonnet, and S. Louca. FAPROTAX2-db: A database of experimentally validated bacterial and archaeal metabolic functions. *in review*, 2026.

[97] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[98] Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic biology*, 68(2):365–369, 2019.

[99] Metin Balaban, Shahab Sarmashghi, and Siavash Mirarab. APPLES: scalable distance-based phylogenetic placement with or without alignments. *Systematic Biology*, 69(3):566–578, 2020.

[100] Jesse RR Zaneveld and Rebecca LV Thurber. Hidden state prediction: a modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. *Frontiers in microbiology*, 5:431, 2014.

[101] Stilianos Louca and Michael Doebeli. Efficient comparative phylogenetics on large trees. *Bioinformatics*, 34(6):1053–1055, 2018.

[102] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[103] Masha V. Korchagina, Claire E. Mullin, Hengameh H. Soufi, Sophia Lambert, Ines G. Moran, Robert Porch, Sage E. Albright, Alexandria S. Montgomery, Leila M. Jones, Nathan Malamoud, Qusheng Jin, A. Michelle Wood, and Stilianos Louca. A genome-resolved metagenomic survey of 500 samples from 56 hot springs across the Western US. *Nature Scientific Data*, In review.

[104] Stilianos Louca, Saulo MS Jacques, Aliny PF Pires, Juliana S Leal, Diane S Srivastava, Laura Wegener Parfrey, Vinicius F Farjalla, and Michael Doebeli. High taxonomic variability despite stable functional structure across microbial communities. *Nature ecology & evolution*, 1(1):0015, 2016.

[105] Stilianos Louca and Claire E Mullin. Effects of vacuum-heat-assisted sample desiccation on microbiome surveys. *Molecular Ecology Resources*, 25(7):e70020, 2025.

[106] William Walters, Embriette R Hyde, Donna Berg-Lyons, Gail Ackermann, Greg Humphrey, Alma Parada, Jack A Gilbert, Janet K Jansson, J Gregory Caporaso, Jed A Fuhrman, et al. Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *Msystems*, 1(1):10–1128, 2016.

[107] Alma E Parada, David M Needham, and Jed A Fuhrman. Every base matters: assessing small subunit rrna primers for marine microbiomes with mock communities, time series and global field samples. *Environmental microbiology*, 18(5):1403–1414, 2016.

[108] Hengameh H Soufi, Daniel Tran, and Stilianos Louca. Microbiology of Big Soda Lake, a multi-extreme meromictic volcanic crater lake in the Nevada desert. *Environmental Microbiology*, 26(2):e16578, 2024.

[109] Hongshan Jiang, Rong Lei, Shou-Wei Ding, and Shuifang Zhu. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics*, 15(1):182, 2014.

[110] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

[111] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[112] Hengameh H Soufi, Robert Porch, Masha V Korchagina, Joseph A Abrams, Jared S Schnider, Ben D Carr, Mark A Williams, and Stilianos Louca. Taxonomic variability and functional stability across Oregon coastal subsurface microbiomes. *Communications Biology*, 7(1):1663, 2024.

[113] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[114] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020.

[115] Jaina Mistry, Robert D Finn, Sean R Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research*, 41(12):e121–e121, 2013.

[116] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[117] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[118] Ya-Jou Chen, Pok Man Leung, Jennifer L Wood, Sean K Bay, Philip Hugenholtz, Adam J Kessler, Guy Shelley, David W Waite, Ashley E Franks, Perran LM Cook, et al. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *The ISME journal*, 15(10):2986–3004, 2021.

[119] Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, et al. A genomic catalog of Earth's microbiomes. *Nature biotechnology*, 39(4):499–509, 2021.

[120] Alex Mira, Ana B Martín-Cuadrado, Giuseppe D'Auria, and Francisco Rodríguez-Valera. The bacterial pan-genome: a new paradigm in microbiology. *Int Microbiol*, 13(2):45–57, 2010.

[121] Chiara Vanni, Matthew S Schechter, Silvia G Acinas, Albert Barberán, Pier Luigi Buttigieg, Emilio O Casamayor, Tom O Delmont, Carlos M

Duarte, A Murat Eren, Robert D Finn, et al. Unifying the known and unknown microbial coding sequence space. *Elife*, 11:e67667, 2022.

[122] Álvaro Rodríguez del Río, Joaquín Giner-Lamia, Carlos P Cantalapiedra, Jorge Botas, Ziqi Deng, Ana Hernández-Plaza, Martí Munar-Palmer, Saray Santamaría-Hernando, José J Rodríguez-Herva, Hans-Joachim Ruscheweyh, et al. Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature*, 626(7998):377–384, 2024.