ADVANCING GENE-CENTRIC APPROACHES FOR MICROBIAL ECOLOGY

by

Ryan M. Moore

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics Data Science

Winter 2024

ADVANCING GENE-CENTRIC APPROACHES

FOR MICROBIAL ECOLOGY

by

Ryan M. Moore

Approved: _____

Cathy H. Wu, Ph.D.
Chair of the Department of Bioinformatics Data Science

Approved: _____

Calvin L. Keeler, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____

Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

K. Eric Wommack, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Shawn W. Polson, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Jennifer F. Biddle, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Nicole M. Donofrio, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to those who have had a significant impact on the completion of this dissertation. Their mentorship, guidance, contributions, and support have been invaluable.

In particular, I extend my thanks to: Eric Wommack and Shawn Polson, my research advisors, for their unwavering guidance, patience, and expertise; Jen Biddle and Nicole Donofrio, my committee members, for their insightful feedback and encouragement; and my fellow lab members, peers, and collaborators, course instructors, and Bioinformatics Data Science faculty and staff for their collegiality, enriching discussions, and academic instruction.

Finally, I would like to make special mention of the continuous support of my family and friends. Their unwavering encouragement and support has made this endeavor possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Metagenomics is a powerful approach that has enhanced our understanding of microbial communities and the roles microbes play in various environments. A deep examination of single genes, particularly protein-coding genes, can add critical insight to metagenomic datasets by providing functional information and allowing for the prediction of observable traits and the formulation of "genome to phenome" hypotheses. However, gene-centric approaches to metagenomics face unique challenges, and the comparative lack of tools and approaches specifically designed to address these problems makes gene-centric analyses of microbial communities less accessible to many researchers. Though data quality issues arise at all stages of the sample-to-sequence-to-discovery pipeline, gene-centric studies are particularly sensitive to issues such as those arising from misannotations of the genes under study, which necessitates time-consuming manual curation, or from the compositional nature of metagenomic data, which requires special statistical care. To address some of the barriers to effective gene-centric analysis in metagenomics, this dissertation introduces three tools: PASV, InteinFinder, and Iroki, as well as a novel framework for examining microbial community diversity. PASV (**p**rotein **a**mino acid **s**ignature **v**alidator) automates the manual curation of homology search results to ensure accurate protein annotation. InteinFinder is a pipeline developed to automatically identify and remove inteins, the protein equivalent of introns, from protein sequences commonly used in gene-centric studies. Together, PASV and InteinFinder significantly reduce the amount of time and domain-knowledge traditionally needed to manually curate single gene datasets. Iroki is a user-friendly tool designed to automatically customize phylogenetic and other types of trees with user supplied metadata, facilitating data interpretation. The introduced diversity framework provides a more comprehensive and scalable view of microbial community

diversity compared to current approaches, particularly for large metagenomic datasets. Overall, these advancements simplify the gene-centric study of microbial communities and enhance the metagenomic analysis pipeline.

Chapter 1

INTRODUCTION

Metagenomics first arose in the 1990s as a method for studying unculturable microbes [356, 131]. The field has since grown dramatically, facilitated by advancements in sequencing technologies. This powerful approach has played a crucial role in the discovery of novel taxa [275, 272] as well as metabolic and functional diversity [25, 137], enhancing our understanding of the roles microbes play in various environments. Metagenomics has expanded our understanding of the microbial contribution to disparate processes from biogeochemical cycling [43, 377, 125] to human health [387, 225].

As the field of metagenomics matures, there is a growing recognition that it is necessary to move beyond high-level surveys and conduct more in-depth analysis. Presently, metagenomic studies typically involve clustering peptide sequences, assembled DNA fragments (contigs) [11], or metagenome-assembled genomes (MAGs) [337] to form population clusters and examining diversity at the population level [124]. However, this broad perspective can sometimes limit the conclusions that can be drawn.

While this data is often used to predict high-level metabolic potential and pathway enrichment of microbial communities (e.g., [11, 415, 340, 405, 127]), a deep examination of single genes, particularly protein-coding genes with extensive biochemical characterization, can add critical insight to the processes that shape microbial communities, adding ecological, evolutionary, and physiological context to metagenomic datasets. Protein-coding genes provide valuable functional information, allowing for the prediction of observable traits and the formulation of "genome to phenome" hypotheses in addition to identifying taxa. For example, studies of DNA polymerase A genes in viral metagenomes have demonstrated a connection between phage lifestyle

and a single amino acid mutation at a critical active site [333, 168]. Genes that have the potential to connect a microbe to its environment are especially interesting to ecologists. For example, researchers can investigate the functional potential of specific aspects of microbial communities by examining the presence and distribution of specific metabolic genes in an environment across taxa [242].

1.1 Challenges in single gene analysis

Single gene approaches have inherent limitations, particularly in the context of viruses where there are no universally shared genes. For example, only about a quarter of viruses are estimated to carry DNA polymerase A genes [395], meaning that any analysis based on these genes covers only a subset of the viral community. However, this narrower scope can allow for more thorough and detailed investigations that incorporate biochemical and functional characterization of the chosen enzymes [333, 224, 251, 135, 168, 401].

Although single gene datasets have a narrower scope than full metagenomic datasets, they still pose distinct challenges that must be addressed in order to fully leverage them.

The first, and potentially largest issue lies in data quality. Gene annotation in metagenome studies typically involves inputting contigs or MAGs into genome annotation software [336] or manually performing homology searches for translated protein sequences against large databases [124]. However, databases commonly used for this purpose such as KEGG, Gene Ontology, and GenBank NR contain high levels of misannotation [155, 335], making this step highly error-prone. This error rate is often acceptable for large metagenomic studies where the focus is at the population level or on predicting high-level metabolic patterns. However, when the focus moves to the level of individual protein-coding genes, accurate annotation is crucial to avoid erroneous conclusions–a case of avoiding the classic "garbage in, garbage out" problem.

Researchers looking at single genes may try to avoid this by setting strict similarity cutoffs for homology, but this runs the risk of missing valuable data because of

low sequence similarity, and still does not protect against misannotated sequences in a query database. Low sequence similarity can especially be an issue for protein coding genes, which may be less conserved than structural genes such as SSU rRNA [150]. For example, while all ribonucleotide reductases share a common ancestor, many show similarities far below the "twilight zone" cutoff of sequence similarity [317, 212]. Low sequence similarity is also a problem when working with viruses, whose genes may have only distant homology to their cellular counterparts [135]. In fact, many viral genes in metagenomic surveys are completely unknown and are even referred to as "viral dark matter" [124, 330].

The effectiveness of homology-based annotations is partially determined by the comprehensiveness, or lack thereof, of the databases used [414, 370, 79]. Databases are biased towards certain environments and organisms, and against others. Many of the common databases suffer from a "research bias", that is, an overrepresentation of sequences similar to early model organisms, culturable microbes, and pathogens, that often leads to poor annotation performance for many environmentally important groups [29, 204]. These database issues can make annotation of environmental sequences more challenging. It is partially to address these concerns that environmental gene catalogs are being compiled (e.g., [253, 406, 216]).

In summary, database errors and bias, and the sequence diversity of the target gene, the target community (e.g., cellular microbes versus viruses), and sampling environment, are all factors that make recovering target genes challenging. Researchers who want to recover as many of their target genes as possible must therefore use highly sensitive homology searches, which increase false positive rates and the number of sequences that must be manually curated to ensure annotation quality. Although expert curation is the gold standard, the process is time-consuming, still error-prone, and increases in difficulty with the size of sequencing datasets. Barriers to manual curation of metagenomic-derived gene datasets are therefore growing and may eventually become an insurmountable obstacle.

Another obstacle to these types of studies is their relative rarity. Microbial

amplicon and metagenomic studies are abundant and popular, which has led to a multitude of tools developed for analysis purposes. For example, the QIIME platform, designed to make amplicon sequence analysis accessible to more researchers, has over forty thousand citations [52, 38]. Comparable tools have not been specifically designed for gene-centric approaches in metagenomes. Instead, researchers must construct bespoke analysis pipelines, from the initial homology search through to diversity analysis, which can be a significant barrier to entry.

A central issue of executing gene-centric studies with metagenomic data therefore lies in accessibility. First, the need for manual annotation means that a high level of domain knowledge is currently required to perform this type of analysis. Second, the growing size of metagenomic datasets is surpassing what current methods can handle, and also makes analyses increasingly difficult for researchers without access to high performance computing resources. And finally, because of lower general interest in this area, there is less development of new tools of methods.

1.2 Addressing these challenges

In response, I have developed a suite of software tools and a new framework for measuring diversity of microbial communities that lower these barriers along the sample-to-sequence-to-discovery pipeline with the goal of making gene-centric analyses more accessible to the scientific community (Fig. 1.1). Each chapter of this dissertation will focus on one of these advancements, beginning with earlier stages of analysis and progressing through the latter. In addition to introducing the tool or framework, each chapter is grounded in practical examples and applications of the methodological advances therein.

First, to address homology search issues, I developed a pipeline for automatic protein partitioning and validation using amino acid signatures, PASV (**p**rotein **a**mino acid **s**ignature **v**alidator). Accurate annotation of proteins is critical, and yet, sensitive homology searches are plagued by the same garbage in, garbage out issues mentioned

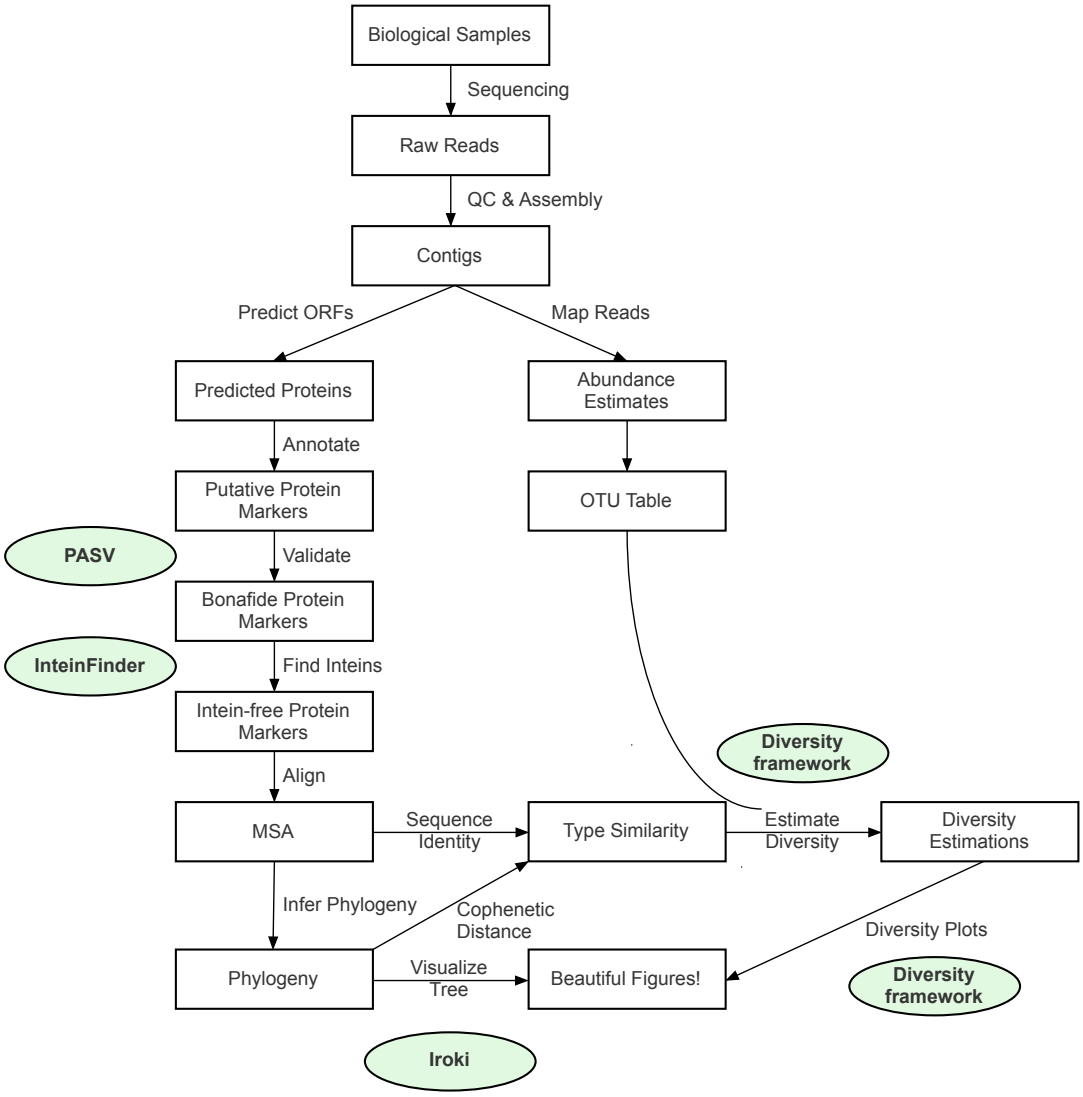

Figure 1.1: An idealized metagenomic sample-to-sequence-to-discovery pipeline. Green ovals represent contributions of this dissertation to the metagenomic sample-to-sequence-to-discovery pipeline.

earlier. Thus, manual curation of homology search results is critical for accurate downstream analyses, even in the context of huge metagenomic datasets with potentially intractable numbers of proteins. Curation generally involves the inspection of sequence alignments for the presence of active sites, binding sites, and other conserved residues. This requires relatively deep domain knowledge regarding the gene of interest and, ideally, alignment visualization software. Once provided with a file of reference sequences and a list of sites to survey, PASV is able to automate the process of manual curation. While some domain knowledge is still required (e.g., positions of conserved residues), this method is shown to be less error-prone and much faster than manual curation. To demonstrate PASV's utility, I analyze misannotations and post-homology search validation in multiple protein-coding genes, and compare the results of the pipeline with that of expert curation.

To help further curate protein sequences, I developed InteinFinder. Inteins, the protein equivalent of introns, are common in certain types of protein coding genes [328]. As mobile elements, they can have separate evolutionary histories from their host proteins and can therefore confound downstream analyses such as phylogenetic inference. Consequently, they must be removed from protein sequences prior to analysis. However, they can be difficult to identify and remove correctly for a researcher who is not familiar with either inteins or the host protein, and so they are likely to be missed during curation. InteinFinder allows for the automatic identification and removal of inteins from protein sequence, saving researchers time and making this step more accessible by lessening the need for expert domain knowledge. I demonstrate the utility of InteinFinder on a dataset of more than 100 million viral proteins, and chart the abundance of inteins across ecosystems and investigate their potential ecological impact on viruses across the biosphere.

To assist in interpretation of data, I created Iroki, which allows for automatic customization and visualization of phylogenetic and other types of trees. While there are several popular tree viewers already in common usage, they either require manual mapping of metadata or have a steep learning curve. Iroki enables researchers to

display a variety of metadata on large phylogenetic trees, allowing them to connect metadata with a summary of community diversity. Iroki was designed to be user-friendly, making gene-centric analyses more accessible to more researchers. Using Iroki, I explore relationships between abiotic factors and samples from the worlds oceanic virome, SSU rRNA abundance, phylogeny, and correlations with metadata in cattle, and phage-host interactions using the phage proteomic tree.

Finally, I developed a diversity framework that provides a more complete view of diversity than is available with current tools and is scalable to large datasets. Diversity is a high-level summary of community structure. Accurate diversity estimates are critical for understanding the effects of treatments and interventions on microbial communities, particularly in human health. However, typical plug-in diversity measures employed in macroecology tend to mishandle the distinctive features of sequencing data [393]. Additionally, accurate estimates of the variance of the diversity measurements is necessary for hypothesis testing. To address these issues, I reengineered a state-of-the-art model for estimating diversity (DivNet [393]) to handle larger metagenomic datasets on commodity hardware, opening up use of the method to users without access to high performance computing.

Different aspects of diversity (e.g., types vs. traits, abundant vs. rare members) give different views into communities. Combining these features with the use of various gene markers enhances the understanding of diversity and its relationship to the experimental or observed conditions of interest. Different gene markers "sense" the environment differently–that is, their different biochemical physiologies are connected in different ways to the ecology of the organisms in which they are present. In a similar way, different weights on similarity between types or abundance of types [304] "sense" the measured diversity in different ways by emphasizing different aspects of the community. Tuning these components therefore allows a measure of community diversity tuned to the specific research questions or outcomes of interest. I then leverage the reengineered DivNet model with varying similarity and abundance viewpoints of multiple gene markers to explore the diversity of cattle hide and fecal microbiomes.

PREFACE TO CHAPTER 2

The work presented in Chapter 2 of this dissertation has previously been published on the bioRxiv with myself as the first author [245]. The original author contribution list published in the manuscript used the CRediT contributor roles taxonomy [1] and is reproduced in full here:

- Ryan M. Moore: Conceptualization, Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing
- Amelia O. Harrison: Conceptualization, Data curation, Writing – review & editing
- Metehan Cebeci: Conceptualization, Data curation, Visualization
- Daniel J. Nasko: Conceptualization, Writing – review & editing
- Jessica Chopyk: Conceptualization, Writing – review & editing
- Barbra D. Ferrell: Conceptualization, Supervision, Writing – review & editing
- Shawn W. Polson: Conceptualization, Funding acquisition, Supervision, Writing – review & editing
- K. Eric Wommack: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

For additional files, refer to the bioRxiv submission [245].

Chapter 2

PASV: AUTOMATIC PROTEIN PARTITIONING AND VALIDATION USING CONSERVED RESIDUES

2.1 Background

Next generation DNA sequencing has continued to yield ever larger sequence datasets, enabling researchers to leverage vast amounts of sequence data in addressing a variety of scientific questions from cataloguing variation in human genomes [371] and connecting the gut microbiome with human health [203] to examining the circadian clock in soybean [199] and surveying viruses of the global ocean [124]. For example, sequencing has led to substantial advancements in understanding the community and population biology of microorganisms in nature. Nevertheless, while generation of data continually improves, accurate and comprehensive data analysis remains a challenge for investigations leveraging large sequence datasets.

Building on the example of microbial ecology, for decades researchers have relied on sequence based surveys of stable RNA genes, such as SSU rRNA, as phylogenetic markers for assessing the composition of cellular microbial communities. However, the focus on stable and highly conserved RNA gene sequences for microbial ecology studies has limited researcher's ability for fine scale delineation of cellular microbial populations from one another [172, 31] and identification of viral populations which do not encode SSU rRNA genes [240]. Use of protein-coding gene sequences as phylogenetic markers for community and population ecology studies can address these shortcomings of SSU rRNA analyses. However, accurate identification of protein-coding genes from either targeted amplicon libraries or shotgun metagenomes remains a significant analytical challenge.

In microbial ecology investigations, both stable rRNA and protein coding marker gene sequences are obtained either through targeted PCR amplification or direct sequencing (i.e., shotgun metagenome sequencing) of environmental DNA. Either approach has limitations that are addressed by the other. Targeted PCR amplification can deeply sample microbial populations within a community, detecting even the rarest of members; however, this approach may miss novel diversity by relying on previously sequenced genes for constructing PCR primers [269, 20, 416]. While every effort is made to ensure marker gene primers capture as much diversity as possible, amplification bias is always present [234]. In contrast, metagenome sequence libraries from shotgun sequencing provide a relatively unbiased picture of microbial diversity, with the caveat of a more limited ability for sampling rare populations [399, 417]. With sequence assembly, this approach also provides the genomic context of marker genes, highly useful information for genome to phenome investigations [251]. Nevertheless, shotgun metagenomics presents significant additional analytical and computational requirements making this approach more expensive and difficult [363, 252]. Furthermore, researchers still must drill down to the level of specific genes within metagenomes, such as those that have undergone extensive biochemical characterization, to uncover interesting biological and ecological patterns from the sequencing data [334, 325, 224, 60]. In the case of either approach, accurately determining the identity of a sequence is critical in preventing subsequent errors in phylogenetic and functional analyses.

Assessing the potential gene functions within a community requires annotation of peptide sequences within metagenomes. Homology-based search tools such as BLAST [9] are the bedrock of sequence annotation, however, functional annotation of proteins based on homology can be error prone [335, 135]. Biochemically annotated proteins are relatively rare in major databases, and usually arise from studies of a few select model organisms [112, 311, 351]. As a result, many environmental sequences are annotated based solely on homology to other computationally annotated environmental sequences rather than to biochemically characterized proteins. Often, such environmental sequence annotations are several steps away from a confident, biochemical

annotation, which can quickly lead to inaccuracies resulting from "error percolation" [112].

Furthermore, highly sensitive homology search tools used for annotating and identifying marker genes within metagenomes can often have high false positive rates [166]. Identifying false positives in functional annotations is an active area of research and many techniques are available. Machine learning algorithms have been used for identifying false positives based on characteristics of multiple sequence alignments (MSAs) [108, 109]. Active site profiling, or examining the characteristics of regions close to a protein's active sites, has been used for sensitive and functionally relevant annotations [97, 196, 133, 171].

Even with accurate functional annotations, researchers need a means for predicting if a peptide sequence represents a functional enzyme. While a protein's function cannot be definitively determined *in silico*, evidence can be gathered by examining active sites, allosteric sites, and other key conserved residues established through biochemical investigations. However, manually validating key residues in thousands of peptide sequences using MSAs is time consuming, especially when considering the large volume of marker gene sequences obtained through amplicon or shotgun metagenome studies [124]. Furthermore, multiple sequence alignment quality degrades as the number of sequences in an alignment increases [236], or when the sequences to be aligned are highly divergent from one another [407].

To address the issue of accuracy in the validation of protein-coding gene sequences, an automated pipeline for protein amino acid signature validation (PASV) was developed. PASV provides researchers with a fast and accurate method for validating protein active sites and point mutations in particular genes of interest. Combining multiple sequence alignment with expert domain knowledge in an automated way, PASV more accurately identifies functional protein sequences within large sequence datasets. In this way, PASV can be used as a post-homology search processing step to eliminate most false positive hits and peptides that are likely to be non-functional. Additionally, PASV can be used to partition proteins into groups based on the residues

present in functionally important positions of an alignment, such as conserved catalytic residues or residues with interesting biochemical properties (e.g., variants in motif B in DNA polymerase I [334]).

The accuracy of PASV was tested using commonly misannotated proteins: ribonucleotide reductase (RNR), alternative oxidase (AOX), and plastid terminal oxidase (PTOX) [214, 258]. In the first case, PASV was used to identify functional RNRs based on active site residues, and to differentiate Class I alpha and Class II RNRs based on a single amino acid residue. In the second case, PASV was used to distinguish two proteins commonly found in plants, AOX and PTOX, which have been previously shown to be difficult to differentiate with homology search alone, but can be readily partitioned using conserved residues [258].

2.2 Methods

2.2.1 PASV pipeline overview

PASV automates the process of aligning query sequences with a set of reference sequences and subsequently validating key residues and regions within the queries (Fig. 2.1). PASV is not a homology search tool. Rather it is a post-homology search filtering program. PASV uses a set of user-defined key amino acid residue positions to review alignment columns within multiple sequence alignments (MSAs). Key positions ideally will be residues that are both essential to the protein's function such as active sites and allosteric binding sites, and highly conserved across the diversity of known protein sequences. In this way, PASV leverages the user's domain knowledge for automated filtering and validation of functional proteins discovered through homology search. Alternatively, key positions may contain residues that, when mutated, display interesting biochemical properties. PASV automatically bins such amino acid variants, providing information on the functional diversity of a given protein. Finally, PASV can automatically filter out query sequences that fail to span a region of interest (ROI) on the reference sequences.

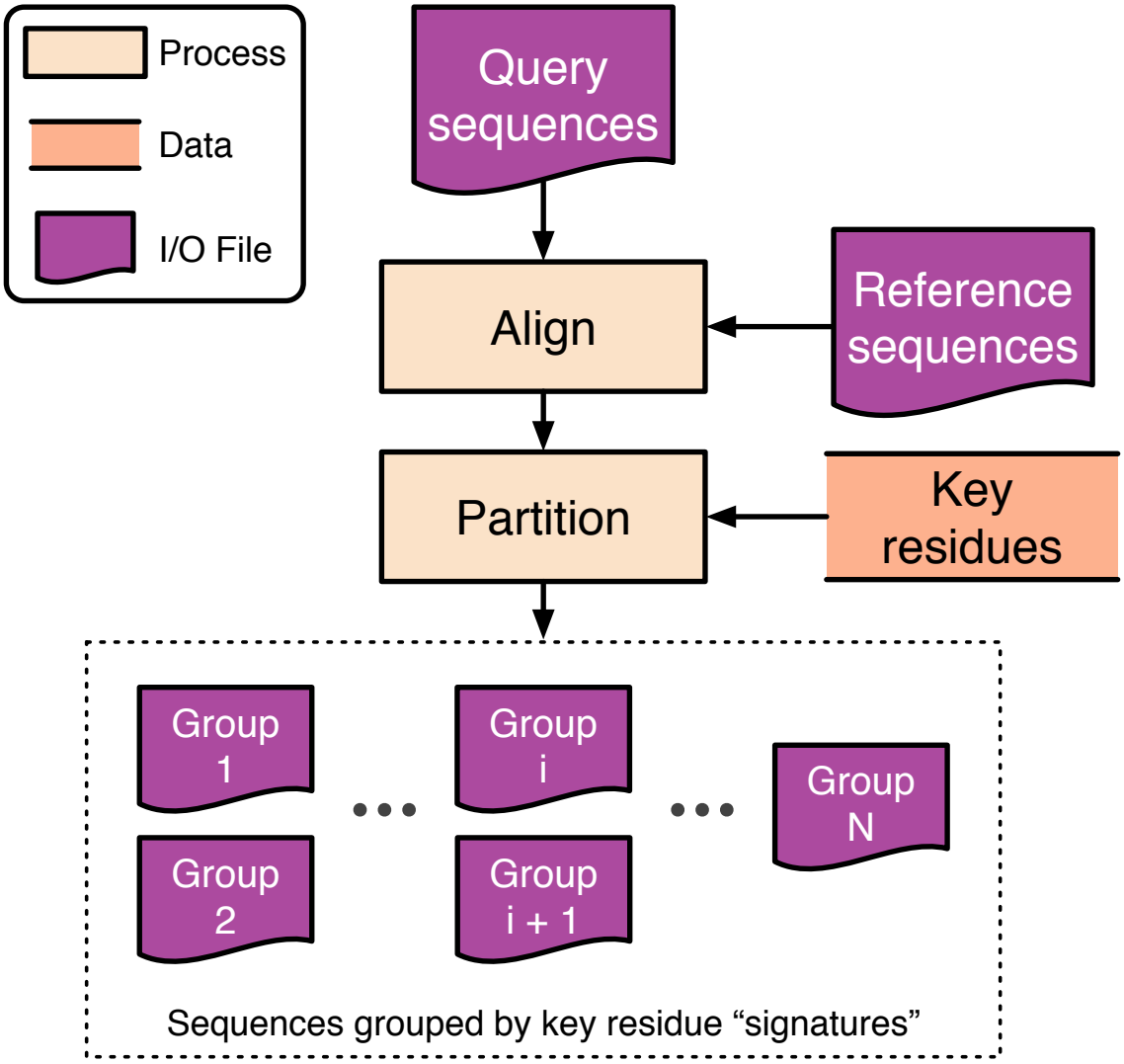

Figure 2.1: PASV conceptual diagram. PASV individually aligns each query sequence with a user-defined set of reference sequences. Then, columns of the resulting multiple sequence alignment are checked for user-defined key residue positions and, optionally, a region of interest (ROI). Finally, query sequences are partitioned into groups based on the amino acids at each of the key residues and whether the sequence spans the ROI.

Prior to using PASV, users must select a set of reference sequences for the alignment. Special care should be taken when choosing a set of reference sequences, as picking an optimal reference set influences PASV's accuracy and runtime (see Results and Discussion sections for best practices). Reference sets are tailored to the protein of interest. That is, a set of references chosen for partitioning ribonucleotide reductase (RNR) sequences would not be the same as a set of references used to partition alternative oxidase (AOX) and plastoquinol terminal oxidase (PTOX). In addition to the reference set, which is developed once for a given protein of interest and then reused, the main input to PASV is a set of query protein sequences, generally obtained via a homology search for a protein of interest within a larger sequence dataset. PASV is especially useful in cases where there are many putative protein sequences to validate. For example, using a highly sensitive homology search tool (e.g., BLAST [9], HMMER [82], MMseqs2 [358], or PSI-BLAST [10]) against a metagenome often returns a large set of putative sequences that would be impractical for manual validation. PASV automates sequence validation avoiding time-consuming and potentially error-prone manual validation.

In the PASV pipeline, each query sequence is individually aligned with the reference sequences. PASV abstracts the process of aligning queries with references and identifying residues present in specific columns. Rather than reimplementing MSA algorithms, PASV leverages existing MSA software for aligning queries and reference sequences. It has built-in support for Clustal Omega [344] and MAFFT [164], but other alignment software can be specified at the command line by providing a custom specification.

For each alignment, PASV checks the residues of the query sequence aligning with the user-provided key residue positions in the reference set. The provided key residue positions are interpreted with respect to the original, unaligned first reference sequence. Each query is assigned a key residue "signature" based on these residues. PASV also optionally checks whether each query sequence spans a user-defined region of interest with respect to the reference sequences. Thus, PASV groups query sequences

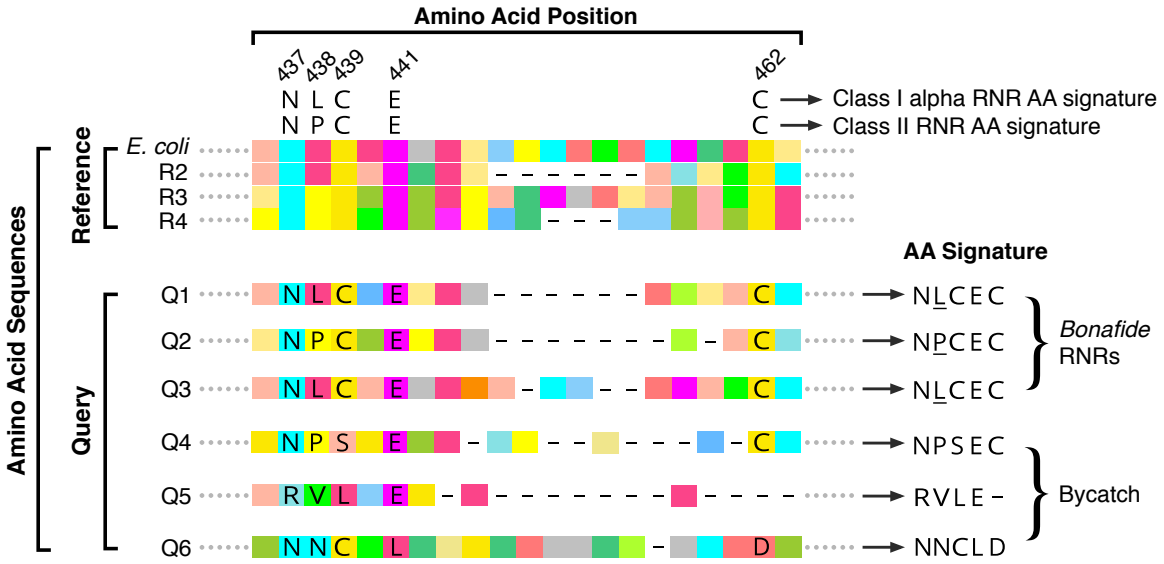

Figure 2.2: RNR classification and partitioning example. PASV aligns each query sequence individually with all reference sequences (in this case, four references). Labelled positions are the user-specified key residues. The coordinates are specified with respect to the original positions on the unaligned first reference sequence (here, *E. coli*). Each query is assigned a signature based on the residues that align in the same columns as the key residues. In the case of RNR, residues N437, C439, E441, and C462 are required, while residue 438 is diagnostic of RNR class (L438 indicates Class I alpha and P438 indicates Class II). In this example, queries 1, 2, and 3 have NCEC in the correct positions and are considered to be bonafide RNRs. Queries 1 and 3 can be classified as Class I alpha based on L438, whereas query 2 can be classified as Class II based on P438. Queries 4, 5, and 6, do not have the required NCEC signature and are thus considered bycatch.

based on the key residue signature, and optionally by ROI spanning status. For example, in the case of RNR, the user may select key residue positions 437, 438, 439, 441, and 462 with respect to the first reference sequence. Then queries will be binned according to the residues that align with the reference sequences at those positions, i.e., their key residue "signatures" (Fig. 2.2).

2.2.1.1 Implementation & source code availability

The PASV pipeline is implemented in OCaml [193]. PASV leverages existing multiple sequence alignment software, such as Clustal Omega [344] or MAFFT [164], thus, a multiple sequence alignment program should be installed prior to running PASV. PASV is open-source software (MIT or Apache license) and is freely available on GitHub (https://github.com/mooreryan/pasv). PASV v1.3.0 (https://github.com/mooreryan/pasv/releases/tag/v1.3.0) was used for all experiments.

2.2.1.2 PASV result network diagrams

Resulting PASV output files were converted to a node-link network diagram with a custom script (available on the PASV GitHub page) and visualized with Cytoscape v3.7.1 [342].

2.2.2 Collecting RNR sequences

2.2.2.1 Retrieving RNR sequences from the RNRdb

All available Class I alpha and Class II RNRs were retrieved from the RNRdb on August 20, 2018 [214]. These 66,209 RNR peptide sequences were dereplicated (exact and substring matches) using CD-HIT v4.6 [106], yielding 29,401 representative sequences. Sequences were then divided into closely related groups (clades) as defined by the RNRdb for manual assessment of active site residues and intein removal [135]. From the 29,401 representative sequences, 286 sequences were removed as they lacked one or more of the four residues essential for RNR function (N437, C439, E441,

C462 with respect to *Escherichia coli* K12 W3110 ribonucleoside diphosphate reductase 1 alpha subunit, accession no. WP_001075164.1) [163, 219, 218, 283]. The 29,133 remaining RNRs were retained for downstream analysis.

2.2.2.2 RNRdb sequence tree & phylogenetic clustering

To reduce the number of sequences used for building a phylogenetic tree of known RNR peptides, the 29,133 bonafide RNRdb sequences were clustered with MMseqs2 (version e1a1c1226ef22ac3d0da8e8f71adb8fd2388a249) [358] at 75% identity over 80% of the alignment length, resulting in a set of 2,579 peptide clusters. Cluster centroids were aligned with MAFFT v7.427 using the FFT-NS-2 method [164]. Columns of the resulting multiple sequence alignment containing >95% gaps were removed. Finally, FastTree v2.1.10 with double precision arithmetic [294] was used to build the tree, and the resulting tree was midpoint-rooted with a custom Python script (https://github.com/mooreryan/midpoint-root) using ETE Toolkit v3 [144]. Different numbers of phylogenetic RNR clusters were generated by collapsing branches whose lengths were below a threshold using iTOL [195]. Six different clustering scenarios were used representing six levels of phylogenetic granularity (4 clusters: collapsed branch length (BRL) < 3.75; 8 clusters: BRL < 3.1; 14 clusters: BRL < 2.85; 19 clusters: BRL < 2.65; 24 clusters: 2.485; and 29 clusters: BRL < 2.34) (Fig. 2.3).

2.2.2.3 Retrieving RNR sequences from the Global Ocean Viromes dataset

The 1,995,784 Global Oceans Virome (GOV) [320] peptides[1] were searched against RNRdb sequences with MMseqs2 (sensitivity: 7, max sequences: 1000, number of iterations: 3, starting sensitivity: 1, sensitivity step: 7, default E-value cutoff: 0.001, defaults for all other options). This search yielded 12,412 virome sequences. Sequences having fewer than 100 amino acids were removed, leaving 9,906 sequences. These sequences were manually curated using a combination of conserved residues, domains,

[1] https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/GOV/Contigs_set, file last modified 2017-04-23

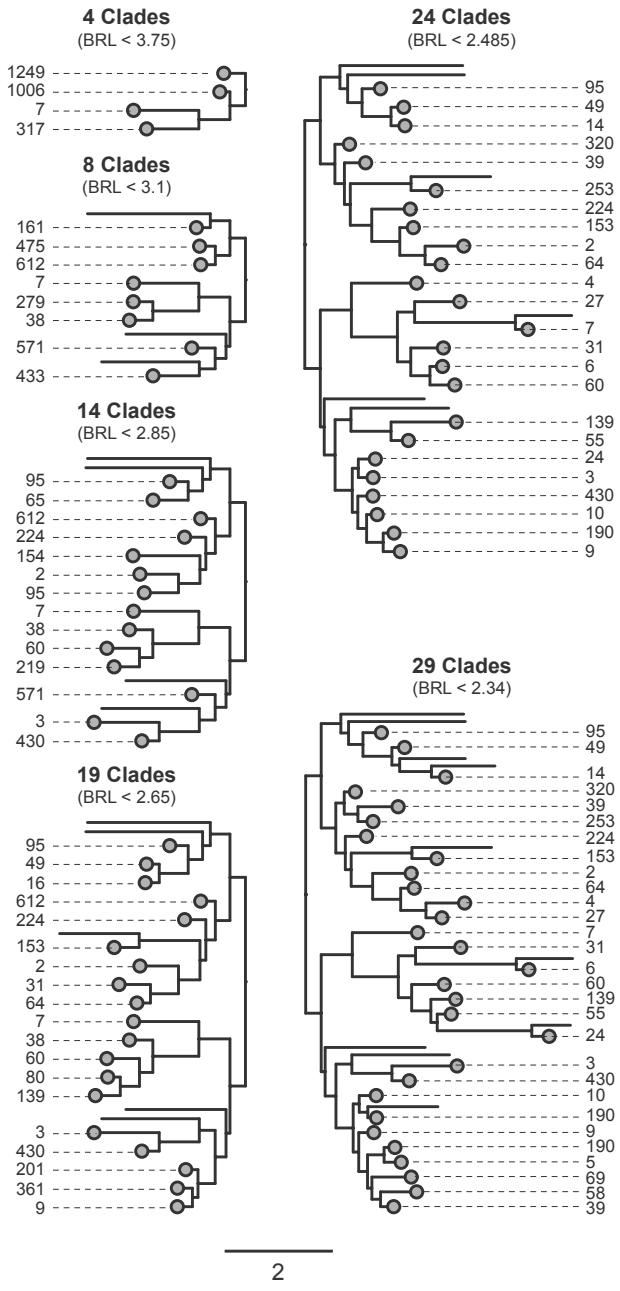

Figure 2.3: Phylogenetic clustering of ribonucleotide reductase proteins. Ribonucleotide reductases (RNRs) from the RNRdb [214] were clustered with MMseqs2 [358] at 75% identity over 80% of the alignment length. Phylogenetic clusters (grey circles) were created in iTOL [195] by collapsing clades with branch lengths (BRL) less than the amount shown. Leaf labels show the number of sequences within the clade. Branches without grey dots represent singleton clusters, and were not included in the pool of potential reference sequences. Scale bar represents amino acid substitutions per site.

and phylogenetic placement (as in [135]) resulting in 2,916 bonafide RNRs and 6,990 non-RNRs.

2.2.3 Reference sets and PASV accuracy

2.2.3.1 Full reference set test

Given that PASV uses MSA for validating key residues, PASV's accuracy is dependent on the chosen reference set and aligner. An experiment testing 1,920 combinations of reference sets, query sets, and aligners was used to determine those variables most affecting accuracy (Fig. 2.4). First, randomly selected reference sequence sets were compared to sets where selection was guided by a phylogenetic tree. For phylogenetically selected references, a tree containing 2,579 RNR sequences was partitioned at six levels of granularity (4, 8, 14, 19, 24, and 29 clusters (Fig. 2.3)). Two approaches were then taken for phylogenetic reference selection. First, phylogenetic reference sets were generated by selecting a single reference sequence from each tree clade (clades defined by various minimum branch lengths (BRL, Fig. 2.3) to test whether increasing the evenness of representation among rarer or divergent clades would improve PASV accuracy. Second, phylogenetic reference sets were generated by weighting the selection of sequences according to the number of sequences within a cluster (one reference sequence for every 200 sequences in the cluster) (Fig. 2.3). For each of the phylogenetically selected reference sets (including weighted and unweighted at all six levels of granularity), size-matched, randomly selected reference sets were included as controls. Finally, for each reference set selection criteria (phylogenetic or random, single or multi, reference set size), ten replicates were generated. Each reference set was tested with two aligners, MAFFT v7.427 [164] and Clustal Omega v1.2.4 [344], and two different query sets (RNRdb queries: 100 bonafide RNRs and 100 invalid RNRs missing key functional residues; Global Ocean Virome (GOV) queries: 200 bonafide RNRs and 100 invalid RNRs missing key functional residues). All experiments were run on an Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.10GHz server with 36 cores (2 threads per core) and 512 GB of ram, with PASV set to use 68 threads (i.e., process 68 queries concurrently).

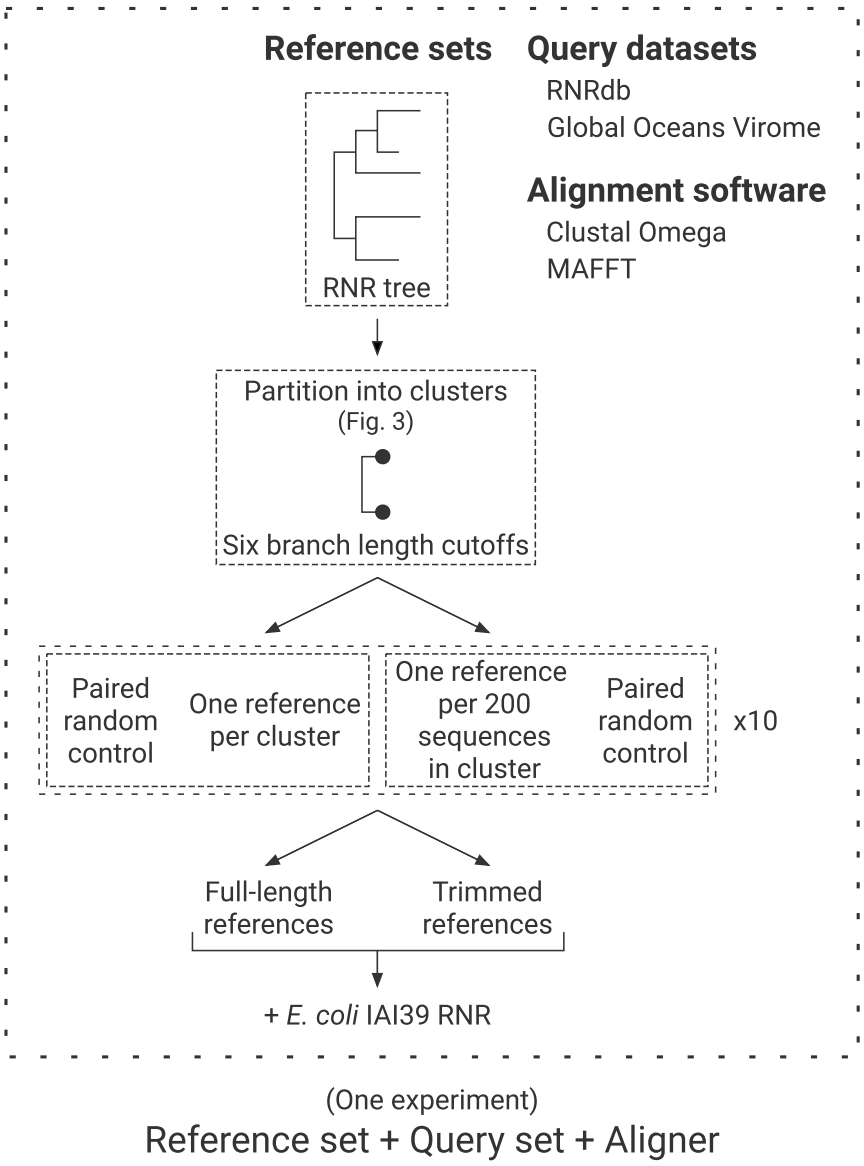

Figure 2.4: PASV reference set test. Conceptual diagram of the validation experiment testing the effects of reference set, query set, and aligner on PASV accuracy. One experiment is a PASV run with a unique combination of a reference set, a query set, and an aligner. The reference sequence selection strategy (phylogenetically-guided or random), the size of the reference set (numbers of sequences and their distribution across the known diversity of a protein), and the length of reference sequences (full length or smaller region of interest) were tested for their impact on PASV accuracy in correctly identifying manually curated sequences. For each reference set category, 10 random samples (i.e., replicates) were generated. For each reference set, two aligners (Clustal Omega [344], and MAFFT [164]), and two query sets (RNRdb [214] and Global Ocean Virome (GOV) [320]) were run.

In summary, a total of 1,920 experiments were conducted. Six levels of phylogenetic tree partitioning were used (6 reference sets) (Fig. 2.3), each generating either a single (unweighted) or multiple (weighted) reference sequences per clade (6 * 2 = 12 reference sets). For each of these twelve groups, ten replicates were made (12 * 10 = 120 reference sets). For each of these 120 reference sets, size-matched, randomly selected reference sets were used as controls (120 * 2 = 240 reference sets). For each of these 240 reference sets, both full-length reference sequences, and reference sequences trimmed to the shorter region of interest (ROI, positions 437 - 605, *E. coli* numbering) were tested (240 * 2 = 480 reference sets). For each of these 480 reference sets, two aligners (Clustal Omega or MAFFT) were tested (480 * 2 = 960 reference sets + aligners). For each of these 960 reference set plus aligner pairs, two different query sets (RNRdb or GOV) were tested (960 * 2 = 1,920 experiments) (Fig. 2.4).

2.2.3.2 Putative GOV RNR queries test

GOV RNR sequences (9,906 sequences) were used to test PASV on a dataset more reflective of an actual use case. Because most of the variables tested in the full reference set test had little effect on PASV accuracy (see Results), and due to the size of the query set, a reduced set of variables was used to generate reference sets. References from three clustering levels (8, 19, 29) with both phylogenetic and random sequence picking were generated in triplicate, yielding 18 reference sets. For the other variables included in the full reference set test, only the top performing options were used in this experiment: Clustal Omega rather than MAFFT, full-length references rather than trimmed, and one sequence per clade vs one sequence for every 200 sequences per clade. All experiments were run on the same server as the full reference set test with PASV set to run 68 concurrent alignment jobs.

2.2.3.3 Data analysis

Data analysis was performed in R v3.6.3 [300] with tidyverse v1.3.0 [390] and ggplot2 v3.3.0 [389]. All true positive and true negative rate linear models were calculated with the `lm` function in R. Model coefficients were considered significant if their p-values were less than 0.05 as reported by the R function `summary.lm`. All box and whisker plots were made using the `geom_boxplot` function from ggplot2. All scatter plot regression lines were made using the `geom_smooth` function from ggplot2 using locally estimated scatterplot smoothing (LOESS, default parameters) with 95% confidence intervals, except for Additional Files 1 and 4 which use linear regression with 95% confidence intervals calculated with `geom_smooth` using `lm`. All point jittering was done using the `geom_jitterdodge` function from ggplot2.

2.2.4 Analyzing putative and bonafide GOV RNRs

2.2.4.1 GOV RNR trees

The 9,906 putative RNR sequences identified through homology search alone, and the 2,914 PASV-predicted bonafide RNR sequences (using the reference set chosen from the best practices according to the full reference set test and the GOV RNR queries test) were aligned with MAFFT v7.427 FFT-NS-2 [164]. Columns with >95% gaps were removed and a phylogenetic tree was inferred with FastTree v2.1.10 double precision arithmetic [294]. The resulting Newick tree files were visualized with Iroki [244].

2.2.4.2 Annotating GOV tree sequences

Sequences were manually selected from clades containing only non-RNRs (according to manual curation) from the phylogenetic tree containing all 9,906 putative RNRs from GOV. Sequences were searched against National Center for Biotechnology Information's Conserved Domain Database (NCBI CDD) v3.18 and the top domain hit by e-value was recorded [221]. All sequences that had a mismatch between manual curation and PASV prediction in any of the 18 full GOV experiments were also

searched against the conserved domain database using Batch CD-Search [222] and the top domain hit was recorded. In the case that multiple domains were identified, the top hit was recorded for each domain (Additional File 10).

2.2.5 Partitioning RNR classes

To test PASV's ability to partition Class I RNR alpha subunit sequences from Class II RNR sequences, the 2,579 clusters from the RNRdb tree (Fig. 2.3) were used as PASV query sequences with the "best practices" RNR reference set. In addition to the same N437, C439, E441, and C462 key residues (*E. coli* numbering) used in previous experiments, residue L/P438 was also included. Any sequence PASV identified as having N**L**CEC was labeled as a Class I alpha RNR, whereas any sequence with N**P**CEC was classified as a Class II RNR. Any sequences with key residue signatures other than the NLCEC for Class I alpha and NPCEC for Class II were grouped into the "Other" category. The PASV predictions were compared with RNRdb assigned class annotations.

2.2.6 Partitioning AOX and PTOX

Alternative oxidase (AOX) and plastid terminal oxidase (PTOX) peptide sequences were collected from a recent study [258]. Sequences from supplemental data sheet 1, containing 14 full-length PTOX proteins that were previously erroneously annotated as AOX, and sequences from supplemental data sheet 2, representing trimmed AOX and PTOX sequences, were obtained. Some of the trimmed sequences in supplemental data sheet 2 had accession numbers with which the corresponding full length sequences could be recovered from NCBI databases using the Entrez Direct efetch [160]. Forty-eight full-length AOX and eight PTOX sequences were recovered in this manner. Recovered full length sequences were combined with trimmed sequences yielding a set of 336 query sequences for PASV testing.

The ability of PASV to classify both AOX and PTOX sequences within a mixed set of peptide sequences was tested with two separate PASV runs: once with an AOX

reference set (UniProt entry IDs O22048, O22049, and E1CIY3; sequences selected from those manually annotated as AOX in [258]) and once with a PTOX reference set (UniProt entry IDs A0A061GHF5, B9RXE2, and Q56X52; sequences selected from those manually annotated as PTOX in [258]) sequences. In the AOX run, all query sequences were checked for conserved residues from AOX motifs 1 (E233, R234, M235, H236, L237, M238, T239) and 2 (L283, E284, E285, E286, A287), and sequences containing the correct residues were labeled as AOX, while sequences with other residues at these positions were labeled as non-AOX (numbering with respect to sequence O22048) [258]. For the PTOX run, all queries were checked for conserved residues from PTOX motifs 1 (G157, W158, R160, R161) and 2 (H177, H178, L179, L180, M182, E183), and any sequences containing the correct residues were labeled as being PTOX, while sequences with other residues at these positions were labeled as non-PTOX (numbering with respect to sequence A0A061GHF5) [258]. Finally, the sequence labels from the AOX and the PTOX run were combined for the final classification. Two positions were excluded from the motifs that were presented in [258] (159 in motif 1, and 181 in motif 2) as these positions were more variable than the other motif positions.

2.3 Results

2.3.1 What factors influence PASV accuracy?

True positive and true negative rates for PASV validated RNR peptide sequences were explored with linear models. For GOV query sequences, aligner and reference trimming had a significant (p-value < 0.05) association with both true positive and true negative rates (Fig. 2.5). Clustal Omega was associated with an 11.1% increase in true positive rate and a 0.2% decrease in true negative rate as compared to MAFFT. Full length references had a 12.6% increase in true positive rate and a 0.07% increase in true negative rate as compared to references trimmed to the region of interest. While statistically significant according to the linear model, variables associated with true negative rate had negligible effect in practice for GOV queries. For RNRdb queries, when all variables were included as predictors, aligner and reference trimming were

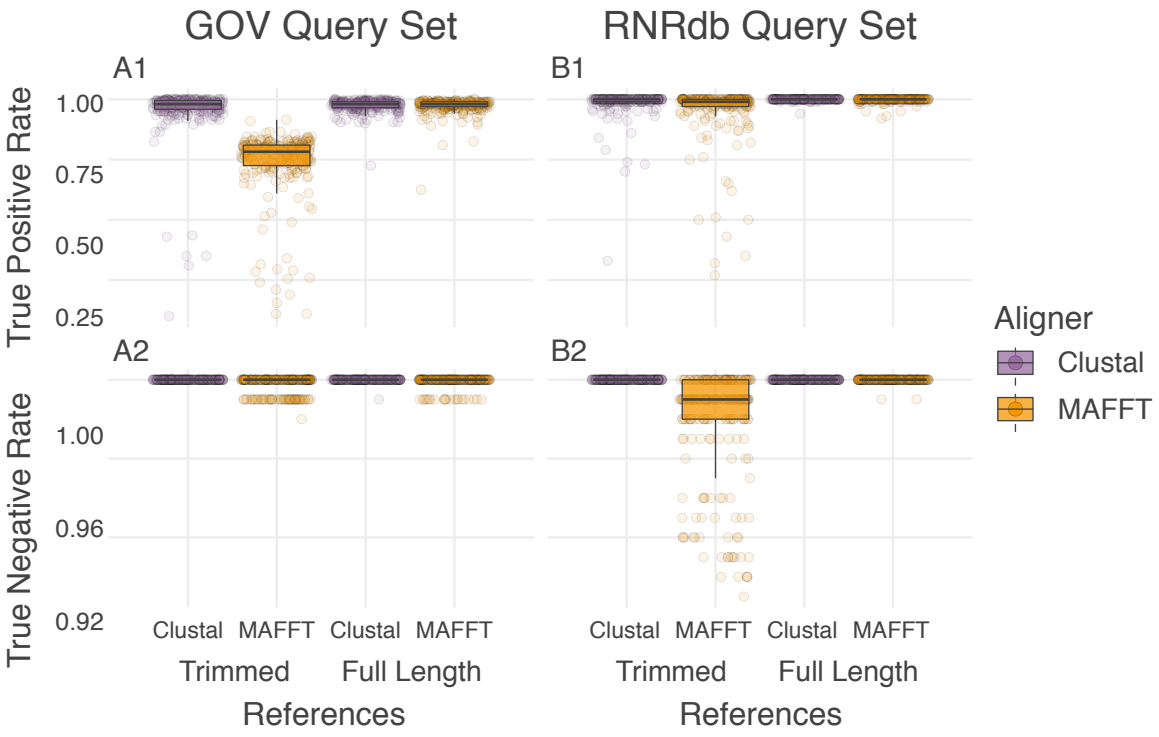

Figure 2.5: PASV accuracy is influenced by aligner and reference trimming. PASV true positive (A1 & B1) and true negative rates (A2 & B2) across reference sets of RNR peptide sequences. Results are shown for the Global Ocean Virome (GOV) query set (A) and the RNRdb query set (B). Each dot represents a single PASV run (i.e., one reference set with an aligner). Box (showing median and interquartile range (IQR)) and whisker (1.5 x IQR) plots are overlaid. Within each panel, PASV tests are partitioned by reference sequence length (full length references vs. those trimmed to the region of interest) and by multiple sequence aligner (Clustal Omega – purple vs. MAFFT – orange).

both significant predictors of true positive rate. Clustal Omega was associated with a 1.5% increase, and full length references were associated with 2.7% increase in true positive rate. For true negative rate, all variables other than replicate were significant; however, all effects were quite small ($< 1.3\%$).

Given that full-length references were superior to those trimmed to a ROI (Fig. 2.5), only full-length references were included in subsequent analysis of covariate effects on PASV accuracy. Full-length reference sets split into groups based on query set (GOV vs RNRdb) and aligner (Clustal Omega vs. MAFFT) were re-run through linear models on the following five remaining covariates: (1) number of tree clusters; (2) number of reference sequences; (3) single or multiple reference sequences chosen per clade (single/multi); (4) random or phylogenetically-guided reference sequence choice (random/phylo); and (5) reference set replicate (Table 2.1).

The PASV true positive rate decreased with the number of tree clusters used in the phylogenetically-guided reference sequence choice approach (Fig. 2.3), but increased with respect to the number of references for GOV-MAFFT, RNRdb-MAFFT, and RNRdb-Clustal groups (Table 2.1). When using MAFFT, picking a single reference from each clade as opposed to weighting the number of references by number of sequences in the clade was associated with a significantly higher true positive rate for both GOV and RNRdb query sets; however, this trend was not seen when using Clustal (Table 2.1). Overall, choosing references randomly (when using MAFFT, but not Clustal) and including more sequences in the reference set were associated with better PASV accuracy. However, the positive effect of the number of reference sequences on true positive rate plateaued after ca. 20 reference sequences (Fig. 2.6). Additionally, the effect of increasing the number of references is more pronounced with the MAFFT aligner than with Clustal Omega (Fig. 2.6).

While using an increasing number of references boosted PASV accuracy, it also increased runtime (Additional File 1), as more sequences needed to be aligned. Using full-length references as opposed to references trimmed to the region of interest also increased the runtime. This is due to full-length references containing more bases that

Table 2.1: Linear model coefficients with p-value < 0.1 for PASV reference set test (full-length references only).

Model	**Variable**	**Coefficient ± Standard Error**			
		GOV-MAFFT	GOV-Clustal	RNRdb-MAFFT	RNRdb-Clustal
True positive rate	Intercept	86.50 ± 1.86	97.60 ± 1.69	97.30 ± 0.55	99.29 ± 0.28
	No. tree clusters[a]	-0.56 ± 0.13		-0.12 ± 0.39	-0.03 ± 0.02
	No. references[b]	0.78 ± 0.15		0.17 ± 0.45	0.05 ± 0.02
	Single vs. multi[c]	4.77 ± 1.33		1.10 ± 0.39	
	Random vs. phylo[d]	0.65 ± 0.36		0.24 ± 0.11	
	Replicate				
True negative rate	Intercept	99.92 ± 0.19	99.99 ± 0.43	99.87 ± 0.06	100.00
	No. tree clusters				0.00
	No. references			0.01	0.00
	Single vs. multi				
	Random vs. phylo	-0.12 ± 0.04			
	Replicate	-0.01 ± 0.01			

[a]Number of tree clusters
[b]Number of reference sequences
[c]Single or multiple reference sequences chosen per clade
[d]Random or phylogentically-guided reference sequence selection

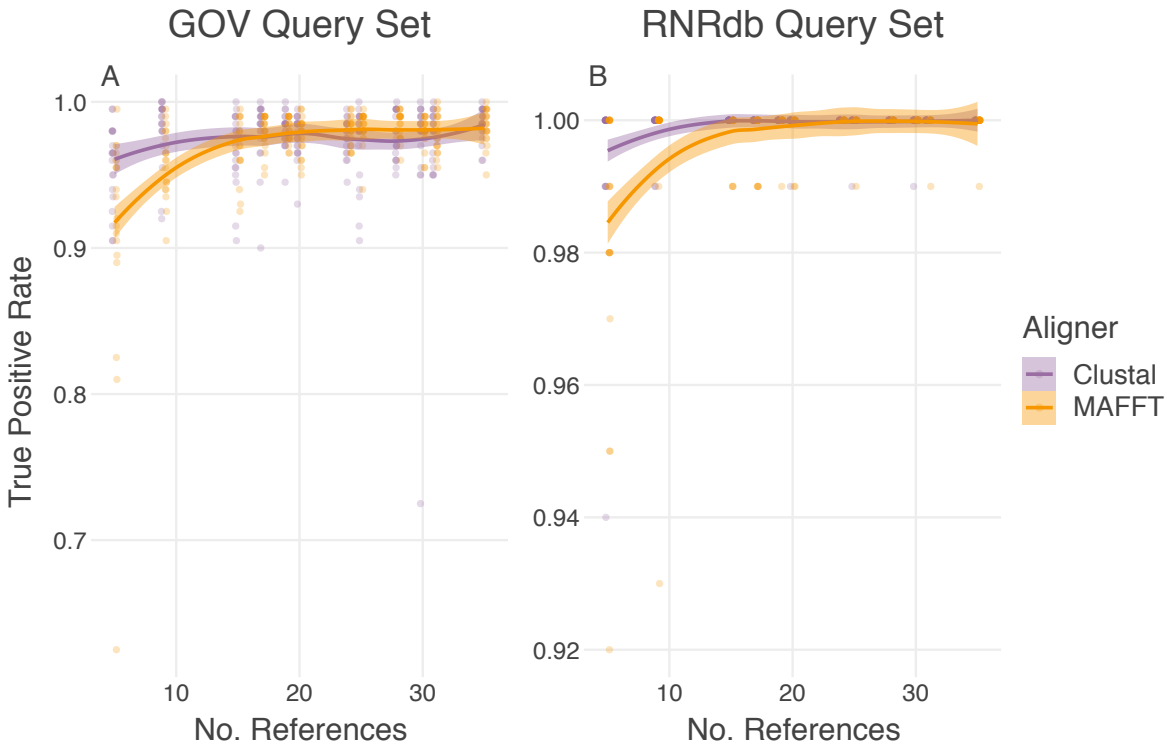

Figure 2.6: PASV true positive rate increases with number of references. Number of references per reference set versus PASV true positive rate for full-length reference sets. GOV query set and RNRdb query set are shown in panel A and panel B, respectively. Each dot represents a single PASV run (i.e., one reference set with an aligner: Clustal Omega – purple, MAFFT – orange). Locally estimated scatterplot smoothing (LOESS) lines with 95% confidence intervals are shown for each aligner. (Note the difference in y-axis scale between panels A and B.)

Table 2.2: Confusion matrix of PASV results for 18 references sets against putative GOV RNR sequences.

PASV Prediction	Manual curation	
	Positive	Negative
Positive	2894.6 ± 5.3	12.5 ± 0.8
Negative	21.4 ± 5.3	6977.5 ± 0.8

Mean $\pm$ 95% confidence interval for 18 PASV runs. Each run is one of 18 reference sets with the full 9,906 sequence Global Ocean Virome (GOV) query set.

need to be aligned. Another consideration for run-time is the alignment algorithm: running PASV with Clustal Omega was faster than with MAFFT (Additional File 1).

In summary, variables that had the most impact on PASV true positive and true negative rate were alignment software (with Clustal Omega outperforming MAFFT) and reference trimming (full-length references performing better than those trimmed to the ROI) (Fig. 2.5, Table 2.1, Additional File 2).

2.3.2 Testing PASV with the full GOV query set

PASV was tested on a large metagenomic query set using best practices determined from the 1,920 reference set tests. The only variable significantly associated with PASV accuracy was phylogenetic vs. random reference picking, which affected the true negative rate; however, the size difference was small (0.027%) (Additional File 3). As the different reference sets all had comparable results, the mean and 95% CI of all 18 reference set runs was used for the confusion matrix. Overall, PASV was highly concordant with the manual curation, with >99% agreement between PASV predictions and manual curation (Table 2.2). As in the full reference set tests, runtime increased with increasing numbers of reference sequences (Additional File 4) (linear model: runtime $= (-13.5 \pm 6.8) + (5.26 \pm 0.3) *$ number of reference sequences).

PASV provides a means for automating the process of validating the identity of peptide sequences collected through homology search. The algorithm partitions

query peptides into bonafide and by-catch sequences (Fig. 2.1). Given this, the impact of including by-catch sequences in a phylogenetic analysis of metagenomic RNR sequences was examined. Phylogenetic trees of putative RNR sequences from GOV (9,906 sequences), and sequences from the putative RNRs that PASV identified as bonafide (i.e., those sequences with N437, C439, E441, C462, *E. coli* numbering) were compared (Additional File 5). For this PASV run, the best performing reference set (hereby referred to as the "best practices" reference set) of the 18 tested on the full GOV query set that also followed the best practices observed in the full reference set test (i.e., full-length, single sequence per clade, random selection) was used. This PASV run yielded 2,914 bonafide RNR sequences (i.e., those sequences with N437, C439, E441, C462, *E. coli* numbering).

The tree including all putative RNRs contained a high proportion of sequences on long branches, indicative of distantly related sequences or sequences with poor alignment (Fig. 2.7A). In contrast, the bonafide PASV sequence tree contained fewer long branches and more reasonable topology [134] (Fig. 2.7B). In the case of both trees, clades with long branches did contain non-target sequences such as helicases, DNA polymerases, terminase, and thioredoxin (Fig. 2.7 and Additional File 6). However, the tree containing bonafide RNR sequences had substantially fewer long branches, and those that were present would be relatively easy to identify and remove. In practice, having fewer long branches reduces the time necessary for manual curation of phylogenetic trees.

Across all 18 GOV PASV runs (1 run per generated reference set), a total of 187 sequences out of 9,906 showed disagreement between PASV predictions and manual curation. These 187 sequences were annotated using NCBI CDD (Table 2.3, Additional File 3). Annotations of the 162 PASV predicted negative, manual curation positive sequences included three Class I RNR alpha subunits, 63 Class II RNRs, and 96 RNRs with unknown subclass. Sequences with hits to the RNR_PFL superfamily were considered to be either Class I alpha or Class II RNRs for two reasons: 1) other members of the supergroup, pyruvate formate lyase (PFL) and Class III RNRs, are

Table 2.3: NCBI CDD annotations of sequences with mismatched PASV prediction and manual curation.

Annotation	Count	
	PASV positive, manual curation negative	PASV negative, manual curation positive
RNR (subclass unknown)	5	96
Class I RNR alpha subunit	*4	3
Class I RNR beta subunit	2	-
Class II RNR	*2	63
Helicase	4	-
Pol I	3	-
Endonuclease	2	-
Terminase	1	-
Ankyrin repeat	1	-
No match	1	-

Counts are totals across 18 PASV runs: the full 9,906 sequence Global Ocean Virome query with 18 different reference sets.

*Sequences erroneously categorized as non-RNR by manual curation

Table 2.4: PASV Class I alpha and Class II predictions.

PASV Prediction	Manual annotation	
	Class I alpha	Class II
Class I alpha[a]	98.96%	0.30%
Class II[b]	0.08%	98.27%
Other[c]	0.96%	1.43%

[a]NCEC sequences with L438 (*E. coli* numbering)
[b]NCEC sequences with P438 (*E. coli* numbering)
[c]Non-NCEC sequences or those with any other residue at position 438

oxygen-sensitive [332, 259], and thus unlikely to be found in the environments sampled in the GOV study [320]; and 2) these sequences grouped with other Class I alpha and Class II sequences on the phylogenetic trees (Fig. 2.7). The 25 remaining mismatched sequences (i.e., PASV predicted positive, manual curation negative) had more heterogeneous annotations. Twelve of these had hits to non-RNR domains: four helicases, three Pol Is, two endonucleases, one terminase, one Ankyrin repeat, and one with no match. Thirteen had hits to RNR domains: four Class I alpha subunit, two Class I beta subunit, two Class II RNR, and five RNRs with unknown subclass. The six sequences annotated as RNR Class I alpha subunits and Class II represent sequences that were likely erroneously categorized during manual curation. To summarize, 187 sequences of 9,906 had mismatched PASV predictions and manual annotations–of those 187, PASV was likely incorrect according to CDD annotations in 175 of the cases (yielding 98.2% accuracy and 1.8% error rate).

2.3.3 Partitioning RNR Class I alpha subunit & Class II sequences

PASV's ability to partition two biochemical classes of RNR sequences (Class I alpha subunit and Class II [305, 259]) was examined. The 2,579 RNRdb sequences used to make the RNR tree for phylogenetic clustering (Fig. 2.3) were partitioned into Class I alpha subunits and Class II sequences using PASV. As the NCEC residues within the RNR PASV profile are required for RNR function [163, 219, 218, 283], any

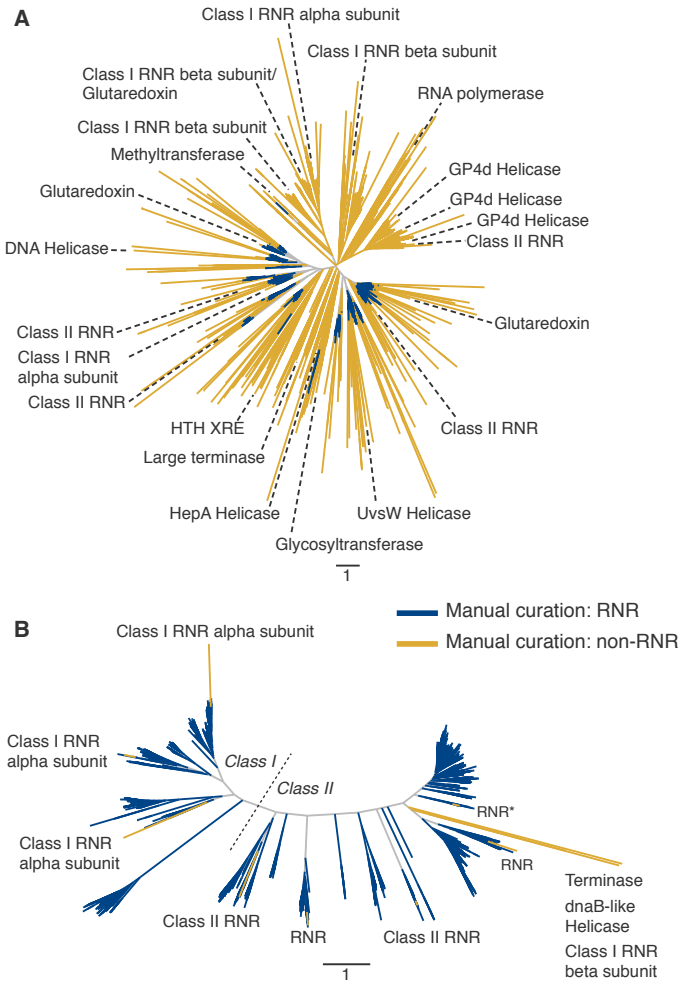

Figure 2.7: Phylogenetic trees of putative and bonafide GOV RNR sequences. Approximately-maximum likelihood trees of (A) 9,906 putative GOV RNR sequences identified by MMseqs2 using sensitive homology search parameters, and (B) 2,914 PASV validated, bonafide GOV RNR sequences (i.e., sequences with N437, C439, E441, C462, *E. coli* numbering). In panel B, the dotted line indicates the divide of Class I and Class II RNR sequences. Branch colors correspond to the results of manual curation. Blue branches indicate sequences manually annotated as RNR, whereas yellow branches represent sequences annotated as non-RNR or non-functional RNR sequences. Labelled sequences represent a sampling of sequences with homology to RNR, but manually curated as non-RNR or nonfunctional RNR. Note that some yellow branches in panel B, which were originally annotated as RNRs through manual curation, but having the correct residues according to PASV, were found to have correct RNR annotations according to the NCBI CDD [221]. The branch labeled "RNR*" in panel B indicates 3 branches annotated as RNR by the CDD.

sequence without NCEC at residues 437, 439, 441, and 462 (*E. coli* numbering) in the PASV run were grouped into the "Other" category. These included five Class I alpha and four Class II sequences. For the remaining 2,570 NCEC sequences, any sequence that PASV predicted as having an leucine at position 438 was labeled as a Class I alpha subunit, whereas any sequence with a proline at that position was predicted to be Class II. These PASV predictions were compared to RNRdb annotations, and the results were recorded in Table 2.4 (Additional File 7). Of the 1,244 annotated Class I alpha sequences, PASV predicted 1,236 of them to be Class I alpha (correct PASV predictions: 98.96%), one to be Class II (0.08%), and seven to be "Other" (0.96%). For the 1,326 annotated Class II sequences, PASV predicted 1,307 of them to be Class II (correct PASV predictions: 98.27%), four to be Class I alphas (0.30%), and 15 "Others" (1.43%).

2.3.4 Partitioning AOX and PTOX sequences

A total of 336 alternative oxidase (AOX) and plastid terminal oxidase (PTOX) peptide sequences were recovered from a previous study examining misannotation of the AOX and PTOX gene groups in plants [258]. These sequences were classified with PASV using residues from the diagnostic, conserved motifs identified in [258]. This experiment tested the ability of PASV for correctly binning a mixed collection of AOX and PTOX peptide sequences. While distinct proteins, AOX and PTOX share regions of homology and are frequently missanotated by standard methods [258]. However, two motifs for each protein, when used in conjunction with MSA, enables correct classification of the proteins. Two reference sets were constructed, one to classify AOX and one to classify PTOX. The entire query set (336 total sequences, 254 AOX, 82 PTOX) was run through the PASV algorithm against both reference sets (Additional Files 8 & 9). In the AOX run, any sequence with the correct residues in the conserved motifs as identified by PASV was considered an AOX (motif 1: E233, R234, M235, H236, L237, M238, T239; motif 2: L283, E284, E285, E286, A287, numbered according to sequence O22048). Any sequence containing any other residue in any of these positions

was considered to be non-AOX. In the PTOX run, sequences that PASV identified as having the correct residues in motifs 1 and 2 were annotated as PTOX (motif 1: G157, W158, R160, R161; motif 2: H177, H178, L179, L180, M182, E183, numbered according to sequence A0A061GHF5). Sequences containing different residues in any of these positions were annotated as non-PTOX. When these two annotations were combined, PASV correctly identified all 254 AOX and 82 PTOX peptide sequences and misannotated none.

2.4 Discussion

Homology tools used for collecting gene sequences from databases and metagenomes, such as BLAST [9], HMMER [82], MMseqs2 [358], or PSI-BLAST [10], are sensitive and have the ability to detect remote homology between sequences. While detecting distant homologs is useful, especially when analyzing environmental metagenomic data, such sensitivity often comes with a price: increased levels of false positive sequences [166]. In the context of viral and microbial ecology, false positives can include non-functional versions of the protein of interest, correctly annotated proteins that do not span a predetermined region of interest, and proteins that share a conserved region or domain with the protein of interest, but are not the desired protein.

Including such false positives in analyses of functional proteins causes a number of problems. False positives interrupt multiple sequence alignments and subsequent phylogenetic analyses, which leads to inaccurate conclusions as to the evolutionary history of a protein [263, 398]. In ecological studies, inclusion of false positive sequences in marker gene phylogenetic analyses can lead to erroneous identification of microbial or viral populations [378, 258, 135].

Manual validation of proteins becomes increasingly error-prone and impractical with increasing dataset size. While larger datasets provide the means for deeper exploration of microbial communities and protein diversity and evolution, they also yield more protein sequences for validation. Sensitive homology searches can result in

thousands of protein sequences from a single metagenome library, making automatic validation an attractive option.

2.4.1 Using RNRs to test PASV

Any protein containing conserved residues, whether these are discovered purely through computational methods or are backed by biochemical characterization experiments can be validated using PASV. Ribonucleotide reductase (RNR), an ancient enzyme with well understood structural biochemical features [156] that is often misannotated in sequence databases [214], was an excellent experimental model for testing PASV's ability to validate and partition putative RNR sequences collected from large sequence datasets by homology search. RNRs contain many immutable residues that have been discovered through decades of structural biology research [173]. There is at least one documented case of a gene with high sequence homology to RNR with mutated active sites that has evolved to perform an alternative function [187].

While RNRs are evolutionarily related, perform the same function, and are biochemically conserved, some share only 10-20% primary sequence similarity, a level below the "twilight zone" of homology search similarity [318, 374, 213]. Searching for RNRs, therefore, requires sensitive homology searches, which can return many false positive sequences. Due to the low level of sequence similarity among RNRs in general, and its many classes and subclasses, RNRs can be difficult to annotate. In one survey of RNRs recovered from GenBank, only 23% were deemed to be annotated correctly and 16% had not been annotated as RNRs at all [214]. Given the frequency of misannotation, low sequence homology, presence of immutable residues, and the RNRdb, a large, hand-curated database of bonafide RNR sequences [214], RNR provided an excellent model system for testing PASV. In addition, RNRs are of interest to researchers in many fields, including evolution, biochemistry, cancer research, and viral ecology [259, 325].

Here, the focus was on Class I and II RNRs, which are the two most closely

related extant RNRs. Class I RNRs are encoded by two genes, one each for the alpha and beta subunits comprising the active protein [156]. The larger alpha subunit is hypothesized to be the direct descendent of Class II RNRs [213], while the beta subunit belongs to the ferritin-like superfamily [13] and bears no homology to either Class I alpha or Class II RNRs. Class I and II RNRs require different cofactors for ribonucleotide reduction, so differentiating the classes is crucial for subsequent ecological analyses [325, 135].

PASV was tested using RNRs from two contrasting datasets: the RNRdb [214] and Global Ocean Viromes (GOV) [320]. The majority of RNRs in the RNRdb are from known organisms within large sequence databases (e.g. GenBank, SwissProt, etc.), with relatively few sequences originating from metagenomes. Virus sequences are relatively rare in curated databases as compared to sequences from eukaryotes and bacteria. In fact, viral sequences make up only 2.7% of the Class I alpha and Class II RNRs in the RNRdb. GOV, in contrast, is an environmental dataset of viral sequences. Thus the RNRdb and GOV represented different challenges for PASV.

2.4.2 Factors influencing PASV accuracy

The most important factors influencing PASV accuracy surrounded the relative length of reference sequences and the approach used for choosing them. Using full length reference sequences, picking references randomly from a pool of potential sequences rather than based on phylogenies, and using more reference sequences all increased accuracy as measured by true positive and true negative rate. The benefit of using more reference sequences, however, plateaued after ca. 20 sequences in the reference set (Fig. 2.6), while the computing time required by PASV continued to increase (Additional File 1).

For each phylogenetically-informed reference set generated, a size-matched set of randomly selected RNRs were chosen to act as a control. It is important to note that while the randomly selected sequences are random with respect to their position on the tree, sequences from the RNRdb are biased with respect to class and subclass

representation. Therefore, the "random" controls can also be seen as weighted by the composition of the RNRdb.

Alignment software was also a factor, with Clustal Omega generally outperforming MAFFT. However, this advantage was mostly lost when using full-length reference sequences rather than references trimmed to the region of interest. This result may also differ depending on the protein to be aligned, as some datasets are more difficult to align than others [386].

Reference sets representing as much of the known diversity of RNRs as possible (i.e., those taken evenly from across major clades of a phylogenetic tree) were hypothesized to increase PASV accuracy. This hypothesis was built on the idea that including diverse RNRs would prevent large irregularities in the alignments from more divergent query sequences. However, including diverse RNRs had the opposite effect and statistical tests showed that randomly selecting full-length reference sequences resulted in greater accuracy. One explanation for this phenomenon is that accuracy of multiple sequence alignment decreases with increasing sequence heterogeneity [407, 236]. As a consequence, forcing divergent sequences into the reference sets likely destabilized the alignments and decreased PASV's accuracy.

2.4.3 Using PASV to eliminate bycatch of non-target sequences

The GOV dataset provided an alternative experimental model for testing how PASV performed as a post-processing step after a homology search of a metagenomic sequence library. PASV effectively filtered out false-positive bycatch sequences recovered from the environmental metagenomes while searching for the gene of interest, RNR. Of the nearly 10,000 putative RNR sequences identified by MMseqs2, only about one-third were validated as functional RNRs by both PASV and manual curation. The other two-thirds were considered bycatch sequences. Common gene families within the bycatch sequences included RNR Class I beta subunits, thioredoxins, glutaredoxins, polymerases, helicases, and terminases (Additional File 6). Given the sensitivity of MMseqs2 [358], it is likely to find significant hits in sequences only distantly related to

RNR or to sequences with domains similar to those occasionally found in RNRs. Some RNR Class I beta subunits are known to contain fused glutaredoxin domains [323]. RNRs may also have regions of remote homology to polymerases, helicases, and terminases as all of these proteins bind DNA. Some RNRs are known to contain zinc-finger domains [205], and at least one of the helicases examined with the CDD contained a zinc-finger domain as well (Additional File 6).

Overall, PASV did an excellent job of removing most bycatch sequences (Table 2.2). Across the 18 reference set experiments that used the full GOV query set, only 187 of 9,906 RNR sequences had PASV predictions that disagreed with manual curation (Additional File 10). In most instances these sequences, annotated as terminases, polymerases, and helicases by NCBI CDD, existed on long branches indicating significant evolutionary distance from true Class I large subunit and Class II RNR sequences (Fig. 2.7B). Many of the false-positives identified by PASV (those sequences that PASV predicted to be RNRs, but manual curation predicted to be non-RNR) were likely RNR sequences that were missed during manual annotation. This can be attributed to the challenge of manually curating thousands of sequences and the problems inherent when performing large multiple sequence alignments.

2.4.4 Partitioning sequences by key residues

PASV was conceived as a tool for validating the identity and functionality of protein sequences following homology searches. However, use cases for PASV extend beyond separation of bonafide and bycatch sequences. PASV provides an automated method for applying domain knowledge of a target protein to a large number of sequences. From this domain knowledge, PASV can partition sequences into groups based on structural characteristics that may be linked with protein biochemistry or phylogeny.

PASV was used in such a way to partition Class I alpha and Class II RNRs. While many amino acid residues in the active and allosteric sites of Class I alpha and Class II RNRs are conserved, other residues may be diagnostic of class [135]. Prior

work based on protein alignments and phylogenetic trees suggests that the residue in position 438 (*E. coli* numbering) may be diagnostic of RNR class. Thus, PASV's ability to leverage this domain knowledge was tested by sorting RNRdb sequences into class based on the identity of the residue in position 438. The function of residue 438 is unknown, but it is known to be conserved and sits within the active finger loop domain that contains the immutable active sites N437, C439, and E441 [84]. The sorting by PASV agreed almost perfectly with the RNRdb class annotations (Table 2.4), with >98% of Class I alpha and Class II sequences correctly identified.

An extension of this use case are peptides that cannot be differentiated by homology searches alone. Alternative oxidase (AOX) and plastid terminal oxidase (PTOX) are membrane-bound di-iron carboxylate proteins that oxidize a quinol substrate [32]. Although the proteins function within different organelles (AOX functions within the mitochondrial electron transport chain [3, 230] while PTOX is a chlororespiration enzyme only found within plastids and cyanobacteria [55]), their shared homology and function has led to high levels of misannotation [258]. However, using the amino acid signatures presented previously [258], PASV was able to sort AOX and PTOX proteins from each other with 100% accuracy. In this way, PASV leverages expert knowledge in an automated fashion.

It has been shown that PASV can accurately partition Class I alpha and Class II RNRs using a residue diagnostic of these classes (Table 2.4), and AOX sequences from PTOX sequences using conserved motifs [258]. Given its success with these two disparate examples, it is likely that PASV could be effectively applied to other gene partitioning tasks as well. For example, a single amino acid mutation at position 762 (*E. coli* numbering) of motif B of DNA polymerase I (Pol I) imparts dramatic changes in either the fidelity or efficiency of replication [368]. Subsequent work has hypothesized that Pol I 762 mutations predict the life history characteristics [334] and the genetic composition of the replication module [251] of bacteriophages using Pol I for genome replication. PASV could be used to automatically partition viral Pol I sequences based on the 762 position, providing a means to further test hypothesized connections between

Pol I biochemistry and phage life history using large metagenomic datasets. There are many examples of point mutation(s) in bacterial proteins that prevent antibiotics from binding and, thus, inhibit the function of the antibiotic (e.g., K88R in *rpsL* [23], C117D in *murA* [72], H526T in *rpoB* [324], Q124K in EF-Tu [420], V246A and V300G in *ndh* [381]). Such point mutations within a protein would not be readily apparent from homology search alone. Thus PASV could be used for validating and grouping these peptide sequences according to key point mutations following identification via homology search.

2.5 Conclusions

Studies using gene sequences of functional proteins collected from metagenomes for investigating microbial diversity provide new challenges not faced when using genes for stable RNAs like SSU rRNA. These challenges include detecting and preventing false-positive bycatch sequences within datasets, validating key functional residues in proteins of interest, and partitioning peptide sequences into groups or classes. The PASV pipeline provides researchers with a means for addressing these challenges in an automated and highly accurate fashion by combining multiple sequence alignment with expert-curated domain knowledge. The PASV program and source code is freely available under the MIT license and can be found, along with documentation and usage examples, on GitHub: https://github.com/mooreryan/pasv.

2.6 Additional information and declarations

2.6.1 Availability of data and materials

PASV source code and documentation are available on GitHub at https://github.com/mooreryan/pasv. The PASV Docker image is available on DockerHub at https://hub.docker.com/r/mooreryan/pasv. Data sets and miscellaneous scripts used in the preparation of the manuscript are available on Zenodo at https://doi.org/10.5281/zenodo.4426410. Additionally, a snapshot of the PASV source code v1.3.0 is available on Zenodo at https://doi.org/10.5281/zenodo.4426410.

2.6.2 List of abbreviations

- AOX: alternative oxidase
- BRL: branch length
- CDD: conserved domain database
- GOV: global ocean virome
- IQR: interquartile range
- LOESS: locally estimated scatterplot smoothing
- MSA: multiple sequence alignment
- NCBI: National Center for Biotechnology Information
- PASV: protein amino acid signature validator
- PFL: pyruvate formate lyase
- Pol I: DNA polymerase I
- PTOX: plastid terminal oxidase
- RNR: ribonucleotide reductase
- ROI: region of interest

2.6.3 Funding

This project was supported by the Agriculture and Food Research Initiative grant no. 2012-68003-30155 from the USDA National Institute of Food and Agriculture, the National Science Foundation Advances in Biological Informatics program (award number DBI-1356374), the National Science Foundation Grant No. 1736030, the Established Program to Stimulate Competitive Research (award number OIA-1736030) from the Office of Integrated Activities, and a Doctoral Fellowship provided by University of Delaware in conjunction with the Unidel Foundation. Computational infrastructure support by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility was made possible through funding from the Delaware Biotechnology Institute, and the Delaware INBRE program with a grant from the National Institute of General Medical Sciences (NIGMS P20 GM103446)

from the National Institutes of Health and the State of Delaware. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

2.7 Acknowledgements

The authors would like to acknowledge current and past members of the Viral Ecology and Informatics Lab at the University of Delaware who tested early versions of PASV and provided valuable feedback: Jacob T Dums (ORCID: 0000-0002-6314-4779), Zach Schreiber (ORCID: 0000-0002-6271-2754), and Michael Dahle (ORCID: 0000-0003-0518-3355).

Chapter 3

INTEINFINDER: AUTOMATED INTEIN DETECTION FROM LARGE PROTEIN DATASETS

3.1 Introduction

Inteins (intervening proteins) are intervening polypeptides found within the coding regions of their host genes and are the protein equivalent of introns [328]. They are transcribed and translated along with host protein fragments (exteins) before autocatalytically splicing out from the precursor protein [189, 277]. Inteins contain multiple domains. Two of these are required and involved in protein splicing (N-terminal and C-terminal domains) [289], while the third is an optional homing endonuclease domain that disrupts the two protein splicing domains and enables intein mobility [26]. Inteins that lack the endonuclease domain are called mini-inteins [328]. Some mini-inteins are split and are transcribed and translated as two separate polypeptides that spontaneously assemble and then ligate their exteins *in trans* [341]. Inteins are distributed widely, if sporadically, across taxa. They are especially common among microbes including archaea, bacteria, single-celled eukaryotes, and viruses, with approximately half of archeal genomes and a quarter of bacterial genomes in NCBI containing inteins [167, 261].

Because of their unique properties, inteins have broad applicability in biotechnology and genetic engineering and are highly studied in those fields. Of particular interest is their ability to regulate extein function at the post-translational level in conditional protein splicing (CPS) systems [246]. In this context, split inteins have been used for controlling transgene expression in *Caenorhabditis elegans* [384], generating

bi-specific antibodies (those with two different antigen-binding arms) [128], and developing tumor-targeting protein delivery systems for protein toxic drugs [58], to name only a few applications.

Inteins were traditionally thought to be parasitic genetic elements providing no benefit to the host organism [117, 250], but recent studies have challenged this idea. Inteins may provide selective advantages to their host organism through post-translational regulation and CPS. It has been posited that some inteins have evolved to function as a "pause button" that controls host protein function: intein presence in the intein-extein precursor pauses the function of the host protein, with conditional intein splicing providing a means of rapid protein activation [189]. Such naturally-occurring CPS has been reported in a variety of systems, controlled by mechanisms including the presence of the extein's substrate [191], the addition of a reducing agent [48, 47], reactive oxygen and nitrogen species [372], and temperature [190, 373]. In some cases, these mechanisms of CPS control are connected to conditions favorable to the host organism, effectively acting as a sort of environmental sensor, preventing splicing until optimal growth conditions for the host organism occur (e.g., [373]).

Lending support to this idea, intein distribution is biased towards specific types of proteins, suggesting selective retention of inteins [262, 261]. Inteins are most commonly found in proteins related to replication, recombination, and repair (e.g., polymerases, helicases, and ribonucleotide reductases) [262], and in viral-specific genes (e.g., the large terminase subunit, which translocates DNA into empty capsids) [167].

Intein (and intron)-encoded endonucleases have been shown to provide competitive advantage to their viral hosts. These endonucleases serve as a mechanism for exclusion processes during competitive infection by multiple viruses [118, 28, 413]. For example, if viruses involved in a mixed-infection differ by the presence of an intein-encoded endonuclease within the same gene, the intein-free copy can be cleaved by the endonuclease resulting in a reduction of fitness of the intein-free virus [75]. Work with giant viruses has demonstrated that intein (and intron)-encoded endonucleases can be anti-viral weapons, rather than simply being genomic parasites [98]. Thus, there is

growing evidence that inteins are not merely selfish genetic elements, but provide a potential benefit to their host.

Accurate intein identification is critical for studying the biology, ecology, and evolution of these fascinating mobile genetic elements and their host proteins. Inteins can be useful for exploring non-vertical evolutionary relationships between the host organisms in which they reside. For example, the distribution of closely related inteins in distantly related organisms, or vice versa, may indicate gene flow and the occurrence of horizontal gene transfer events [352, 353]. The presence of inteins with endonucleases especially may indicate frequent genetic exchange in the types of genes that carry them [117, 80]. Additionally, there is no clear picture of the distribution of inteins themselves across taxa, despite recent studies in this area [167, 261].

Intein identification is also important for investigations where intein presence can confound phylogenetic and ecological analyses. Ecological studies of viruses and cellular microbes often utilize peptide coding genes as a means for assessing diversity and population dynamics [2, 395]. In particular, ecological and phylogenetic studies of viruses often use genes that commonly contain inteins, such as DNA polymerase [143, 334, 388, 251, 168], ribonucleotide reductase [80, 325, 135, 401], and terminase [167, 209]. The presence of inteins in these viral genes confounds their phylogenetic analysis as the intein sequences themselves provide little phylogenetic information for their extein sequence [117]; because they are mobile, most inteins generally do not cluster based on the phylogenetic or taxonomic classification of their host organism or virus [281]. Thus, intein sequences should be identified and removed prior to evolutionary or population-scale analyses [68, 243, 325, 135].

Studies that have identified inteins within a set of genomes or other large peptide datasets have followed a similar general workflow [352, 261, 167, 121]: creating a reference intein database (e.g., using intein sequences from InBase [282] or conserved

domains); identifying intein-containing sequences by homology search against the reference intein database; and validating putative inteins through conserved residues or annotation using methods such as CD-Search [221] or InterPro [101]. To date, intein discovery and validation workflows have not been consolidated into a single pipeline that automates this multi-step bioinformatic process. InteinFinder provides a standardized and automated pipeline for identifying, cataloging, and removing inteins from peptide sequences. The pipeline can handle large datasets with millions of peptide sequences and can be incorporated into existing workflows focused on phylogenetic analysis of marker gene sequence data.

3.2 Methods

3.2.1 Building search databases

InteinFinder uses two databases in the initial search stage of the pipeline. One consists of full length intein sequences (intein sequence database – ISDB), and the other includes conserved domain models associated with inteins (intein conserved domain database – ICDDB).

3.2.1.1 Intein sequence database (ISDB)

InteinFinder's intein sequence database (ISDB) contains experimentally and computationally predicted intein sequences. All intein sequences from InBase [282] (InBase last updated Nov. 5, 2010) were downloaded. Resulting sequences were checked manually and errors corrected (e.g., FASTA header lines included inside the sequence definition, or errors in sequence IDs). A subset of intein sequences in InBase included up or downstream extein residues (e.g., -1 and +1 extein residues). For consistency, any extein residues included in the intein sequences were removed. Inteins identified in other studies [167, 121] were also collected and subjected to similar manual post-processing. Finally, all intein sequences were combined into a single, non-redundant database by clustering all sequences using CD-HIT version 4.6 [106] at 100% global

sequence identity over 100% coverage of the smaller sequence to remove exact and substring matches.

3.2.1.2 Intein conserved domain database (ICDDB)

Conserved domain models (CDs) were selected from NCBI's Conserved Domain Database (CDD) FTP (accessed March 27, 2017) based on their membership in the following superfamilies often to be associated with inteins: Hint superfamily (Hint: cl22434), HNHc Superfamily (HNHc: cl00083), Intein splicing (cl25944), and LAGLIDADG WhiA (cl08299). A full listing of included CDs can be found in Table 3.1.

3.2.2 InteinFinder pipeline

The InteinFinder pipeline consists of three parts (Fig. 3.1): (1) homology search against known intein sequences and intein-associated conserved domain models, (2) validation of conserved splice junction residues [281], and (3) refinement of predicted intein region(s). Each step of the pipeline yields additional evidence for the presence of inteins in query sequences. Through multiple rounds of validation, InteinFinder partitions query sequences into different confidence tiers based on the evidence for an intein sequence.

InteinFinder is implemented in OCaml [193] with source code and precompiled binaries for MacOS and Linux available on GitHub[1] under the MIT or Apache license. It uses the following third-party software for homology searches and sequence alignments: MMseqs2 [358], MAFFT [164], and RPS-BLAST [221].

3.2.2.1 Defining putative intein regions

InteinFinder identifies query peptides containing putative intein sequences using two separate homology searches. Query peptides are searched against ISDB using MMseqs2 [358], and searched against ICDDB using RPS-BLAST [221]. Queries that

[1] https://github.com/mooreryan/InteinFinder

Table 3.1: Superfamilies and conserved domain models included in the InteinFinder Intein Conserved Domain Database (ICCDB).

Superfamily	**Superfamily ID**	**CD**	**CD ID**
Hint	cl22434	Hint	cd00081
Hint	cl22434	Hint_2	pfam13403
Hint	cl22434	HintN	smart00306
Hint	cl22434	intein_Nterm	TIGR01445
Hint	cl22434	Vint	pfam14623
HNHc	cl00083	Colicin-DNase	pfam12639
HNHc	cl00083	Csn1	cd09643
HNHc	cl00083	DUF1524	pfam07510
HNHc	cl00083	EndA	COG2356
HNHc	cl00083	Endonuclease_1	pfam04231
HNHc	cl00083	HNH	pfam01844
HNHc	cl00083	HNH_2	pfam13391
HNHc	cl00083	HNH_3	pfam13392
HNHc	cl00083	HNH_4	pfam13395
HNHc	cl00083	HNHc	cd00085
HNHc	cl00083	HNHc	smart00507
HNHc	cl00083	McrA	COG1403
HNHc	cl00083	PRK11295	PRK11295
HNHc	cl00083	PRK15137	PRK15137
HNHc	cl00083	TIGR02646	TIGR02646
HNHc	cl00083	WHH	pfam14414
HNHc	cl00083	zf-His_Me_endon	pfam05551
Intein_splicing	cl25944	HintC	smart00305
Intein_splicing	cl25944	Hop	COG1372
Intein_splicing	cl25944	intein_Cterm	TIGR01443
Intein_splicing	cl25944	Intein_splicing	pfam14890
LAGLIDADG_WhiA	cl08299	Hom_end	pfam05204
LAGLIDADG_WhiA	cl08299	LAGLIDADG_3	pfam14528
LAGLIDADG_WhiA	cl08299	LAGLIDADG_WhiA	pfam14527

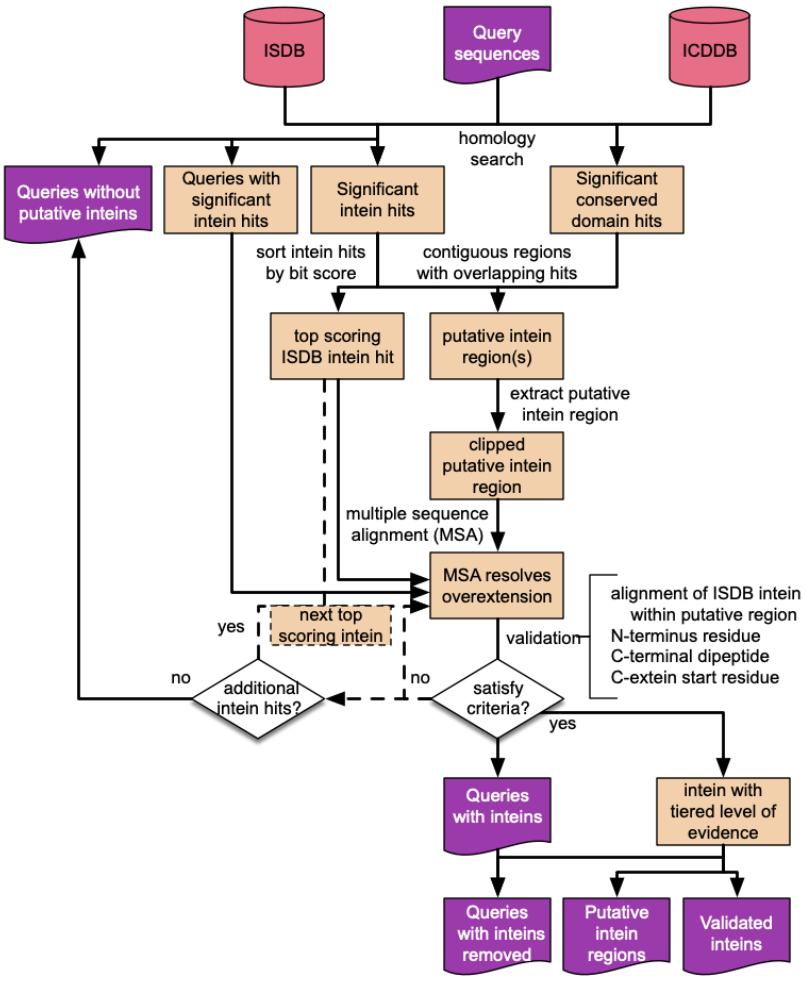

Figure 3.1: InteinFinder conceptual diagram. Query sequences are searched against two databases. One consists of a set of curated, dereplicated intein sequences (Intein Sequence Database–ISDB), and the other consists of conserved models of protein domains typically associated with intein sequences (Intein Conserved Domain Database–ICDDB). Overlapping significant hits from these searches are used to predict and extract ("clip") putative intein regions on query sequences. Query sequences that had significant hits to curated intein sequences (ISDB) are aligned with the top scoring hit and the clipped putative intein to further refine the boundaries of the putative intein, repeating this process with the next top scoring hit until the intein region N- and C-terminal boundaries are validated or all hits have been tested. The ensemble homology approach, alignment, and region refinement, bins query sequences into groups based on user-specified tiers of evidence that an intein is present, e.g., putative intein region, bonafide intein, etc.

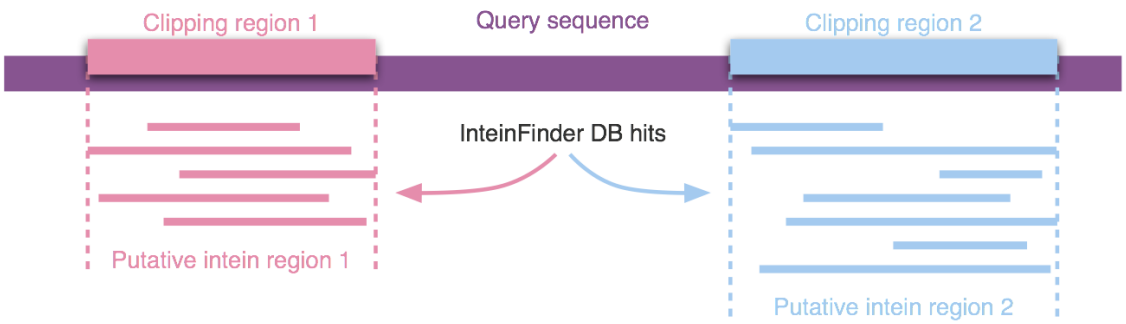

Figure 3.2: Putative intein regions are defined by overlapping significant hits to InteinFinder's databases. Putative intein regions represent contiguous regions on the query sequence that are covered by significant hits to InteinFinder's databases: Intein Sequence Database (ISDB) and Intein Conserved Domain Database (ICDDB). Hits are "tiled" along the query sequence, and regions of the query sequence with unbroken coverage are considered putative intein regions. These regions are extracted as "clipping regions" used in InteinFinder alignments.

show significant homology (E-value $\leq$ 1e-3 by default) to subject sequences in these two search databases are retained for further analysis.

Next, putative intein regions are identified on query sequences showing significant homology to a known intein sequence or conserved domain. A putative intein region is defined as the portion of a query sequence having overlapping hits to either database in the homology search (Fig. 3.2). Query sequences may have more than one putative intein region. For example, if query A has three significant hits spanning positions 100-200, 150-250, and 200-250, then the putative intein region for query A would be positions 100-250. If query A also had hits spanning region 500-600 and 550-700, then query A would have an additional putative intein region in position 500-700. Any putative intein regions longer than a user-specified length cutoff (default $\geq$ 100 peptides) are retained for further analysis.

3.2.2.2 Conserved residue validation

For each putative intein region that has a significant hit to an intein in the ISDB, sequence alignment is used for refining and verifying conserved intein residues using the following procedure. (Putative intein regions that only have hits to conserved domain models in the ICDDB do not go through the refinement and verification stages of the pipeline.)

3.2.2.2.1 Procedure

First, the putative intein is "clipped" from the query sequence using the boundary defined by the tiling procedure described above. All intein subject sequences from ISDB having significant homology to the putative intein region are sorted according to hit quality (as measured by bit score), and the top ISDB intein hit is selected. Next, the full-length query sequence, the clipped putative intein, and the selected ISDB hit are aligned using MAFFT. Including the clipped region in the multiple sequence alignment helps guide the alignment in cases where low-complexity regions cause issues or in cases where the query sequence contains multiple inteins (Fig. 3.3). Finally, using this alignment, four validation criteria are checked: (1) alignment of the ISDB intein within the putative intein region of the query peptide, (2) the presence of the intein N-terminal residue, (2) intein C-terminal dipeptide, and (4) the C-extein start residue. If at least one validation criteria is not met, then the process is repeated with the next highest scoring intein. This process repeats until a putative intein region is found that satisfies all four criteria, or until all significant ISDB intein hits are exhausted.

3.2.2.2.2 Confidence tiers

After validation, putative inteins are assigned Pass/Fail values for each validation check. The Pass values for residue checks (N-terminal, C-terminal, and C-extein start) are more granular and include user-defined confidence tiers. These tiers allow the partitioning of residues based on user requirements. For example, an N-terminal

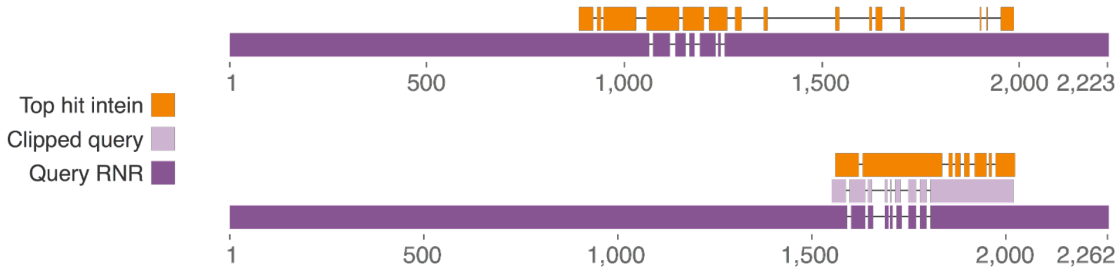

Figure 3.3: Clipped query sequences resolve overextension of alignments and refine putative intein boundaries. The alignment (top) of a query sequence (purple) to its top scoring homologous hit (orange) to Intein Sequence Database (ISDB) often results in a spurious, extended alignment, possibly due to low complexity regions, multiple inteins per query sequence, etc. InteinFinder resolves this issue and refines intein boundaries by including the clipped putative intein region (lilac) defined by overlapping hits from the ensemble homology search. Numbers at the bottom of the alignments represent the alignment column.

cysteine is assigned a Tier 1 Pass by default, as it is commonly found in known inteins and is well-supported in the literature. Other residues may be placed in lower confidence tiers as required. An overall validation result is determined by the results the four validation checks. If a putative intein fails at least one check, then its overall check also fails; however, if it passes each check, then its overall check also passes. The overall check is also assigned a confidence tier determined by the lowest tier among the individual checks. Putative intein regions that pass the overall check are referred to as "bonafide" inteins in later sections, whereas any that do not are referred to as putative inteins or putative intein regions.

3.2.2.2.3 Region refinement

For those putative intein regions that pass the overall check (i.e., bonafide inteins), the regions on the query sequence predicted from the initial homology search are refined. That is, boundaries for the putative intein region are adjusted to the positions of the first and last non-gap position of the subject ISDB intein sequence aligned against the query sequence. These intein sequences may also be removed from query

sequences using the `RemoveInteins` program included in the InteinFinder repository.

3.2.3 Constructing test query data sets

Various controlled test data sets were used to examine InteinFinder's effectiveness at identifying inteins. After construction, each of the following data sets were used as queries in the InteinFinder pipeline.

3.2.3.1 UniProt-test-data

UniProt provides a list of 104 SwissProt entries with one or more inteins, with a total of 118 (114 unique) inteins referenced as of the March 2018 release. These protein sequences were downloaded, run through the InteinFinder pipeline, and the results were manually compared to the SwissProt annotations.

3.2.3.2 RNR-real-test-data

A subset of 100 viral ribonucleotide reductase sequences (RNR) from the RNRdb [214] were selected and manually screened for inteins. Sequences were selected so that twenty sequences contained at least one intein and 80 had no inteins.

3.2.3.3 RNR-*in-silico*-test-data

A subset of 1000 manually-screened, intein-free Class I alpha subunit RNR sequences were selected from the RNRdb [214]. ISDB inteins were added *in silico* to a randomly selected subset of 500 RNRs: 250 of these received a single, randomly selected ISDB intein inserted at position 101; 250 received two randomly selected inteins (each unique) inserted at positions 101 and 201 of the original RNR sequence. Directly following each added intein, one of S, T, or C amino acids was inserted in the RNR sequence, as InteinFinder checks the C+1 extein residue.

3.2.3.4 RNR-ISDB-test-data

An *in silico* intein-containing RNR intein dataset was constructed using the full-length inteins from the ISDB. A set of intein-free RNR sequences taken from the same

set of intein-free RNR sequences used in the RNR-*in-silico*-test-data were randomly selected, one for each intein in the ISDB. A single intein was added to each of the intein-free RNR sequences, in the same way as in the RNR-*in-silico*-test-data.

3.2.4 InteinFinder sequence database characteristics

Various analyses were conducted to describe the properties of the 792 inteins in InteinFinder's ISDB.

3.2.4.1 ISDB sequence similarity

To examine the relationship of ISDB inteins to one another, all-versus-all global percent identity was calculated using a custom program (`align` version 1.0.0 available on GitHub[2]) that leverages the pairwise global aligner from Rust-Bio v1.1.0 [176] with the BLOSUM62 scoring matrix, a gap-open penalty of 10, and a gap-extend penalty of 1. Percent identity scores were visualized with a heatmap in R v4.1.2 using the `heatmap.2` function from the gplots package v3.1.3. Heatmap dendrograms based on sequence percent identity (converted to distances with $100 - P$ where P is the percent identity) were made using the `hclust` function in R using the Ward D2 option for dendrogram agglomeration. Data range was calculated with the quantile function in R.

3.2.4.2 Sensitivity of individual ISDB inteins

Rather than using the full ISDB in each InteinFinder run, a reduced intein sequence database containing a single entry from the standard ISDB was used (i.e., the InteinFinder pipeline was run once for each of the 788 full-length inteins in the ISDB, using a reduced intein sequence database containing a single intein from the ISDB). The query set used for each run was the RNR-ISDB-test-data, excluding the RNR with the intein currently being used as the target database. The number of putative and bonafide inteins was recorded for each run, along with their percent

[2] https://github.com/mooreryan/align

identity to the single sequence in the target database. These scores were then plotted using kernel density estimation from the ggplot2 R package. To evaluate the differences in percent identities between bonafide and putative inteins, the Wilcoxon rank sum test for difference in location was used.

3.2.4.3 Intein collectors curve

As a proxy for the comprehensiveness of ISDB, a collectors curve was generated from the ISDB intein sequences. The ISDB sequences were clustered at 30% peptide identity using MMseqs2 [358] followed by rarefaction analysis in QIIME [52]. Finally, the curve was visualized using ggplot2 in R.

3.2.5 IMG/VR Methods

The IMG/VR database of uncultivated viral genomes (UViGs) v4.1 (released Dec. 2022) from the JGI Genome Portal, which included a set of 112,567,455 high-confidence peptide sequences, was obtained [49]. IMG/VR uses the first four levels (Ecosystem, Ecosystem Category, Ecosystem Type, and Ecosystem Subtype) of the JGI GOLD classification system [151] to provide detailed ecosystem information for the UViGs.

3.2.5.1 Identifying inteins from IMG/VR peptide sequences

InteinFinder (version 1.0.0-SNAPSHOT [7a303c7], default settings except – MMseqs2 number of search iterations: 1, MMseqs2 sensitivity: 4) was used to identify putative inteins from this data set. Third-party software versions used in the pipeline were as follows: MAFFT version v7.490, MMseqs2 version 5ae55, rpsblast and make-profiledb version 2.13.0+. InteinFinder pipeline databases, ISDB and ICDDB, were the default databases for the version of InteinFinder used for the experiment.

3.2.5.2 Intein distribution across ecosystems

Intein under- and over-representation was estimated using a ratio of bonafide inteins to background sequences per environment (Bonafide to Background ratio). The

Bonafide to Background ratio for each ecosystem was calculated as follows: (1) calculate the proportion of all bonafide inteins in that were identified in that ecosystem, (2) calculate the proportion of total UVIGs originating from that environment, (3) calculate the $\texttt{log}_2$-ratio of the value from (1) to the value from (2). Note that the number of UVIGs and the number of protein sequences per ecosystem were highly correlated ($r = 0.97$), and so the former was used as an estimate of the latter, as the data was more readily available (Fig. 3.4). The result is a value ranging from $[-\infty, \infty]$. A value of zero indicates an environment with exactly the expected number of inteins given the size of that ecosystem. Positive numbers indicate over-representation and negative numbers indicate under-representation. For example, if an environment contained 50% of all bonafide inteins, but only 25% of total UVIGs from the IMG/VR dataset, its Bonafide to Background ratio would be $\texttt{log}_2(0.5/0.25) = 1$, indicating about twice as many inteins as expected.

3.2.5.3 IMG/VR functional annotation

Functional annotation of IMG/VR sequences was done using the GhostKOALA automatic KO assignment and KEGG mapping service version 2.2 (released May 15, 2019) [159], with the search parameters "genus_prokaryotes + family_eukaryotes + viruses". Given the 500,000 sequence limit per annotation job, random samples with 500,000 sequences each of the IMG/VR peptides individually were submitted. The first five samples were generated using the `sample_seqs` program (version 1.0.0 [accf101], random seed: 53643). At a later date, twenty additional samples were generated with `sample_without_replacement` program (version 1.0.0 [90a4be8]), with random seed 294342) Both programs are available on GitHub.[3]

KO terms were mapped to higher level KEGG terms using custom scripts to parse the mapping provided by the KEGG BRITE database (accessed 2023-03-01). In this way, each annotated sequence was assigned one or more "paths" through the BRITE hierarchy: KO term → C level term → B level term → A level term → root.

[3] https://github.com/mooreryan/sampling

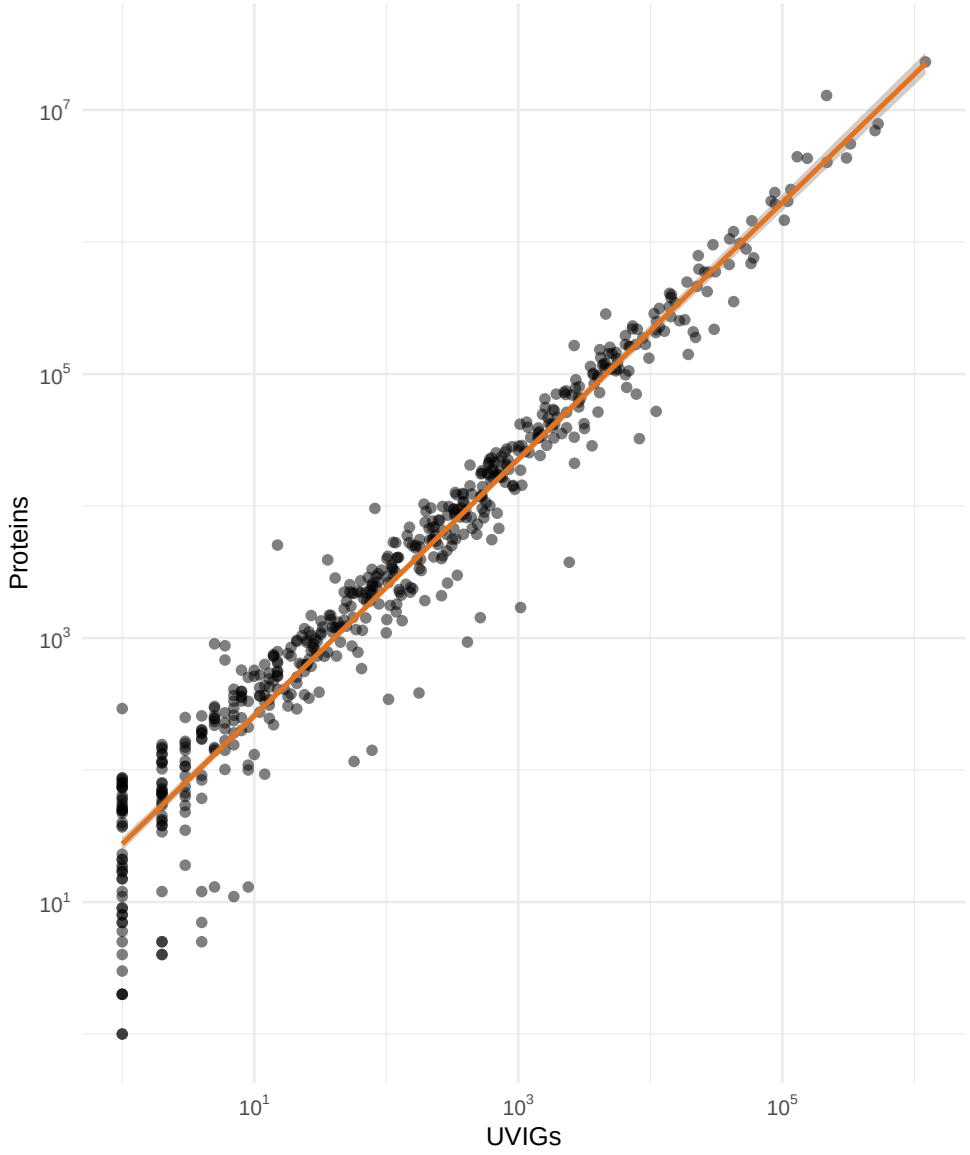

Figure 3.4: Number of UVIGs is highly correlated with number of proteins per ecosystem. Scatterplot of the number of uncultivated viral genomes (UVIGs) and the number of proteins for each ecosystem in IMG/VR. Orange line represents log-log linear regression of number of proteins on number of UVIGs. The solved equation is $\texttt{log}_{10}(y) = 1.44 + 0.97 \times \texttt{log}_{10}(x)$, with $R^2 = 0.94$, where x is the number of UVIGs in a given ecosystem and y is the number of proteins in a given ecosystem.

Note that many KO terms (the lowest level of the hierarchy) have multiple paths to the root. That is, nodes can have multiple parent nodes in addition to parent nodes having multiple children. Annotations were limited to those with the following A level terms: 09100 Metabolism, 09120 Genetic Information Processing, 09130 Environmental Information Processing, 09140 Cellular Processes.

Principal components analysis (PCA) on the centered log-ratio proportions of KEGG annotations at three levels, A, B, and C, were used to generate KEGG profile ordinations for the various environments.

3.3 Results & Discussion

Inteins are autocatalytically self-splicing protein introns that are of interest to a broad community of scientific researchers. The ability to break and form peptide bonds has made inteins powerful biotechnological tools with applications in molecular biology, protein chemistry, and other fields [341, 208, 385]. Evolutionary biologists have an interest in identifying intein sequences, either because researchers need to remove them prior to analyses (e.g., [135]) or because of their utility in tracking gene transfer events [352, 167]. Inteins are also useful to ecologists, as it has been demonstrated that inteins likely provide a means of post-translational regulatory control of their host peptide sequences (exteins) through environmental "sensing" [373, 372, 192]. This suggests that inteins have a marked impact on the fitness and ecology of the organisms and genes in which they reside, which is in contrast to the early days of their discovery and study when inteins were generally thought to be neutral or selfish genetic elements [117, 250].

A common feature of intein studies, regardless of their specific focus, is the need to identify inteins in protein data sets ranging from genome collections like NCBI RefSeq [264] to individually collected metagenomic samples. These studies generally follow a similar set of steps for identifying inteins: creating or using an existing reference database of intein sequences, searching the sequences of interest against that

database, and validating the putative inteins for important protein features like conserved residues or splice junctions (e.g., [167, 261]). Given the common features of these studies, a similar procedure was consolidated into a reusable intein-discovery pipeline. As the significance of inteins in various fields grows, an automated intein-discovery pipeline will become increasingly valuable.

3.3.1 InteinFinder pipeline considerations

InteinFinder combines common practices from many large intein surveys and adapts them to form a single pipeline (e.g., [352, 261, 167, 121]). InteinFinder's goal is to lessen the burden of manual identification and curation of inteins in large peptide data sets by standardizing the search for inteins and increasing the accessibility to a greater number of researcher groups. InteinFinder is scalable to datasets containing hundreds of millions of peptide sequences, making it well suited to the large metagenomic datasets common in microbiome and environmental research.

InteinFinder includes two databases: an intein sequence database (ISDB) of experimentally and computationally predicted intein sequences, and an intein conserved domain database (ICDDB). The InteinFinder pipeline combines multiple homology search methods to increase its ability to identify inteins. Query sequences are searched against a curated set of validated intein sequences (ISDB) using MMseqs2 [358], and against a set of intein-associated domain models from the NCBI conserved domain database (ICDDB) using RPS-BLAST [408]. Its ensemble method allows InteinFinder to identify a greater amount of inteins than using either intein or conserved domain homology methods individually. Any query sequence that has a hit to at least one target in either of InteinFinder's databases and passes the user-defined length filter will have a putative intein region predicted.

3.3.2 Validating the InteinFinder pipeline

The first consideration in developing a discovery and identification pipeline for any protein is ensuring the results are accurate. Multiple validation experiments were

used to test InteinFinder's ability to recapitulate the results of manual intein curation. To this end, three different tests were conducted to evaluate InteinFinder's accuracy in identifying and demarcating inteins within query sequences. Pipeline results were compared with the known true locations of inteins.

First, InteinFinder was tested against SwissProt (March 2018 release) which contained 104 peptide sequences having one or more inteins, with a total of 118 (114 unique) intein sequences (UniProt-test-data). InteinFinder correctly identified all 118 intein regions, with all regions marked as bonafide (Table 3.2). InteinFinder did not identify any inteins that did not exist in the test data. SwissProt and InteinFinder annotations disagreed in four sequences by one amino acid. Manual inspection of these sequences found that SwissProt included the up- and down-stream extein residue(s) in the annotation, whereas InteinFinder did not. Additionally, InteinFinder extracted one intein sequence (P74750 (1)) as a single intein whereas SwissProt separated this intein in two (P74750 (1 & 2)). This split intein interrupts the DNA polymerase III alpha subunit from *Synechocystis* sp. (strain PCC 6803 / Kazusa). The protein is encoded by two separate genes (dnaE-N and dnaE-C) 745 kb apart in the genome and the two halves are spliced *in trans* by the split intein [400, 90], so it is logical that the SwissProt annotation is also split. Given that the protein itself is presented as a single entry in the database, InteinFinder's start and stop annotations accurately reflected the intein boundaries within the protein sequence.

Second, a set of 20 manually-annotated, intein-containing ribonucleotide reductase (RNR) sequences from the RNRdb [214] were selected, containing 26 total intein regions (RNR-real-test-data). An additional 80 manually-screened, intein-free RNRs were randomly selected from RNRdb. InteinFinder identified all 26 intein regions within the 20 intein-containing RNRs, and no inteins within the 80 intein-free sequences. Twenty-four of the 26 intein regions were marked as bonafide, and two regions were marked as putative. However, the two putative intein regions did contain the true intein within the predicted start and end position.

Third, intein sequences were added to RNR sequences *in silico* creating a

Table 3.2: SwissProt intein test set proteins with mismatches (UniProt-test-data)

Acc. No.	Protein	Intein No.	InteinFinder		SwissProt	
			Start	End	Start	End
O33845	DNA polymerase	2	855	1392	856	1392
O55716	Ribonucleoside-diphosphate reductase large subunit	1	272	610	272	611
P74750	DNA polymerase III subunit alpha	1	775	933	775	897
P74750	DNA polymerase III subunit alpha	2	-	-	898	933
P74918	DNA polymerase	2	901	1289	901	1282
Q58445	DNA-directed RNA polymerase subunit Rpo1N	1	460	911	461	911
Q9F5P4	Replicative DNA helicase	1	16	350	16	351

database containing 500 intein-free RNRs, 250 RNRs with one ISDB intein inserted at position 101, and 250 RNRs with two ISDB inteins inserted at positions 101 and 201 (RNR-*in-silico*-test-data). InteinFinder did not identify inteins within the 500 intein-free sequences. It correctly identified the start and end positions of 743 (99.1%) of the 750 total inteins added to the dataset. All seven incorrect inteins were found on RNRs with two intein regions. While the exact start and end coordinates were incorrect, the actual intein location was within the putative intein region that InteinFinder predicted.

Through these experiments, InteinFinder was shown to be both sensitive and specific. Additionally, even in the few cases in which the boundaries of predicted inteins disagreed with manual annotation, the true intein was within the boundary predicted by InteinFinder. That is, the intein was still correctly identified, even while not being perfectly delimited in the extein sequence.

3.3.2.1 Parameter and sensitivity tuning

Users must consider the tunable parameters of the homology search tools when using InteinFinder. While the default parameters were shown to work well in the above tests, it is likely that that the specific scientific question being addressed could call for adjustments. Less stringent E-value cutoffs, higher sensitivity levels, and greater numbers of search iterations all lead to a larger pool of putative intein sequences and may capture inteins with distant homology to InteinFinder's databases. This may be desirable in cases where query sequences originate from environments not well represented in the InteinFinder database, or in a more exploratory or discovery based setting. This increase in sensitivity may also increase the number of false positives [40]. However, more sensitive searches could still be applied without increasing the amount of false positive non-inteins by giving increased manual scrutiny to sequences to which InteinFinder gives a lower confidence tier.

3.3.2.2 Database comprehensiveness & efficacy

Expanding on the previous point regarding the tradeoffs between sensitivity and specificity, any homology-based database-backed discovery pipeline will be limited in its ability to discover novel features by the comprehensiveness of its database [414, 370, 79]. Many of the inteins in the ISDB are from InBase [282], which includes user-submitted intein sequences from across the tree of life, including viruses and phages. InBase has historically contained many inteins found in the literature, i.e., inteins that have high importance in biochemical operations and those from commonly studied laboratory organisms or genes. Though InBase contains inteins from a variety of organisms, it likely suffers from the same "research bias" that affects other protein databases, that is, sequences similar to early key model organisms will have a much higher annotation rate, and many environmentally important groups will be underrepresented [29, 204].

To add valuable diversity to ISDB, inteins from large-scale surveys of mycobacterium phage [167] and microeukaryotes [121] were included. Acquisition of these sequences was straightforward as the authors provided data in a format amenable to data collection. At the time of data collection (2018), data or inteins from other existing large-scale studies were challenging to incorporate or were not amenable to the curation process. However, these studies and other more recent intein surveys (e.g., [141]) could be added in later updates to the InteinFinder databases. Notably, users can easily add new intein sequences or profiles to InteinFinder's databases or even opt to use their own databases entirely, decoupling InteinFinder's usefulness from depending on a central curated intein database. While there have been large-scale studies of intein discovery across protein datasets such as NCBI's Gene database [262] and RefSeq [261], to our knowledge, there are no large scale, multi-environment environmental sequencing studies in which the focus has been intein identification. Thus, determining a baseline of global intein diversity across ecosystems remains an open question. Regardless, care was taken to include inteins from multiple studies of different kinds of organisms, and the collector's curve suggests that the environments in ISDB may be

Table 3.3: Intein sequence database (ISDB) intra-cluster percent identity quantiles.

Cluster	No. seqs	PID Quantiles		
		2.5%	50%	97.5%
1	49	43.3	58.2	98.8
2	48	19.5	26.0	77.0
3	88	9.9	18.8	52.9
4	47	27.7	42.1	99.1
5	47	27.2	35.0	99.6
6	15	42.3	62.5	100.0
7	39	19.1	29.7	100.0
8	13	54.2	73.9	100.0
9	20	40.4	48.2	100.0
10	21	18.3	35.8	100.0
11	119	12.9	18.3	56.1
12	286	13.3	18.5	27.2
All	*792*	*7.6*	*16.2*	*26.7*

fairly well represented[4] (Fig. 3.5).

Various analyses of ISDB intein characteristics were conducted. First, the similarity between ISDB inteins was examined. Global percent identity was calculated for each combination of sequences in the ISDB. While most intein pairs had low percent identity scores (median: 16.2%, 95% quantile range: [7.6%, 26.7%]), there were localized groups of highly similar inteins (Fig. 3.6, Table 3.3).

Next, the sensitivity of each ISDB intein for identifying novel inteins was examined. Each full-length ISDB intein (788) was used as a target database against the corresponding intein-containing RNR query sequences (RNR-ISDB-test-data). The median number of bonafide inteins predicted per run was 57 (95% quantile range: [3, 159]), while the median number of putative inteins was 773 (95% quantile range: [773, 776]). The median percent identity for predicted bonafide inteins across all runs was

[4] The comprehensiveness of the ISDB was examined using a collector's curve of 30% clusters of ISDB inteins. While the curve does not plateau completely, it does begin to level off as the number of inteins included in the analysis increases.

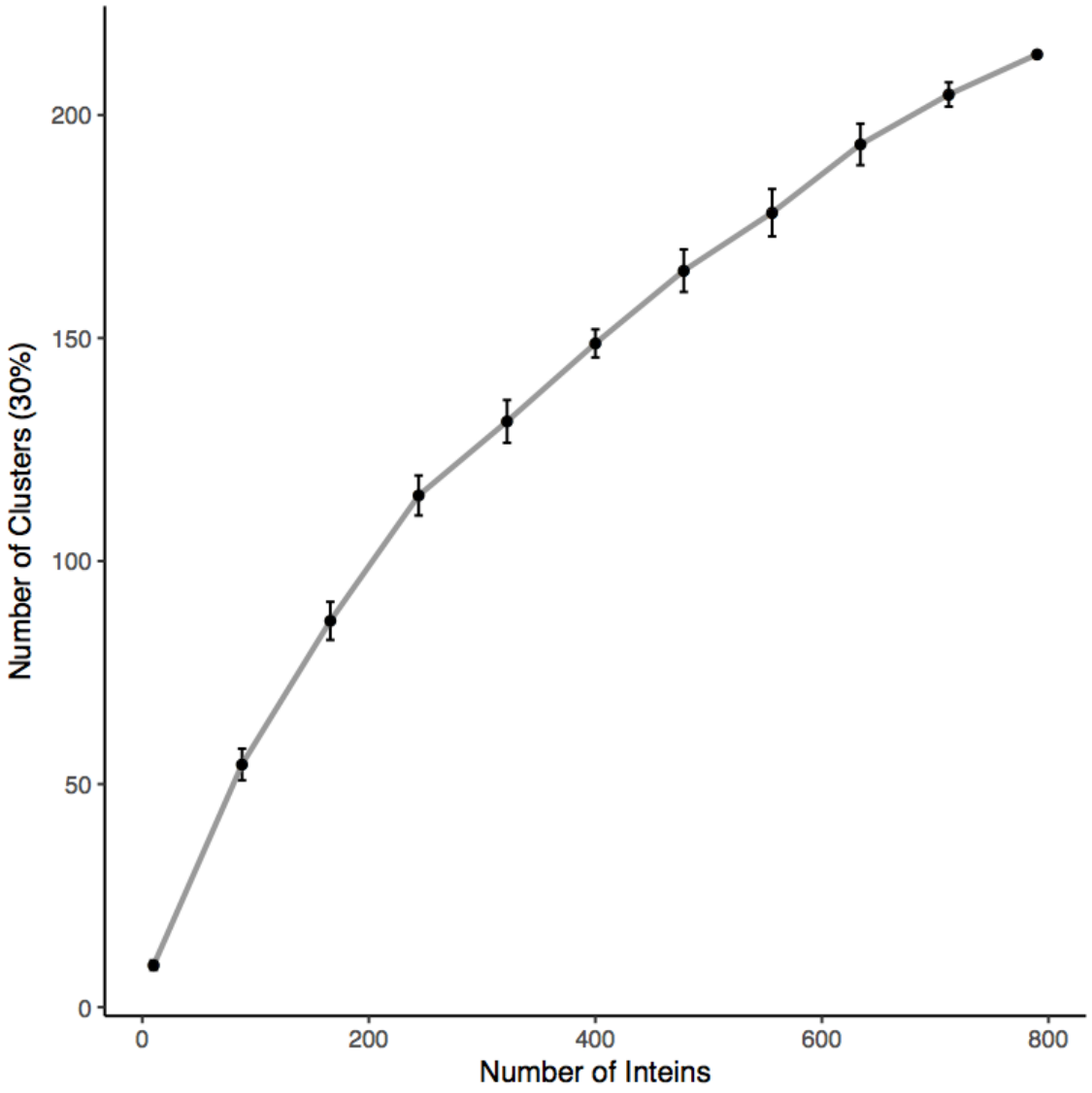

Figure 3.5: The collectors curve of 30% amino acid clusters of inteins from InteinFinder intein sequence database (ISDB) suggests that model systems and environments sampled in ISDB are well represented. ISDB inteins were sampled at 10 steps between 10 and 790 sequences, 10 iterations each. Intein sequences from each subset were clustered with MMseqs2 at 30% identity. The mean number of clusters and standard deviation was plotted at each rarefaction level.

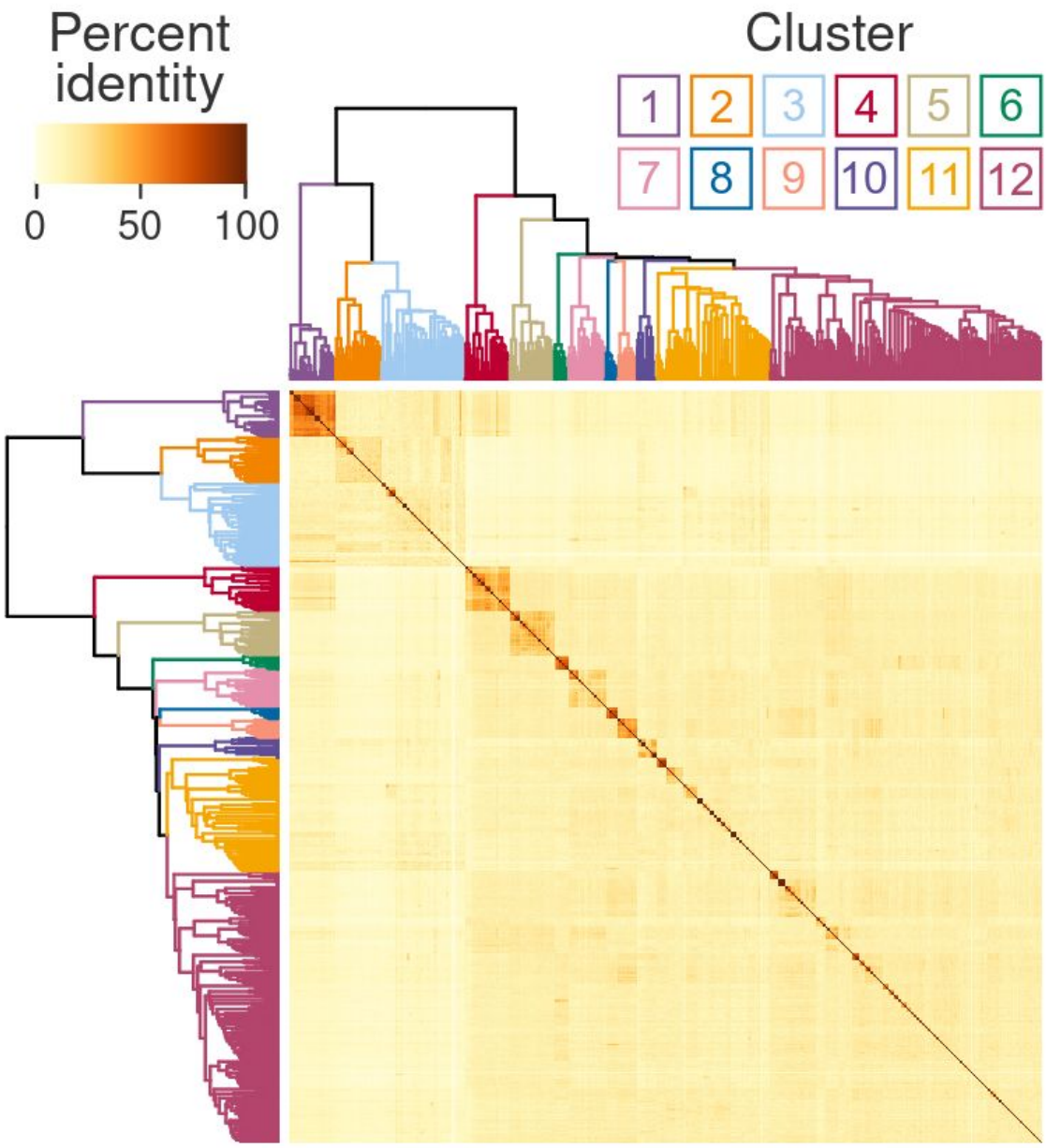

Figure 3.6: Sequence percent identity of Intein Sequence Database (ISDB) shows a module-like network structure. Localized groups of highly-similar inteins cluster together, and are highly dissimilar to most inteins outside of the cluster. The similarity network is displayed as a heatmap, where the yellow-orange-brown color scale represents the global percent identity of each intein pair, and dendrograms are hierarchical clustering of ISDB intein sequences based on the percent identity. Clusters defined by the dendrogram are labeled from 1-12 and colored for clarity.

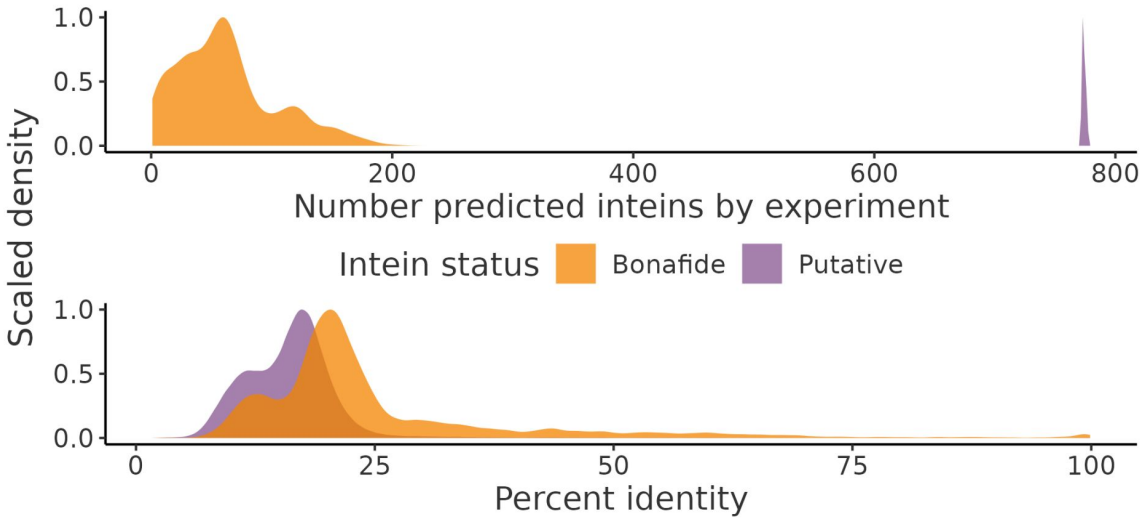

Figure 3.7: Sensitivity of ISDB sequences at identifying novel inteins. Kernel density estimates of the distributions of number of predicted inteins and the percent identity of bonafide (orange) and putative (purple) inteins from in the RNR-ISDB-test-data set. Distributions were significantly different according to the Wilcoxon rank sum test for difference in location.

21% (95% quantile range: [10%, 68%]), and median for predicted putative inteins was 16% (95% quantile range: [8%, 26%]). Wilcoxon rank sum test indicated a significant non-zero location shift between the bonafide and putative inteins, with estimated difference in location of 5.45 (95% confidence interval: [5.40, 5.51]) (Fig. 3.7). Notably, InteinFinder was still able to identify putative inteins in sequences even when those sequences were quite distant from the sequence in the target database.

On average, the inteins in the ISDB have low sequence similarity scores. However, rather than the scores being uniformly low, the similarity network has a module-like structure in which localized groups of highly-similar inteins cluster together, and are highly dissimilar to most inteins outside of the cluster, with some cases of inter-cluster similarity (Fig. 3.6, Table 3.3). There are many possible explanations for this structure. First, though inteins are a diverse set of proteins [366], they have common regions (i.e., shared blocks) [287, 88], and many share common protein domains like HNH homing endonucleases [26, 116, 85], which could be driving the connections both

within and between intein modules. Second, inteins show a mix of vertical and horizontal descent [288, 341, 167, 121]–inteins transferred vertically may be more similar than those transferred horizontally. Finally, inteins that incorporate in the same proteins at the same positions are known to be more similar to one another than to inteins from different peptides or even to other inteins at a different insertion site in the same type of protein [281, 366]. Still, further work is needed to fully understand the mechanisms behind the network structure.

The ICDDB includes intein-specific domains and domains for mobile elements like endonucleases, which are associated with many inteins. Conserved domain searches are highly sensitive and increase the ability to identify full length putative inteins that are dissimilar to known inteins included in the ISDB. For example, in the IMG/VR dataset (described below), of queries with at least one significant hit to one of the InteinFinder databases, 84% of them only had hits to the ICDDB. A broad conserved domain search will increase the search sensitivity, but can also increase false positives because sequences with similar domains to inteins, that themselves are not inteins, may be recovered. However, only those queries that also have hits to an intein sequence in the ISDB are retained for alignment and validation, and false positives are reduced through InteinFinder's tiered evidence system.

As an example of the utility of the ensemble approach, the predicted intein region of an RNR sequence from *Fimbriimonas ginsengisoli* (acc. AIE87195.1) based only on homology to sequences from the ISDB was 815-849. This was too short to be included in InteinFinder's refinement steps. However, including hits to the ICDDB expanded the putative intein region to 481-852. The final predicted intein region after validation and refinement was 483-849, which matched manual annotation.

There were cases during pipeline development in which aligning the full length query sequence one-by-one with its ISDB hits led to errors in the putative intein region start and end positions due to over-extension of the alignment. InteinFinder automates the alignment process without user supervision by leveraging the putative intein regions defined by overlapping homology search results. The putative intein region is "clipped"

out of the query sequence, and aligned with the full length query sequence and top scoring hit from the ISDB. The clipped intein region guides the alignment, preventing over-extension even in cases where queries had multiple inteins (i.e., Fig. 3.3).

Through this alignment process, key criteria of the putative intein region are checked: N-terminal residue and position, C-terminal dipeptide and position, and the extein C-terminal +1 residue. For example, an intein with a C-terminal dipeptide of HN is both common among known inteins and supported in the literature, so by default it is labeled as a Tier 1 pass, whereas other peptide pairs may be placed in lower tiers based on their level of confidence (e.g., has literature support but is rare in ISDB, or is common in ISDB but has no direct literature support). Thus, the subset of putative inteins retained for analysis can be customized based on the strictness of the research questions. If false positives are highly detrimental to the question at hand, users may restrict downstream analysis to only putative inteins with the highest levels of support. Contrastingly, more lenient selections may be made if a more exploratory approach with further manual curation is desired.

The InteinFinder pipeline does not currently consider several other noteworthy intein features. Inteins have various conserved motifs critical for proper functioning, commonly referred to as blocks, in which certain groups of residues tend to be highly conserved [281]. Residues in these blocks could be leveraged to provide insights into splicing dynamics and the intein-extein relationship [104]. By identifying these blocks in predicted inteins, valuable information can be added to downstream analyses. Apart from blocks, specific intein residues and non-conserved regions have been identified as critical for intein functioning [190]. Particular extein residues have even been implicated in intein functionality [268]. A future version of InteinFinder could leverage this and other information from the literature to predict and annotate these features on predicted inteins. Many of the inteins currently in the ISDB come from InBase and are well annotated. These annotations could be incorporated into InteinFinder's output to enrich the information included with the predicted inteins. Inteins have different splicing mechanisms and there are multiple intein families [354, 141]. Automatically

annotating these features would provide insights into possible biochemical properties of predicted inteins. Sensitivity, or the ability to identify inteins highly divergent from those in then InteinFinder database, may be increased by leveraging latent signatures in the primary sequence structure of queries identified using large language models (LLMs), or other machine learning techniques that could be integrated into the InteinFinder pipeline (e.g., [249, 411]).

In addition to extensions to InteinFinder's core discovery pipeline, the intein databases included with InteinFinder could become an independent resource. In this study, over 70,000 bonafide inteins were identified from across the biosphere, greatly expanding the known intein sequence space. Ideally, these inteins would be cataloged and incorporated into a central repository like InBase; however, InBase is no longer maintained. Given the ease at which InteinFinder allows for intein discovery, an updated central repository for inteins would be a welcome addition to the field.

3.3.3 Viral intein diversity across the biosphere

While inteins possess unique characteristics that make them highly valuable for various biotechnology applications, [385], there are two major reasons that inteins can provide useful insights to microbial ecologists: (1) inteins are mobile elements and are markers for gene flow and horizontal gene transfer events [352], and (2) inteins are likely to be post-translational regulators of extein peptide function (e.g., [373, 190, 192]). Given the mounting evidence of their ecological importance, patterns of intein distribution and ecology were explored using IMG/VR, a database of peptides from viral genomes and viral metagenomic contigs identified in microbial metagenomes submitted to the Joint Genome Institute's Integrated Microbial Genomes portal [49]. IMG/VR provided the potential for novel intein discovery and ecological insight as it spans thousands of metagenome experiments across many of Earth's ecosystems [49], is supported by comprehensive metadata, and focuses on viruses, which are still comparatively understudied as compared to other microbial life [364].

InteinFinder was used to identify inteins within IMG/VR. There were 72,249 inteins that passed InteinFinder's checks spread across 69,605 unique extein sequences. In addition to these "bonafide" inteins, InteinFinder identified 980,336 putative inteins. Of these, 70,716 of the bonafide and 936,406 of the putative inteins occurred on extein sequences that had associated metadata regarding ecosystem of origin.

To analyze the functional profile of sequences in IMG/VR, 12.5 million sequences were submitted to GhostKOALA for KO term annotation. Of these, 1,098,487 sequences were successfully annotated with 493,603 sequences falling under one of the four A level KEGG BRITE terms used in this study: 09100 Metabolism, 09120 Genetic Information Processing, 09130 Environmental Information Processing, 09140 Cellular Processes. While the magnitude of KEGG annotated proteins is low compared to the total number of sequences in the IMG/VR database, bootstrap analysis showed little variation in the proportional makeup of A term annotations (Fig. 3.8), indicating that additional samples likely would not be drastically different in their annotations as compared to the sequences that were annotated.

3.3.4 Environmental bias in intein distributions

For IMG/VR proteins with environmental annotations, the ratio of the distribution of the bonafide-intein-containing sequence subset (designated intein-containing peptides, ICP) to the distribution of background sequences[5] was compared across environments. ICP and background distributions differed within and across ecosystems, indicating that environmental effects on intein distribution were likely non-random (Fig. 3.9). These observations may have multiple explanations: (1) despite best efforts, InteinFinder's databases are sufficiently biased that inteins in underrepresented environments are too dissimilar to be identified, (2) the environments with more/fewer ICP than expected have more/fewer of the proteins in which inteins are more common, thus leading to their over/under representation, or (3) there is environment-specific

[5] As estimated by the number of UViGs, see Section 3.2.5.2.

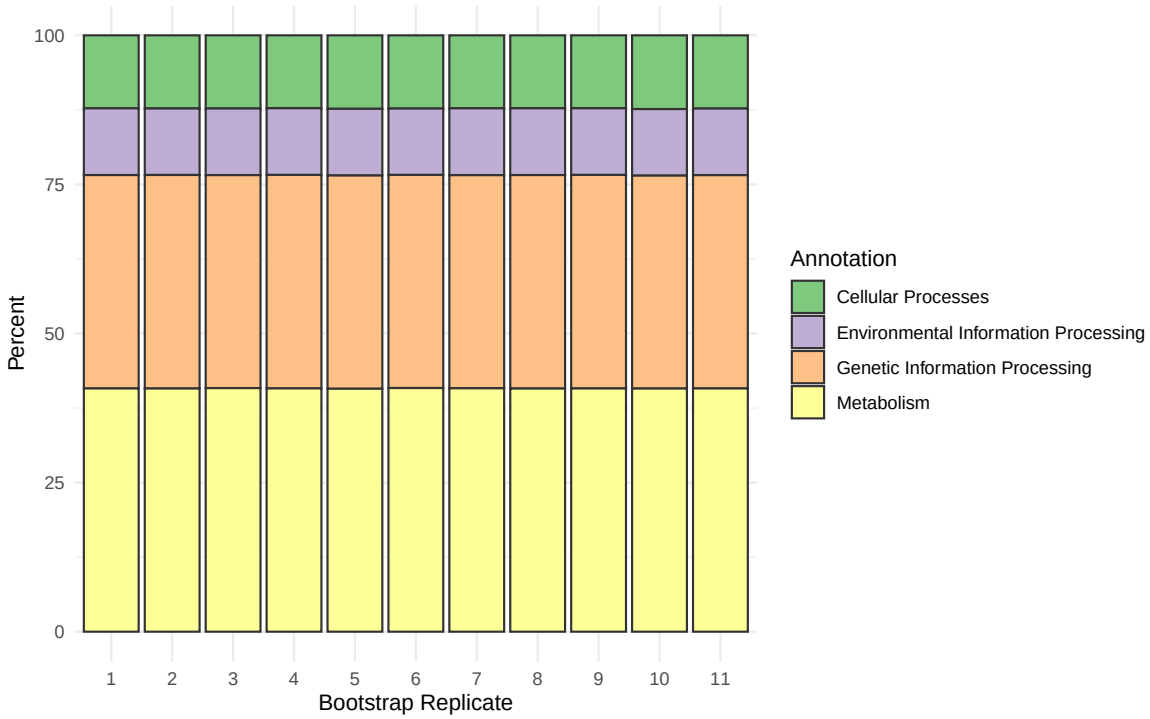

Figure 3.8: Bootstrap analysis of KEGG annotation of IMG/VR sequences. Bootstrap analysis of KEGG annotation proportion in the annotated IMG/VR sequences. Bars represent proportion of proteins in each bootstrap replicate annotated as the given term, with color indicating the annotation (green: cellular processes, purple: environmental information processing, orange: genetic information processing, yellow: metabolism). Observed variation across bootstraps was very low.

selection at work, i.e., there is some change in environmental pressures that affects retention/fixation of inteins in the members of that environment.

The simplest explanation for the decreases in ICP compared to the background observed in some environments is that inteins in those ecosystems are less likely to be identified through homology because they are underrepresented in InteinFinder's databases. The observed median identity score between ISDB inteins (16%) is lower than the median percent identity found in the RNR-ISDB-test-data experiment (21%). That is, even among known inteins in the InteinFinder database, there exist high levels of sequence diversity. Thus, it could be the case that that many environmental inteins are too distant from the sequences in the ISDB to be identified as bonafide inteins. However, similar, though less pronounced trends in over- and under-representation are still observed when all putative intein regions retained, rather than restricting the analysis to bonafide inteins only (data not shown), indicating that the observed differences between ICP and background proportions are likely not an artifact of overly strict similarity thresholds. While including lower-confidence, putative intein regions increases the probability that some of the putative intein regions are actually non-intein mobile elements, the similarity in trends between bonafide and putative predicted inteins suggests that a more permissive search would not markedly change the interpretation of the result.

Another potential explanation is that environments in which inteins are more abundant than expected contain a proportionally higher number of proteins that commonly harbor inteins. Inteins are not evenly distributed across protein types, rather, they are more likely to be found in replication, recombination, and repair (RRR) and nucleotide metabolism proteins [262, 261]. In this study, KEGG term annotations were used as a proxy for protein function in the various ecosystems. Ordinations were used to examine any potential connections between ecosystems, their A, B, and C level KEGG functional profiles, intein under- or over-representation, and total number of inteins (Fig. 3.10). However, no strong patterns were observed in the ordinations. Bonafide/total ratios (point color), and bonafide intein counts (point

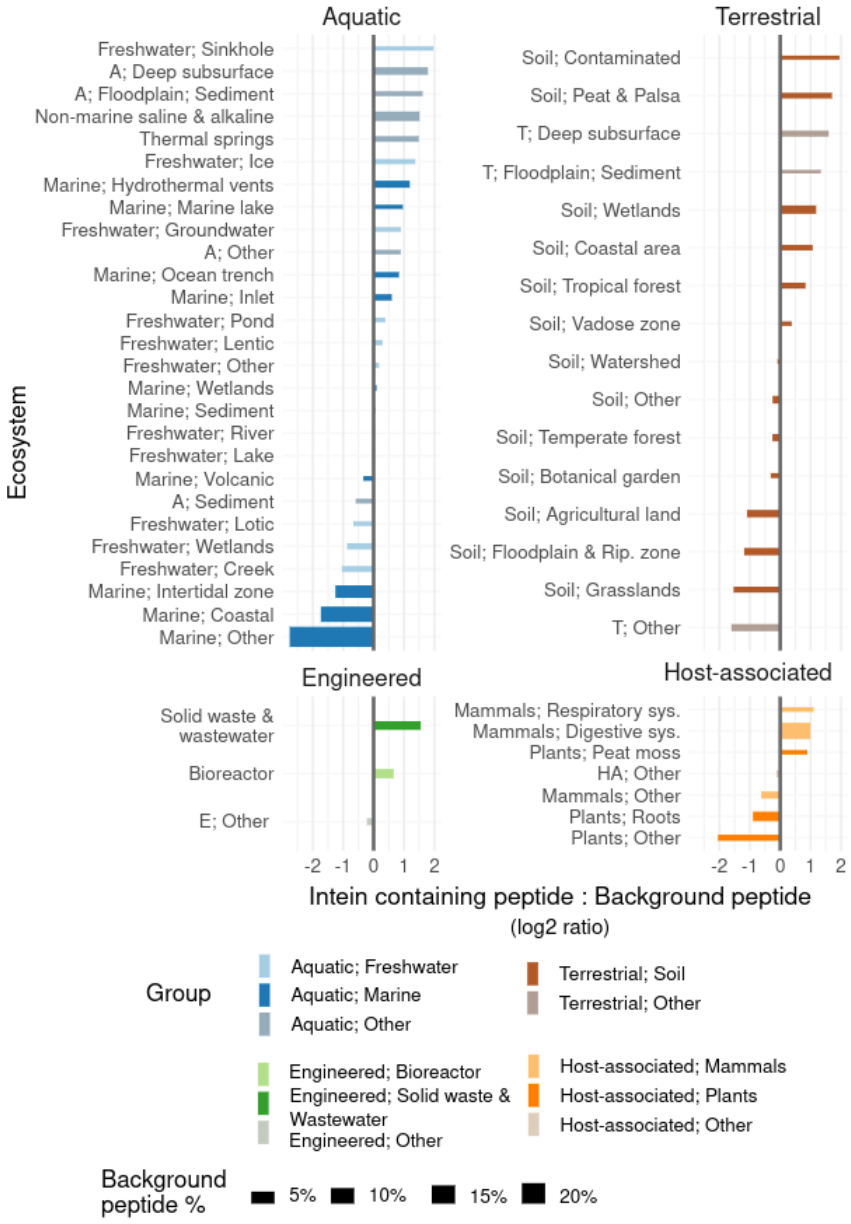

Figure 3.9: Proportional under- and over-representation of intein containing peptides across various ecosystems. Proportional shifts between intein-containing peptides (ICP) and the background across aquatic (blue), terrestrial (brown), engineered (green), and host-associated (orange) ecosystems. Within panels, color shades indicate more granular ecosystem designations. Bar width indicates ecosystem proportion in the IMG/VR dataset, with wider bars indicating a larger proportion of total IMG/VR peptides originating from that ecosystem. A positive $\log_2$-ratio indicates higher than expected numbers of inteins in that ecosystem, whereas as negative $\log_2$-ratio indicates a lower than expected number of inteins.

size) were distributed haphazardly in all three ordinations, suggesting that there is no measurable link between the KEGG functional profiles and the measured under- and over-representation of inteins in specific environments. Therefore, it is unlikely that the differences in intein proportional abundance among ecosystems are due to variations in the protein composition of those ecosystems.

An alternative explanation for the observed differences in ICP and background proportions across environments is that they are driven by an ecological or evolutionary force, such as HGT or stress-response. Horizontal gene transfer (HGT) is a common mechanism by which populations obtain inteins [167], and intein presence can be considered a marker of HGT events [353]. HGT has been observed more frequently in extreme environments such as hot springs, certain sediments, and oil wells [107] and in response to fluctuating environmental conditions and stressors [102, 42, 198]. Furthermore, the environmental stressors that trigger mass HGT events can require rapid survival responses [42], creating conditions in which inteins could provide particular benefit to their host populations, as post-translational protein regulation provides a rapid mechanism for responding to changing conditions [154, 360]. Some inteins have also been shown to act as environmental sensors, splicing out of the extein in response to external changes [373, 191, 190], making inteins yet more valuable in dynamic environments.

In this study, inteins were more commonly enriched in environments that are more extreme (e.g., high temperature, salinity, or alkalinity) or more variable (e.g., floodplains and some host-associated systems) (Fig. 3.9). Many of the environments that were enriched for inteins and have been shown or hypothesized to have high levels of HGT include thermal springs [107], sediments [107], wastewater [161, 148], solid waste [410], contaminated soils [349], hydrothermal vents [12], and the human microbiome [348, 126].

While some environments were enriched for inteins, others showed fewer than expected. The marine environment, including oceanic and coastal habitats, showed the largest proportional decrease between ICP and the background (Fig. 3.9). This

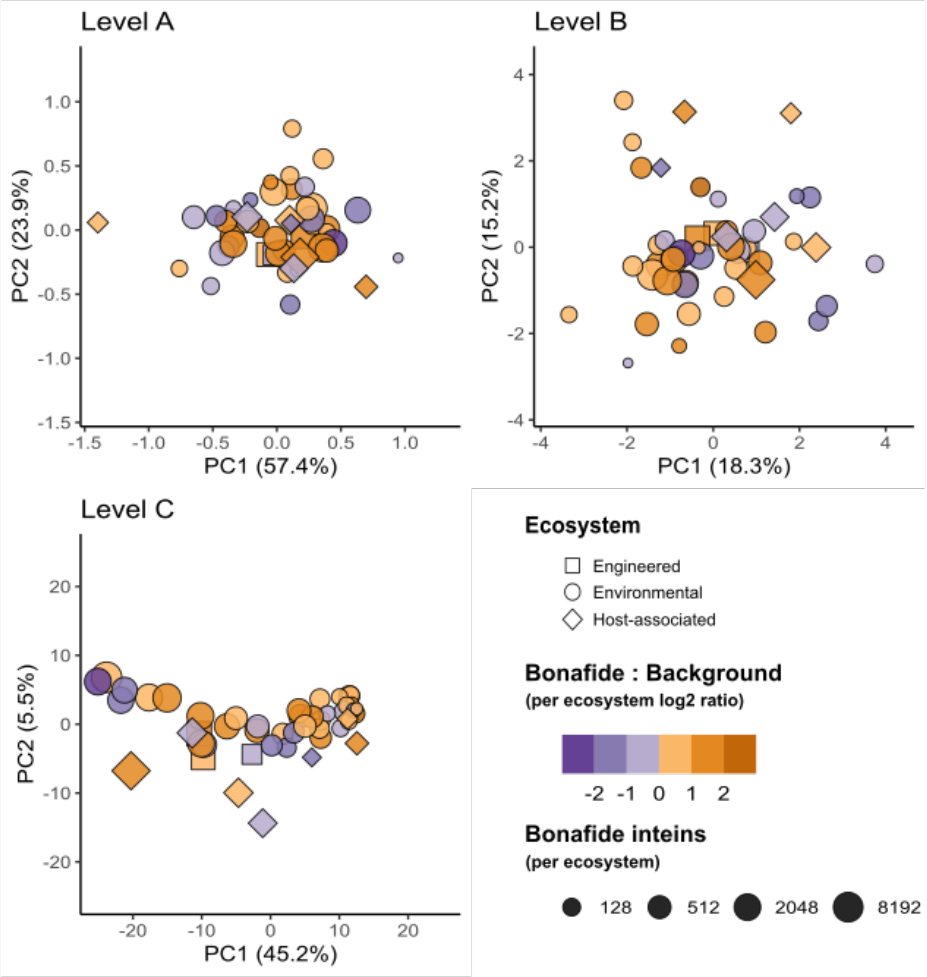

Figure 3.10: Ordination of ecosystems using KEGG annotations does not reveal clustering according to the per-ecosystem ratio of bonafide inteins to total background sequences. Principal component analysis (PCA) of centered log-ratios of KEGG functional annotation proportions for ecosystems in the IMG/VR dataset. Each panel shows a different level of annotation, with decreasing granularity from A to C. Point shape indicates high level ecosystem (square: engineered, circle: environmental, diamond: host-associated). Point color represents the $\texttt{log}_2$-ratio of the proportion of bonafide inteins in that environment compared to total number of inteins in all environments and the proportion of proteins in that environment to total number of proteins (i.e., showing the over- or under-representation of inteins in each environment). Positive numbers (orange) represent more inteins that expected, whereas negative numbers (purple) represent fewer inteins than expected. (E.g., an environment with 50% of all bonafide inteins, but only 25% of total proteins in IMG/VR would have a Bonafide : Background ratio of $\texttt{log}_2(50/25) = 1$; that is, there were twice as many inteins as expected in that environment.) Point size represents the absolute number of bonafide inteins identified in that ecosystem.

decrease in intein enrichment aligns with the hypothesis that the cost of HGT may outweigh the evolutionary benefits, particularly in nutrient-poor or more static environments [24, 303]. Though one study did report high rates of HGT in oceanic marine environments [229], these results were based on the ability of marine microbes to incorporate gene transfer agents in a laboratory setting, rather than a comprehensive survey of HGT events in marine microbial communities. Even if high levels of HGT do exist, it is possible that the relative stability of the ocean over short timescales may make the rapid post-translational response mechanisms provided by inteins less necessary compared to other environments. In such cases, the potential downsides and costs of maintaining inteins may outweigh any benefits, potentially leading to their removal via genome streamlining, which is common among marine bacteria [113, 273].

Some of the intein distribution patterns are best explained by factors other than HGT rates and ecosystem stability. For example, the environment with the second-lowest ratio of ICP to the background was plant-associated microbial communities, specifically the "Plant; Other" category (Fig. 3.9). This category largely consists of samples from the phyllosphere (leaf surface), which is considered to be a hostile environment [140] and a hotbed for HGT [379], both between microbes [36, 146] and their host plant [290], and among bacteria, plants, and fungi [200]. Together, these factors would seem to indicate that the phyllosphere should be enriched for inteins. Interestingly, the phyllosphere is virtually free of archaea [100, 170], which carry more inteins than any other domain of life, with nearly half of archaea containing at least one intein [261]. Thus, the limited proliferation of inteins in the microbial community of the phyllosphere may be caused by lower numbers of microbes that tend to harbor higher numbers of inteins.

In addition to carrying high numbers of inteins, archaea are also known to have high levels of horizontal gene transfer [119]. Notably, archaea are abundant in many of the environments enriched for inteins, like floodplains [308], terrestrial and marine deep subsurface [351], and hydrothermal vents [78]. However, the relative abundance of archaea in an environment cannot fully explain the observed intein enrichment patterns.

For example, archaea constituted roughly 12% of the rice paddy rhizopshere [170], but made up less than 10% of several peatland communities [338]; however, in this study, both ecosystems were observed to be highly enriched for inteins. It should also be noted that this study examined viral proteins, meaning that the relationship between archaea, inteins, and the environment can only be examined indirectly, if at all.

The enrichment of inteins in certain environments may be driven simply by the presence of archaea–archaea carry more inteins than any other branch of life, and are known for thriving in extreme and highly variable environments. However, it is possible that archaea harbor so many inteins precisely because these are the types of environments that the rapid, post-translational response mechanism provided by inteins would be most beneficial. Future work is needed to examine the origin and distribution of inteins in archaea across environments to attempt to untangle this causality dilemma.

3.4 Conclusions

InteinFinder is an easy-to-use tool for the automatic identification and removal of inteins from peptide sequences. In this study, InteinFinder's speed and accuracy was demonstrated using IMG/VR, a large environmental dataset containing more than 100 million peptide sequences, a number that would make less automated intein identification infeasible. This allowed for a survey of inteins and their host proteins across a broad range of habitats, revealing that inteins are enriched in some environments and reduced in others, supporting recent sentiments that inteins are not simply selfish genetic elements. These enrichment patterns were unlikely to be artifacts and were in agreement current perspectives in horizontal gene transfer, a method by which inteins are commonly distributed among organisms. There is still much to be discovered about the environmental, taxonomic, and genomic distributions of inteins, as well as their ecological and evolutionary impacts on their host proteins and organisms. InteinFinder provides the potential for expanding the scope and depth of future studies to examine the many questions surrounding inteins.

PREFACE TO CHAPTER 4

The work presented in Chapter 4 of this dissertation has previously been published in PeerJ with myself as the first author [244]. The original author contribution list published in the manuscript is reproduced in full here:

- Ryan M. Moore and Sean M. McAllister conceived the project.
- Ryan M. Moore wrote the manuscript and implemented Iroki with assistance from Amelia O. Harrison.
- K. Eric Wommack and Shawn W. Polson guided the project and edited the manuscript.
- All authors read, edited, and approved the final manuscript.

Chapter 4

IROKI: AUTOMATIC CUSTOMIZATION AND VISUALIZATION OF PHYLOGENETIC TREES

4.1 Introduction

Community and population ecology studies often use phylogenetic trees as a means to assess the diversity and evolutionary history of organisms. In the case of microorganisms, declining sequencing cost has enabled researchers to gather ever-larger sequence datasets from unknown microbial populations within environmental samples. While large sequence datasets have begun to fill gaps in the evolutionary history of microbial groups [345, 248, 181, 183, 402], they have also posed new analytical problems, as extracting meaningful trends from high dimensional datasets can be challenging. In particular, scientific inferences made by visual inspection of phylogenetic trees can be simplified and enhanced by customizing various parts of the tree. Many solutions to this problem currently exist. Standalone tree visualization packages allowing manual or batch modification of trees are available (e.g., Archaeopteryx [129], Dendroscope [147], FigTree [302], TreeGraph2 [361], Treevolution [329]), but the process can be time consuming and error prone especially when dealing with trees containing many nodes. Some packages allow batch and programmatic customizations through the use of an application programming interface (API) or command line software (e.g., APE [270], Bio::Phylo [383], Bio.Phylo [369], ColorTree [57], ETE [145], GraPhlAn [17], JPhyloIO [362], phytools [310], treeman [30]). While these packages are powerful, they require substantial computing expertise, which can be an impediment for some scientists. Current web based tree viewers are convenient in that they do not require the installation of additional software and provide customization and management features

(e.g., Evolview [138], IcyTree [380], iTOL [194], PhyD3 [177], Phylemon [367], PhyloBot [132], Phylo.io [314]), but often have complex user interfaces or complicated file formats to enable complex annotations. Iroki strikes a balance between flexibility and usability by combining visualization of trees in a clean, user-friendly web interface with powerful automatic customization based on simple, tab-separated text (mapping) files. Given its focus on automatic customization and a core set of key features, Iroki's user interface can remain lean and easy-to-learn while still enabling complex customizations. In addition to specifying simple color gradients directly in the mapping file, Iroki also provides a dedicated module allowing the user to generate custom gradients to embed their data into color space, enhancing visualization. Iroki stays responsive even when customizing large trees, and it does not require an account or uploading potentially sensitive data to an external service.

Here, Iroki was used to customize large trees containing hundreds to thousands of leaf nodes according to extensive collections of metadata. These applications demonstrated the utility of Iroki for distilling biological and ecological insights from microbial community sequence data. The particular use cases included examinations of phage-host interactions, relative abundance of populations across sample types, and comparisons of viral community composition across environmental gradients.

4.2 Methods

Iroki is a web application for visualizing and automatically customizing taxonomic and phylogenetic trees with associated qualitative and quantitative metadata. Iroki is particularly well suited to projects in microbial ecology and those that deal with microbiome data, as these types of studies generally have rich sample-associated metadata and represent complex community structures. The Iroki web application and documentation are available at the following web address: https://www.iroki.net, or through the VIROME portal (http://virome.dbi.udel.edu) [396]. Iroki's source code is released under the MIT license and is available on GitHub: https://github.com/mooreryan/iroki.

4.2.1 Implementation

Iroki is built with the Ruby on Rails web application framework. The main features of Iroki are written entirely in JavaScript allowing all data processing to be done client-side. This provides the additional benefit of eliminating the need to transfer potentially private data to an online service.

Iroki consists of two main modules: the tree viewer, which also handles customization with tab-separated text files (mapping files), and the color gradient generator, which creates mapping files to use in the tree viewer based on quantitative data (such as counts) from a tab-separated text file similar to the classic-style OTU tables exported from a JSON or hdf5 format biom file [231].

4.2.2 Tree viewer

Iroki uses JavaScript and Scalable Vector Graphics (SVG), an XML markup language for representing vector graphics) to render trees. The Document Object Model (DOM) and SVG elements are manipulated with the `D3.js` library [41]. Rectangular, circular, and radial tree layouts are provided in the Iroki web application. Rectangular and circular layouts are generated using D3's cluster layout API (`d3.cluster`). For radial layouts, Algorithm 1 from [18] was implemented in JavaScript. In addition to the SVG tree viewer, Iroki also includes an HTML5 Canvas viewer with a reduced set of features capable of displaying huge trees with millions of leaf nodes (Supplementary Materials Sec. 4).

Iroki provides the option to automatically style aspects of the tree using a tab-separated text file (mapping file). Entries in the first column of this file are matched against all leaf labels in the tree using either exact or substring matching. If a leaf name matches a row in the mapping file, the styling options specified by the remaining columns are applied to that node. Inner nodes are styled to match their descendant nodes so that if all descendant nodes moving towards the inner parts of the tree have the same style, then quick identification of clades sharing the same metadata is possible.

Aspects of the tree that can be automatically styled using the mapping file include branches, leaf labels, leaf dots, bar charts, and arcs.

Inner node labels may represent support values (e.g., bootstrap results) or other comments that describe the inner nodes. If inner labels are numeric, then inner nodes can be decorated with filled and unfilled circles that allow quick identification of branches with high support. The semantics of support labels are key to proper tree representations [69]. As Iroki currently does not implement tree rerooting, Iroki handles these specifics implicitly rather than giving the option to map inner node labels to branches or to the nodes themselves.

While Iroki is focused mainly on automatic customization via mapping files, some interactive features are included such as node selection and the ability to modify labels after a tree has been submitted. Finally, various aspects of the tree can be adjusted directly through Iroki's user interface.

4.2.3 Color gradient generator

Iroki's color gradient generator accepts tab-separated text files (similar to the classic-style count tables exported by VIROME [396] or QIIME 1 [53]) and converts the numerical data (e.g., counts/abundances) into a color gradient. Several single-, two-, and multi-color gradients are provided including cubehelix [122] and those from ColorBrewer [44].

Iroki reads numerical data from tab-separated text files. Similar to the mapping file for the tree viewer, the first column should match leaf names in the tree, and the remaining columns describe whatever aspect of the data is of interest to the researcher (e.g., counts or abundance). In a dataset with M observations and N variables, the input file will then have $M + 1$ rows (the first row is the header) and $N + 1$ columns (the first column specifies observation names). From this data, Iroki can generate color gradients in a variety of ways.

4.2.3.1 Observation means

A color gradient is generated based on the mean value of each observation across all variables. In this case, an observation i would be represented as $\mu_i = \sum_{j=1}^{N} c_{ij}$, where c_{ij} is the value of observation (row) i for variable (column) j.

4.2.3.2 Observation evenness

A color gradient is generated based on the "evenness" of observation i across all N variables. Then, each observation i is represented by Pielou's evenness index [286] calculated across all variables: $E_i = H_i / H_{\max}$, where H_i is the Shannon entropy for observation i with respect to the N variables specified in the input file, and H_{max} is the maximum theoretical value of H_i. In this case, H_{max} occurs when observation i has equal values c_{ij} across all N variables. Thus, Pielou's evenness index for an observation i is calculated as

$$E_i = \frac{-\sum_{j=1}^{N} p_{ij} \log_2 p_{ij}}{\log_2 N},$$

where N is the number of variables and p_{ij} is the proportion of observation i in variable j (i.e., $c_{ij} / \sum_{j=1}^{N} c_{ij}$).

In this way, the user can map observations with high evenness (i.e., an observation with approximately the same value for each variable) to one side of the color gradient and observations with low evenness (i.e., an observation with high values in a few variables and low values in most others) to the other side of the gradient for easy identification.

4.2.3.3 Observation projection

Data reduction can be a powerful method for extracting meaningful trends in large, high-dimensional data sets. Given that microbiome or other studies in microbial ecology can have hundreds of samples and a rich set of metadata associated with those samples, data reduction often proves useful. Thus, Iroki provides a method to project the data into a single dimension and then map that projection onto a color gradient.

For data reduction, Iroki conducts a principal components analysis (PCA) calculated via the singular value decomposition (SVD) using the LALOLib scientific computing library for JavaScript [184]. Briefly, performing singular value decomposition on the centered (and optionally scaled) count matrix X, with observations as rows and variables as columns, the following decomposition is obtained: $X = USV^T$, where the columns of US are the principal component scores, S is the diagonal matrix of singular values, and the columns of V are the principal axes. To illustrate as much variance as possible in a single dimension, the first principal coordinate is mapped onto the chosen color gradient.

4.3 Results & Discussion

4.3.1 Bacteriophage proteomes, taxonomy, & host phyla

Viruses are the most abundant biological entities on Earth, providing an enormous reservoir of genetic diversity, driving evolution of their hosts, influencing composition of microbial communities, and affecting global biogeochemical cycles [365, 316]. Due to their importance, there is a growing interest in connecting viruses with their hosts through the analysis of metagenome data. As such, researchers have used a variety of computational techniques to predict viral-host interactions including CRISPR-spacer [321, 67, 256] and tRNA matches [27, 321, 67, 256], sequence homology [321, 67, 256], abundance correlation [67], and oligonucleotide profiles [322, 321, 247].

Iroki was used to examine phage-host interactions at the taxonomic scale by constructing a tree based on proteomic content [315] from a subset of viral genomes from the Virus-Host DB [238] using ViPTree [257] (Fig. 4.1; Supplementary Materials Sec. 1). A proteomic tree clusters phage based on relationships between the collection of protein-encoding genes encoded within their genomes [315, 255, 397]. Specifically, ViPTree bases its clustering on normalized tBLASTx scores between genomes following the method of [241].

Tree branches were colored by host phyla and virus family was indicated by a ring surrounding the tree using Iroki's bar plot options (Fig. 4.1; Supplementary

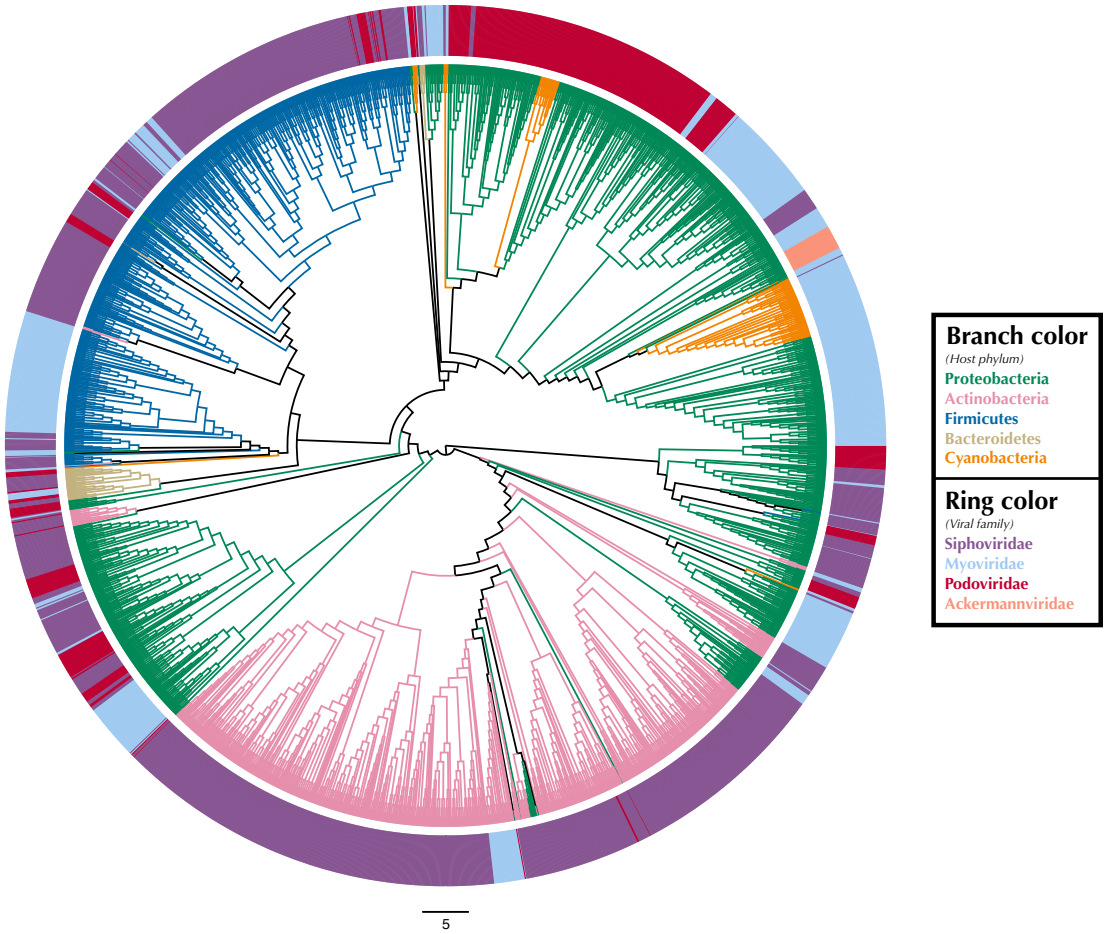

Figure 4.1: Proteomic cladogram of viruses from Virus-Host DB. Proteomic cladogram of viruses infecting Actinobacteria, Bacteroidetes, Cyanobacteria, Firmicutes, and Proteobacteria. Branches are colored by host phylum. Outer ring colors represent virus taxonomic family. Virus-host data is from the Virus-Host DB [238].

Materials Sec. 1). As shown by the branch coloring, host phyla mapped well onto the proteomic tree (i.e., large clusters of viruses that are similar in their proteomic content often infect the same host phylum). Firmicutes-infecting phage (represented by blue branches of the tree in Fig. 4.1) are confined almost exclusively to a large cluster in the top-left quadrant of the tree. This large cluster of mostly Firmicutes-infecting viruses can be further partitioned according to virus family, with a distinct group of myoviruses clustering separately from the other clades which include mostly siphoviruses. The Actinobacteriophage (pink) also cluster near each other with most viruses being confined to a few clusters at the bottom of the tree. The tight clustering of the Actinobacteriophage phage is likely explained by the fact that many of the viruses infect a limited number of hosts including Propionibacterium and *Mycobacterium smegmatis* from the SEA-PHAGES program (`https://seaphages.org`) [291]. In contrast, the Proteobacteria-infecting viruses (green) are clustered in a few locations across the tree, with each cluster showing high levels of local proteomic similarity.

Homology and similarity-based methods have previously been shown to be effective in predicting a phage's host [83], perhaps because viruses that infect similar hosts are likely to have more similar genomes [382]. Given this and the fact that the proteomic tree clusters viruses based on shared sequence content using homology and multiple sequence alignments [315], it is unsurprising that viruses infecting hosts from the same phylum often cluster near each other on the proteomic tree. In fact, previous studies have used proteomic distance [256] and other measures of genomic similarity [382] to transfer host annotations from viruses with known hosts to metagenome assembled viral genomes with unknown hosts. In contrast, virus taxonomy is primarily based on multiple phenotypic criteria including virion morphology, host range, and pathogenicity, rather than on genome sequence similarity [346, 347]. One study found that for prokaryotic viruses, members of the same taxonomic family (as defined by phenotypic criteria) were divergent and often not detectably homologous in genomic analysis [6]. In particular, multiple viral families in the order Caudovirales were interspersed in their dendrograms. Similar results can be seen in Fig. 4.1, in which several

Caudovirales viral families are intermixed in clusters throughout the tree.

4.3.2 Bacterial community diversity & prevalence of *E. coli* in beef cattle

Shiga toxin-producing *Escherichia coli* (STEC) are dangerous human pathogens that colonize the lower gastrointestinal (GI) tracts of cattle and other ruminants. STEC-contaminated beef and STEC cells shed in the feces of these animals are major sources of foodborne illness [130, 54]. To identify possible interactions between STEC populations and the commensal cattle microbiome, a recent study examined the diversity of the bacterial community associated with beef cattle hide [61]. Hide samples were collected over twelve weeks and SSU rRNA amplicon libraries were constructed and sequenced on the Illumina MiSeq platform [91]. The study found that the structure of hide bacterial communities differed between STEC-positive and STEC-negative samples.

To illustrate Iroki's utility for exploring changes in the relative abundance of taxa in conjunction with metadata categories, a subset of cattle hide bacterial operational taxonomic units (OTUs) were selected from the aforementioned study (Supplementary Materials Sec. 2). A Mann-Whitney U test comparing OTU abundance between STEC-positive and STEC-negative samples was performed. Cluster representative sequences from any OTU with a p-value < 0.2 (selected to limit the number of OTUs on the tree and to demonstrate Iroki's features by coloring branches based on test significance) from the Mann-Whitney U test were selected and aligned against SILVA's non-redundant, small subunit ribosomal RNA reference database (SILVA Ref NR) [296] and an approximate-maximum likelihood tree inferred using SILVA's online Alignment, Classification and Tree (ACT) service (`https://www.arb-silva.de/aligner/`) [295]. Iroki was then used to display various aspects of the data set (Fig. 4.2; Supplementary Materials Sec. 2). Branches of the tree were colored based on the p-value of the Mann Whitney U test examining change in relative abundance with STEC contamination (dark green: $p \leq 0.05$, light green: $0.05 < p \leq 0.10$, and gray: $p > 0.10$). Additionally, bar charts representing the log of relative abundance of each OTU (inner bars) and

the abundance ratio (outer bars) of OTUs in samples positive and negative for STEC are shown. The color gradient for the inner bar series was generated using Iroki's color gradient generator. Finally, leaf labels show the order and family of the OTU and are colored by predicted OTU phylum using one of the color palettes included in Iroki.

Decorating the tree in this way allows the user to explore the data and look for high-level trends. For example, Firmicutes dominates the tree (e.g., Bacillales, Lactobacillales, Clostridiales). Members of Clostridiales are at low-to-medium relative abundance compared to other OTUs on the tree. Some Clostridiales OTUs (e.g., a majority of the Ruminococcaceae) tend to be at higher abundance in STEC-positive samples, whereas other Clostridiales OTUs, namely those classified as Lachnospiraceae, tend to be at lower abundance in STEC-positive samples. Previous studies have also identified significant positive associations between STEC shedding and Clostridiales OTU abundance in general [419] and Ruminococcus OTUs abundance more specifically [412]. In contrast, other studies have found certain Ruminococcus OTUs associated with shedding cattle and other Ruminococcus OTUs associated with non-shedding individuals [404]. Apparent contradictions may be explained by the fact that the various studies were examining the bacterial microbiome associated with different locations on the cow (e.g., GI tract, recto-anal junction, hide). In fact, significant spatial heterogeneity in community composition exists even among different sites along the gastrointestinal tract [220]. Other potential explanations include methodological differences, or that variation associated with STEC presence may be better explained by using more granular groupings than taxa and OTUs (e.g., amplicon sequence variants) [46].

In this dataset, more of the OTUs had a higher average relative abundance (brown bars) in STEC-negative samples than in STEC-positive samples (blue bars). Similarly, in a study of the upper and lower gastrointestinal tract microbiome of cattle, a majority of differentially abundant OTUs were found to be at higher abundance in animals that were not shedding *E. coli* O157:H7 [412]. In contrast, another study found that over 75% of differentially expressed OTUs were at greater abundance in STEC shedding cattle [404].

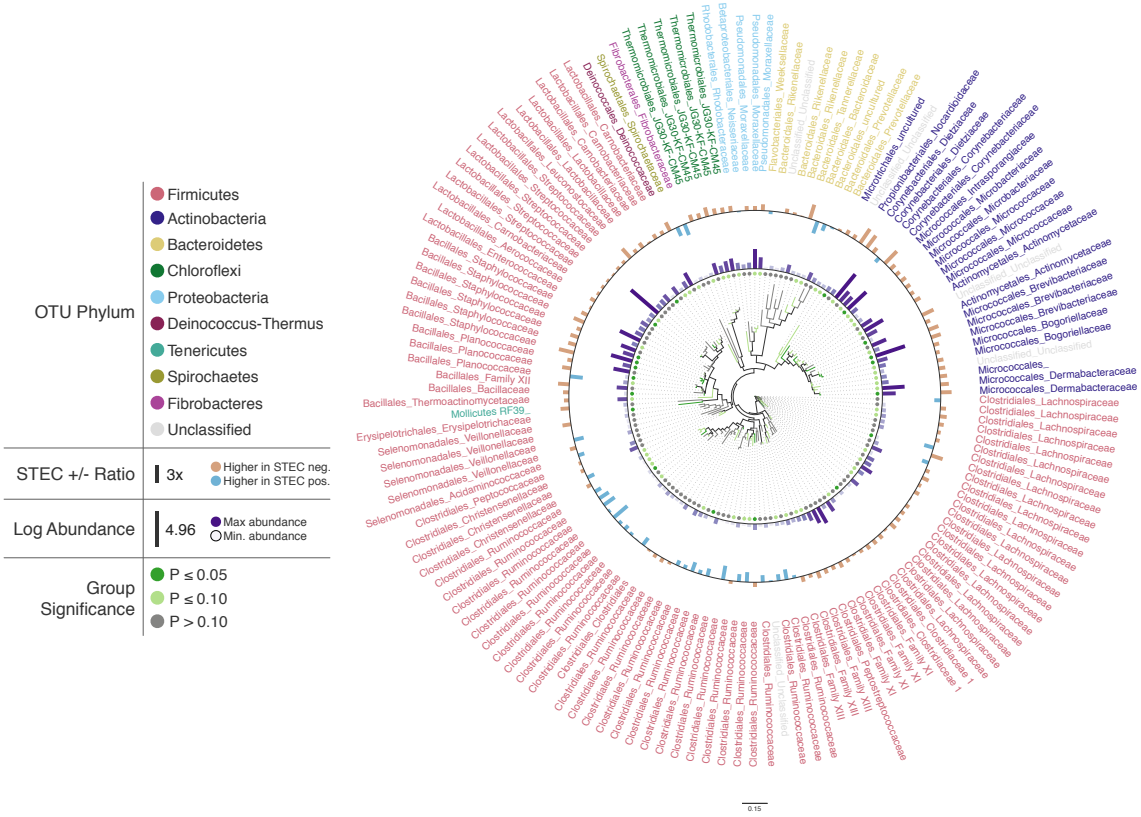

Figure 4.2: Changes in OTU abundance in two sample groups. Approximate-maximum likelihood tree of hide SSU rRNA OTUs that showed differences in relative abundance between STEC positive and STEC negative cattle hide samples. Branch and leaf dot coloring represents the p-value of a Mann-Whitney U test (dark green: $p \leq 0.05$, light green: $0.05 < p \leq 0.1$, gray: $p > 0.1$) testing for changes in OTU abundance between STEC-positive samples and STEC-negative samples. Inner bar heights represent log transformed OTU abundance, and outer bars represent the abundance ratio between STEC-positive and STEC-negative samples (blue bars for higher abundance in STEC positive samples and brown bars for OTUs with higher abundance in STEC negative samples). Taxa labels show the predicted Order and Family of the OTU and are colored by the predicted phylum using the Paul Tol Muted color palette included with Iroki.

4.3.3 *Tara* Oceans viromes

The ribonucleotide reductase (RNR) gene is common within viral genomes [81] and RNR polymorphism is predictive of certain biological and ecological features of viral populations [326, 136]. As such, it can be used as a marker gene for the study of viral communities. To explore viral communities of the global ocean, RNR proteins were collected from the *Tara* Oceans viral metagenomes (viromes). The *Tara* Oceans expedition was a two-and-a-half year survey that sampled over 200 stations across the world's oceans [39, 284]. Forty-four viromes were searched for RNRs (Supplementary Materials Sec. 3). Of these, three samples contained fewer than 50 RNRs and were not used in the subsequent analysis. In total, 5,470 RNR sequences across 41 samples were aligned with MAFFT [165] and post-processed manually to ensure optimal alignment quality. Then, FastTree [293] was used to infer a phylogeny from the alignment. Using this tree, the unweighted UniFrac distance [210] between samples was calculated using QIIME [53]. A tree was generated from this distance matrix in R using average-linkage hierarchical clustering. Additionally, Mantel tests identified that conductivity, oxygen, and latitude were significantly correlated ($p < 0.05$) with the UniFrac distance between samples (Supplementary Materials Sec. 3). Finally, Iroki was used to generate color gradients and add bar charts to visualize the data (Fig. 4.3). Coloring of the dendrogram with the Viridis color palette (a dark blue, teal, green, yellow sequential color scheme) was based on a 1-dimensional projection of sample conductivity, oxygen, and latitude calculated using Iroki's color gradient generator. The color gradient generator was also used to make the color palettes used for the bar charts.

Coloring the dendrogram based on a projection of the environmental conditions of the samples results in samples with similar environmental metadata being similar in color. For example, the station 66 surface and deep chlorophyll maximum (DCM) samples are nearly identical to one another with respect to conductivity, oxygen, and latitude and have the same dark bluish branch color. In contrast, surface samples from stations 31 and 32 both have a lighter yellowish-green branch color. As the bar charts indicate, these two samples are very similar to one another with respect to the

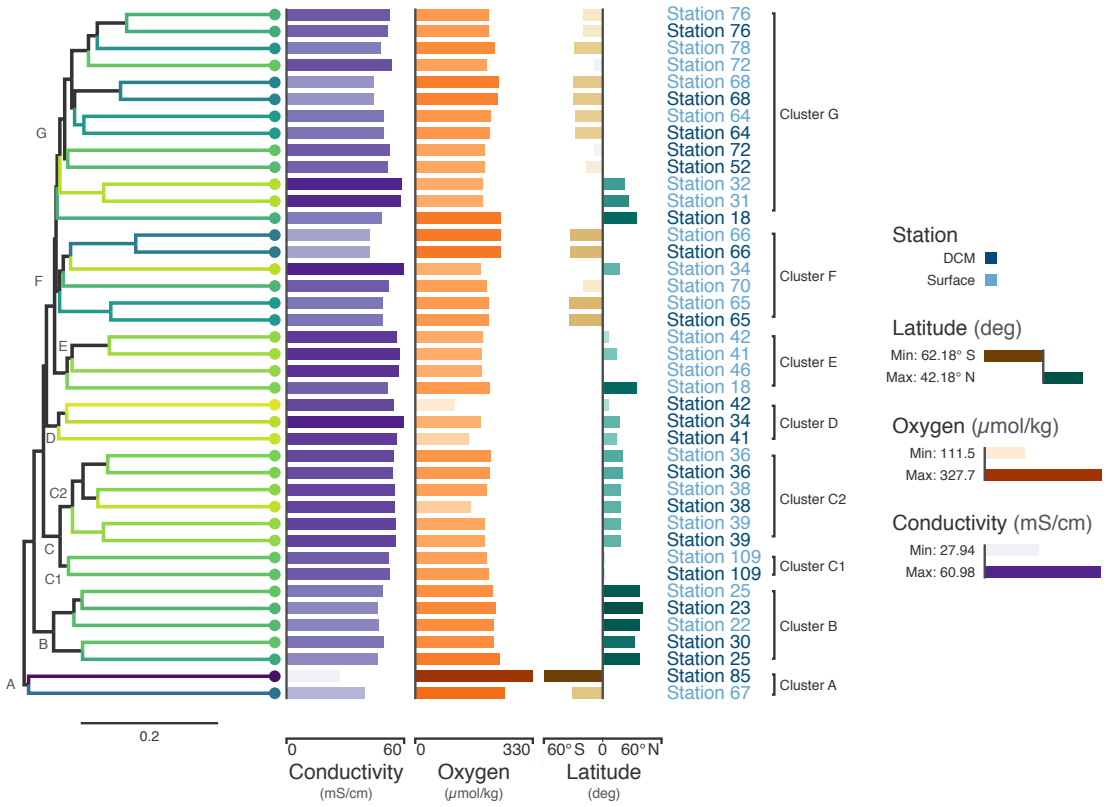

Figure 4.3: *Tara* Oceans virome similarity with associated metadata. Average-linkage hierarchical clustering of sample UniFrac distance based on RNR sequences mined from 41 *Tara* Oceans viromes. Major and sub-clusters of samples (A-G) are labeled. Branch color is based on a scaled, 1-dimensional projection of sample conductivity, oxygen, and latitude onto the cubehelix color gradient. Samples that are more similar to each other in branch color represent those that are more similar to each other with respect to the environmental parameters in the ordination. The first bar series (purple) represents sample conductivity (mS/cm), the second bar series (orange) represents sample dissolved oxygen levels (µmol/kg) and the third bar series (brown/green) represents sample latitude (degrees). For the first two bar series, shorter bars with lighter colors indicate lower values, while longer bars with darker colors indicate higher values. For the third series, longer, dark brown bars indicate samples with extreme negative latitudes, whereas longer, dark blue bars indicate samples with extreme positive latitudes. Samples with intermediate latitudes are represented by shorter, light colored bars. Sample labels represent the station from which the virome was acquired and are colored by sampling depth, with light blue representing surface samples and dark blue representing samples from the deep chlorophyll maximum at that station.

metadata (hence their similar coloring), but are rather different from the station 66 samples in branch color, reflecting the differences in metadata between the two groups.

The combination of dendrogram coloring and bar charts assists in finding trends in the data. Since the dendrogram is based on UniFrac distance between samples based on RNR OTUs, samples that cluster together on the tree have more similar viral communities, according to RNR gene allele content, than samples that are far from one another. In contrast, dendrogram branch coloring and the bar charts show environmental information about the samples themselves (conductivity, oxygen, and latitude). Combining these two aspects of the samples enables visualization of the relationship between the similarity of RNR-containing viral communities and the environments in which they are found.

For example, the samples in the bottom half of the tree are, in general, from northern latitudes, whereas samples towards the top tend to be from southern latitudes. In a previous study of the T4-like viral communities of Polar freshwater lakes, no significant correlation between latitude and viral community diversity was found in the Antarctic samples [71]. Though the Arctic lakes were not tested among themselves for significant associations between latitude and viral community richness (presumably due to the small latitudinal variation in Arctic sampling locations), Arctic and Antarctic lakes were tested against one another; however, no significant difference in viral diversity was seen with respect to pole of origin. The Antarctic samples from the study ranged from 67.84 degrees S to 62.64 degrees S, whereas the *Tara* Oceans viromes used to build the tree in Fig. 4.3 ranged from 62.18 degrees S to 41.18 degrees N. The increased range of samples from the *Tara* survey may have enabled this shift in diversity to be detected. Additionally, the previous study used *g23*, the gene for major capsid protein, to survey the viral community. It is possible that a functional protein like RNR is more connected with environmental conditions than a structural protein such as the T4-like major capsid protein. RNRs reduce ribonucleotides, the rate-limiting step of DNA synthesis [175, 4]. There are several different types of RNR,

each with specific biochemical mechanisms and nutrient requirements [260]. Accordingly, the type of RNR carried by a cell or virus often reflects the environmental conditions in which DNA replication occurs [306, 65, 326, 355, 136]. A survey based on RNR, then, may provide more sensitivity in detecting environmental effects on viral community structure. A significant relationship between T4-like viral communities and bacterial assemblages was found however [71], and numerous other studies have reported a significant relationship between bacterial community diversity and latitude (e.g., [180, 301]), latitudinal variation in bacterial communities is likely linked to viral community variation.

Certain clusters have been marked on the tree for further analysis. Cluster A (Station 85 DCM, Station 67 surface) contains the samples with the most divergent RNR-containing viral populations (Fig. 4.3) according to the dendrogram. Station 85 DCM is also the sample with the lowest conductivity, highest dissolved oxygen, and most southerly latitude, suggesting that the divergent conditions of the sample with respect to the other included samples could be influencing the divergent RNR-containing viral population. Clusters B and C also offer a good point of comparison (Fig. 4.3). In addition to the similarity of their RNR-containing viral populations, samples in cluster B have highly similar conductivity, oxygen, and latitude (as shown by their highly similar branch color and bar charts), suggesting a close connection between sample composition and viral population. Cluster C is separate from cluster B on the dendrogram, implying their RNR-containing viral populations are less similar. The sample metadata between the two clusters is less similar as well, with Cluster B having on average a lower conductivity and higher dissolved oxygen content than samples from cluster C.

Connections between viral community composition and environment have been seen before. Salinity, which can be estimated from measurements of electrical conductivity [278, 279], has been shown to affect viral-host interactions. In a viral-host system of halovirus SNJ1 with its host, *Natrinema* sp. J7-2, viral adsorption rates and lytic/lysogenic rates were measured at varying salt concentrations. Adsorption and

lytic rate were found to increase with salt concentration, whereas the lysogenic rate decreased [237]. In a system of tropical coastal lagoons, salinity was found to be one of the main factors positively affecting viral abundance [158]. Viral community structure has also been associated with shifts in salinity in various environments [33, 87, 394, 99]. These shifts likely effect a change in the host communities, which is reflected in the shifts in viral communities.

Cluster C can be further divided into two clusters, C1 and C2. While the samples in C1 are closer to those in C2 than to those in cluster B in terms of their RNR-carrying viral populations, the samples in C1 are more similar to the samples in cluster B with respect to their metadata projection. The similar branch coloring between samples in clusters B and C1, despite their large differences in latitude, occurs because more of the variation in the first principal component (the principal component on which the Viridis coloring is based) is explained by conductivity and oxygen than by latitude (Fig. 4.4; full ordination: Supplementary Figure S1). More striking examples can be found elsewhere in the tree. For example, station 66 surface, station 66 DCM, and station 34 surface cluster together on the dendrogram based on viral community similarity (cluster F), but the conductivity, oxygen, and latitude values for sample 34 surface are quite different from the station 66 samples. Thus, while these three metadata categories were significantly correlated with sample UniFrac distance, other factors also play a role in shaping the viral communities. Overall, using Iroki to add color and bar charts based on environmental metadata to the dendrogram based on RNR-carrying viral community structure helps visualize that high-level viral community structure can be influenced by the environmental parameters of the sample from which they originate.

4.4 Conclusions

Iroki is a web application for fast, automatic customization and visualization of large phylogenetic trees based on user specified, tab-delimited configuration files with categorical and numeric metadata. Through the use of simple configuration files, Iroki provides a convenient way to rapidly visualize and customize trees, especially in cases

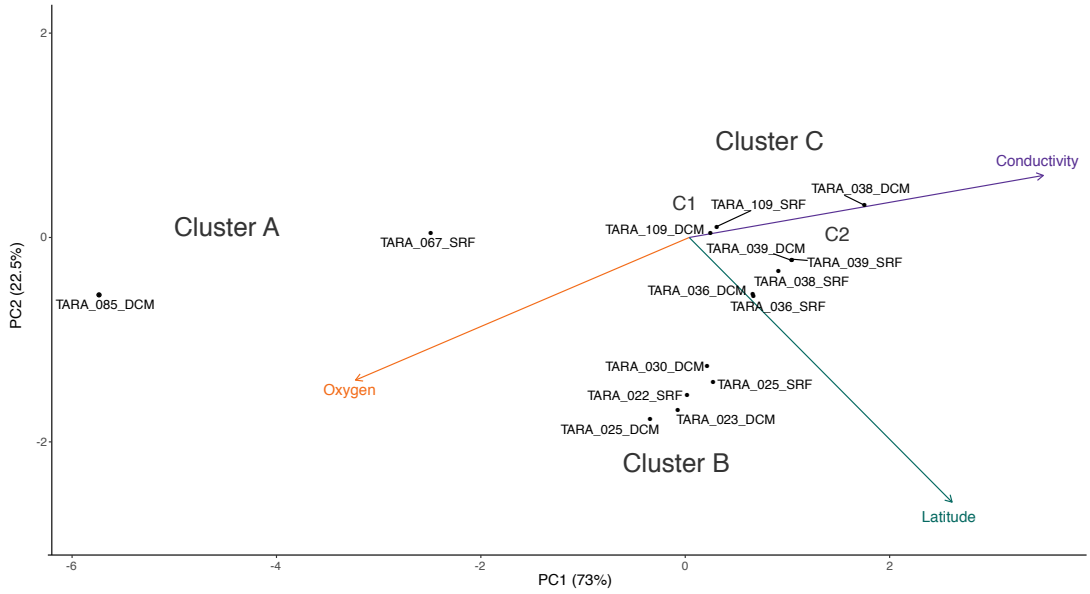

Figure 4.4: PCA biplot of *Tara* Oceans virome clusters A, B, and C. Principal components analysis biplot of *Tara* Oceans viromes based on sample oxygen, conductivity, and latitude. Ordination was done on all viromes, but only those from clusters A, B, and C are shown here for clarity.

where the tree in question is too large to annotate manually or in studies with many trees to annotate. While Iroki includes many key features, future work is planned to increase its utility. There is no mechanism within Iroki to handle rerooting trees. As such, users must use an external program to reroot their tree before viewing it in Iroki. Customizing the tree is mainly handled by modifying the mapping file, however, Iroki could be made more interactive by allowing the user to edit certain aspects of the tree "by-hand" without having to reupload a new mapping file. Currently, Iroki allows editing leaf labels after a tree is submitted. More interactive features, such as editing label and branch styles, are planned for a future release. Finally, bringing the full feature set of Iroki's SVG based viewer to the Canvas viewer will allow users to visualize and customize huge trees quickly and easily.

Various datasets from microbial ecology studies were analyzed to demonstrate Iroki's utility. Iroki simplified the processes of data exploration and presentation by facilitating the mapping of various aspects of the data directly on the tree. Though these examples focused specifically on applications in microbial ecology, Iroki is applicable to any problem space with hierarchical data that can be represented in the Newick tree format.

4.5 Additional information and declarations

4.5.1 Availability of data and materials

Data used to generate figures for this manuscript are available for download on Zenodo at the following URL: https://doi.org/10.5281/zenodo.3458510.

4.5.2 Funding

This project was supported by the Agriculture and Food Research Initiative grant no. 2012-68003-30155 from the USDA National Institute of Food and Agriculture, the National Science Foundation Advances in Biological Informatics program (award number DBI_1356374), the National Science Foundation Grant No. 1736030, the Established Program to Stimulate Competitive Research (award number OIA_1736030) from

the Office of Integrated Activities, and a Doctoral Fellowship provided by University of Delaware in conjunction with the Unidel Foundation. Computational infrastructure support by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility was made possible through funding from the Delaware Biotechnology Institute, and the Delaware INBRE program with a grant from the National Institute of General Medical Sciences (NIGMS P20 GM103446) from the National Institutes of Health and the State of Delaware.

4.5.3 Acknowledgments

The authors would like to acknowledge Barbra D. Ferrell for manuscript editing, and the reviewers for their constructive feedback. This content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Chapter 5

A COMPOSITIONAL DIVERSITY FRAMEWORK WITH APPLICATIONS TO CATTLE MICROBIOME

5.1 Introduction

Robust measures of diversity of microbial communities must account for important properties including the compositional nature of next-generation sequencing (NGS) data, the sparsity of count tables generated from NGS sequencing, and other biases across the sample-to-sequence-to-discovery pipeline.

5.1.1 Common issues in microbial diversity analysis

5.1.1.1 NGS data are compositional

The sequencing data commonly used to analyze microbial communities is inherently compositional [114]. Features (taxa, operational taxonomic units, amplicon sequence variants, etc.) are subject to the constant sum constraint induced by the sequencing procedure itself [297]. There is a finite number of sequencing reads, resulting in a finite number of possible observations per sequencing run. The "counts" of reads, amplicons, contigs, etc. are not really counts at all–they represent proportions of the total sequencing effort. Consequently, differences in counts among samples, treatments, environments, or other metadata groupings do not necessarily indicate changes in biology or ecology, but rather reflect variations in sampling effort, sequencing efficiency, and other factors.

Ignoring the compositional nature of the data can result in critical issues during subsequent statistical analysis and interpretation, potentially leading to erroneous conclusions [114, 298]. Because of the constant sum constraint, an increase in the abundance of one feature necessarily requires a corresponding decrease in the abundance

of another feature. This introduces a negative bias among all features, violating the assumption of feature independence in many statistical methods.

Because common bioinformatic tools, pipelines, and statistical analyses generally treat data as real numbers within Euclidean space rather than as proportions constrained to the simplex, novel specialized methods and techniques are required to analyze NGS data. Although originally developed for other fields such as geology, compositional data analysis (CoDA) approaches can be applied to microbial community diversity analyses as well. CoDA methods typically employ the log-ratio transformation on the data, enabling the use of traditional statistical models downstream [7, 8].

The additive log-ratio (ALR) is a simple transformation in which a single reference feature is selected, and all other features are treated proportionally to that reference. Assuming there are Q features, and given a reference feature D selected from the Q features, the additive log-ratio transformation is defined as

$$\texttt{alr}(x_q) = \log\left(\frac{x_q}{x_D}\right), q = 1, \ldots, D-1, D+1, \ldots, Q \tag{5.1}$$

Another common transformation is the centered log-ratio (CLR) transformation. Let $g(\mathbf{x})$ be the geometric mean of all features in the given sample, then the CLR is defined as

$$\texttt{clr}(x_q) = \log\left(\frac{x_q}{g(\mathbf{x})}\right), q = 1, \ldots, Q \tag{5.2}$$

Other more complex transformations, such as the isometric log-ratio (ILR) are also used when appropriate [312]. Though each of the transformations address the fundamental issues of compositional data, each comes with specific considerations in terms of their implementation and interpretation. For example, it has been argued that the ALR is sub-optimal due to its lack of isometry[1]; however, others have argued that the simplicity in its interpretation, and the fact that it is a nearly isometric

[1] An isometry is a distance preserving transformation from one mathematical space to another

transformation in the context of typical microbiome experiments, overcomes any lack of mathematical purity [123].

In addition to the manual application of log-ratio transformations that form the basis of CoDA methods, there are some methods specifically designed to handle microbiome data in a compositional framework, including ANCOM [201], ALDEx2 [94, 95], propR [299], balances [313], and DEICODE [228]. These have made CoDA methods more accessible to more researchers by providing alternatives for more traditional bioinformatic tools, e.g., ALDEx2 and ANCOM can be utilized for differential abundance analysis, replacing non-CoDA tools such as DESeq2 [114].

5.1.1.2 Microbial communities are sparse

Microbial communities are compositionally complex, generally characterized by a small number of highly abundant features, and an extremely long tail of type- and trait-level diversity [350]. This long tail of microbial life, often referred to as the rare biosphere, encompasses an enormous range of taxonomic, phylogenetic, and functional diversity [86, 215]. It includes a diverse set of microbes from all domains of life, and is hypothesized to act as a bank of microbial diversity–an ancient and vast source of genetic and functional diversity existing at a low abundance until conditions are favorable allowing them to thrive and drastically increase in abundance [19, 350]. Thus, members of the long tail are likely critical to the long term maintenance and functional resilience of microbial communities [350, 339, 331].

The sparsity of microbial communities poses challenges for estimating their diversity, and must be accounted for. Study design considerations such as increasing sampling effort, including biological replicates, and potentially combining untargeted shotgun approaches with approaches that specifically target rare members, should be considered and employed when appropriate to help mitigate the issue [350, 188]. Even in well-designed studies, the sequenced microbial community will likely yield a highly sparse count tables [276] that requires careful bioinformatic processing to effectively handle [267]. This sparsity necessitates special care even when using CoDA methods,

which require a zero-replacement step. Zero-replacement can cause under-sampled communities to appear similar to each other purely due to similar levels of sparsity, rather than reflecting actual biological or ecological similarity, and should be considered when interpreting results [211].

5.1.1.3 Measuring microbial community is subject to various biases

Microbial communities can be complex and highly variable, even among similar sampling sites, making it challenging to obtaining accurate measurements of their composition. The measurement of microbial community diversity is subject to random and systemic errors, as well as biases introduced at every stage of the sample-to-sequence-to-discovery pipeline [234, 254]. Sampling effects, artifacts, and biases can arise due to the spatial heterogeneity and variability of microbial communities, especially when under-sampling occurs [16, 188]. Additionally, biases are introduced during DNA extraction and amplification that affect DNA yield and cause preferential extraction, which can cause under- and over-representation of certain taxonomic groups, and generation of chimeric sequences [223, 269, 64, 35]. Further, bioinformatic and computational methods can introduce bias into downstream results [254, 111]. This can occur through the use of statistical tests and models that are not appropriate for the type of data being analyzed [114], reliance on incomplete or error-filled reference databases [233], software bugs [70], and other factors. Additionally, certain bioinformatic methods designed to mitigate these issues, such as rarefaction and normalization, can introduce biases and alter the community structure [235, 392]. As these biases have a measurable effect on the diversity estimates of microbial communities [64, 35, 343], it is crucial to address them.

5.1.2 Addressing these issues

To effectively address these problems, it is important to consider the multiple levels at which they exist [188]. Improving experimental design and sampling procedures can help minimize the effects of random errors by addressing small sample sizes

and lack of replication. Optimizing DNA extraction and amplification protocols can address primer bias and DNA extraction inefficiencies. Proper quality control in bioinformatics including removal of low-quality reads, chimeras, and contaminant sequences, utilizing multiple bioinformatics tools for comparative analyses, and employing CoDA methods all contribute to a robust discovery pipeline [114, 34].

However, there are still many outstanding problems related to estimating the diversity of microbial communities. Due to biases and the inherent complexities of microbial communities, diversity estimates are noisy, and typically used plug-in point estimates of diversity[2] do not account for the specific challenges presented by microbial communities [393]. Accurate estimation of variability and variance in the estimates themselves is critical in assessing the meaning and significance of measured levels of diversity.

Another technical challenge that must be addressed is the presence of zero counts, which are highly prevalent in microbial communities due to their sparsity. In particular, CoDA methods, which involve log-ratio calculations, cannot handle data with zero counts, and they must be replaced [7, 266, 297]. At the high levels of sparsity common in microbial communities, the zero-replacement procedure used in CoDA methods can lead to detectable distortions in the inferred community structure and diversity estimates [211].

In addition to the technical challenges, philosophical issues regarding what aspects of diversity should be measured must also be considered. A comprehensive understanding of community diversity requires consideration of both the abundant and the rare microbes that make up the long-tail of the community [274]. Furthermore, there is an ongoing debate about the relative importance and utility of type-level versus trait-level diversity in understanding microbial community structure [120].

By approaching these questions in a principled and unified manner, it is possible

[2] E.g., the Shannon index and Simpson index are commonly used plug-in estimates of alpha diversity.

to generate interesting hypotheses about community ecology and draw robust conclusions even when using summary measures such as α- and β-diversity. To this point, the remainder of the chapter addresses the following issues: (1) accurate estimation of variability and variance in diversity estimates, (2) incorporating measures of both type- and trait-level diversity, (3) considering both abundant microbes and members of the "rare biosphere" in diversity estimates, and (4) handling high levels of sparsity in microbial datasets.

5.2 A framework for measuring diversity of microbial communities

The proposed framework for exploring microbial community diversity integrates a state-of-the-art compositional method for estimating community structure, diversity indices and their variance [393], and similarity aware diversity measures [304] parameterized by an abundance viewpoint parameter (e.g., [375]) and a newly introduced similarity viewpoint parameter, combined with careful selection of protein-coding gene markers to survey communities.

5.2.1 Background

Microbial communities are highly diverse and often highly uneven, with few taxa numerically dominating the community and many low abundance and frequently unobserved taxa filling out the long tail of diversity. Common tasks in microbiome studies include linking taxon abundance with biological, ecological, or clinical data, detecting correlation between taxa, and metabolic pathway analysis These tasks are challenging for standard analytic methods due to specific aspects of the data generated by the sequencing procedure. NGS surveys of microbial communities generally generate data in the form of a high-dimensional count table with supporting covariate information describing the biotic or abiotic conditions in which the communities were observed [197]. In addition to the high dimensionality, the data vectors are sparse and compositional, i.e., subject to the constant sum constraint. Thus, statistical methods that explicitly take these criteria into account are required.

A common approach for analyzing high-dimensional ecological community composition data is diversity analysis. Consider a community of C features present in relative abundances $\boldsymbol{z} = (z_1, \ldots, z_C)$. In a typical microbiome study, C will be on the order of thousands to hundreds of thousands. An α-diversity index is a function that summarizes the relative abundances, $\boldsymbol{z}$, of a single sample $f : \mathbb{S}^{C-1} \to \mathbb{R}$, where $\mathbb{S}^d$ is the d-dimensional simplex.[3] A β-diversity index is a function that summarizes community composition information from two communities: $g : \mathbb{S}^{C-1} \times \mathbb{S}^{C-1} \to \mathbb{R}$. α-diversity indices are within-community structure summaries, whereas β-diversity indices are between-community structure summaries.

5.2.1.1 α-diversity

Many different measures of α-diversity exist in the literature, each of them emphasizing different features of the community. While analyses focusing on α-diversity are ubiquitous in microbial ecology, their statistical formalization remains an area of active research. Diversity is commonly estimated using statistical methods that assume observed counts are drawn from a multinomial distribution with an unknown probability vector $\boldsymbol{z}$ (e.g., [418, 142, 50]). Additionally, most estimates of community α-diversity are a function of the abundance vector of a sample in isolation, and do not utilize information from the full community or the measured covariates that describe experimental conditions. However, some methods use more flexible and appropriate models from the compositional data analysis literature, including those that enable the modeling of taxa co-occurrence [74, 15, 309, 51, 393].[4]

[3] $\mathbb{S}^{C-1}$ is the simplex for a community with C features.

[4] De'ath [74] introduces a multinomial generalized linear model that extends the logistic regression model from two to two or more response categories and can link Shannon diversity to environmental, spatial, and temporal predictors. Arbel et al. [15] use a nonparametric Bayesian model that uses the structure of the full count table as well as covariate information; it was specifically designed to deal with estimating community response to environmental variables. Unfortunately, this method is computationally prohibitive and the original model focuses on a single covariate, though the authors claim that the process could be extended to multiple covariates. Ren et al. [309] use

5.2.1.2 β-diversity

Similar to α-diversity, a large number of β-diversity measures also exist, each highlighting different aspects of shared community structure [185]. Unlike with α-diversity, there is little work on statistical estimation of β-diversity indices, rather β-diversity estimates are generally performed with the plug-in estimates only [393]. At a high-level, small values of β-diversity correspond with compositionally similar communities, or communities that share many features, whereas large values of β-diversity indicate communities that are more compositionally dissimilar, or those that share few features.

5.2.2 DivNet model overview

DivNet is a recent method that utilizes a compositional framework to model and estimate microbial community structure, diversity indices, and their variance [393]. To improve the accuracy of diversity estimates, it leverages the networked structure of microbial communities and aggregates information across samples. Finally, the DivNet model explicitly accounts for the compositional nature of sequencing data, which is critical for accurate statistical inferences [115, 114].

The network structure of microbial communities can have a marked impact on diversity estimates. An ecological network describes the patterns of co-occurrence seen in microbial communities, including competition for resources, predator-prey dynamics, symbiotic cooperation, and viral-host interactions. These patterns of interaction are a hallmark of ecological communities and are repeated across different environmental settings [92, 96]. Previous methods to estimate ecological networks exist, including

---

a nonparametric Bayesian model for W given Z, though this method cannot handle continuous covariates. Cao et al. [51] uses the full observed count (abundance) matrix W to obtain a low-rank estimate of the true (and unobserved) relative abundances Z, though no publicly available software implements this method. The DivNet method of Willis et al. [393] will be described in detail later in this chapter.

SparCC [105] and SPIEC-EASI [179]. A key aspect of the DivNet model is accounting for these network effects when estimating microbial community composition and diversity.

While DivNet provides improved performance of diversity estimates, it is not practical for large microbiome datasets, particularly for research groups with limited access to high-performance computing environments. To overcome this limitation, DivNet was reengineered and reimplemented with a focus on performance and parallelization in Rust, a compiled, statically-typed, non-garbage collected language well suited to high-performance numerical programs. Additionally, this work leverages DivNet's community composition estimates to calculate similarity-aware measures of diversity, a task not addressed in the original work.

The Rust implementation, divnet-rs, introduced in this manuscript and publicly available on GitHub[5], uses the same model of community composition and parameter estimation methods as the original, though with an emphasis on improved runtime and memory efficiency. Therefore, this section provides only a high-level overview of the model and parameter estimation procedure. For a detailed explanation of the DivNet model and parameter estimation procedure, readers are referred to the original manuscript [393].

5.2.2.1 Compositional data models

The standard model for compositional data is the multinomial distribution, which implies a negative covariance between counts of different features. An alternative approach suggested by Aitchison [7] is to model the data in a way that explicitly accounts for the constant-sum constraint and other aspects unique to compositional data: the log-ratio model. This models the count matrix W as independent draws from a multinomial distribution, where the true composition matrix Z of all communities in question is unknown:

[5] https://github.com/mooreryan/divnet-rs

$$p(W|Z) \propto \prod_{i=1}^{N} \prod_{q=1}^{Q} Z_{iq}^{W_{iq}} \tag{5.3}$$

where N is the number of samples, Q is the number of features, $W \in \mathbb{N}^{N \times Q}$ is the N-sample by Q-feature observed count matrix, W_{iq} is the observed count of feature q in sample i, $Z \in \mathbb{R}^{N \times Q}$ is the latent random variable that gives the underlying true composition of the communities, and Z_{iq} is the unobserved, true proportion of feature q in sample i. Note that rows of Z are the relative abundances or proportions of the features in each sample, and so $\sum_{q=1}^{Q} Z_{iq} = 1$ for each sample i. Then, the additive log-ratio transformation is performed using a baseline reference feature D:

$$Y_{iq} = \phi(Z_{iq}) = \left\{ \log\left(\frac{Z_{iq}}{Z_{iD}}\right) \right\}_{q=1,\ldots,D-1,D+1,\ldots,Q} \tag{5.4}$$

A goal of DivNet is to account for the networked structure of microbial communities. Thus, the multivariate normal distribution is used to model the log-ratios, which allows co-occurrence of the features via the covariance parameter, Σ, of the probability mass function:

$$f(\boldsymbol{Y}_i|\mu, \Sigma) \propto |\Sigma|^{-1/2} \, \texttt{exp}\left\{ -\frac{1}{2} \left(\boldsymbol{Y}_i - \mu_i\right)^T \Sigma^{-1} \left(\boldsymbol{Y}_i - \mu_i\right) \right\} \tag{5.5}$$

where μ are the means, and Σ is the covariance matrix of the features, and $|\Sigma|$ is the determinant of the covariance matrix Σ.

Finally, the mean of $\boldsymbol{Y}_i$ is linked to the covariates by the following equation

$$\mu_i = X_i^T \gamma \tag{5.6}$$

where $X \in \mathbb{R}^{N \times P}$ is the covariate matrix associated with the samples ($P \geq 1$ is the number of covariates) and $\gamma \in \mathbb{R}^{p \times (Q-1)}$; that is γ_{ip} gives an expected increase in $\texttt{log}(\frac{Z_{iq}}{Z_{iD}})$ for one unit increase in X_{ip}.

An assumption of this model is that counts are conditionally independent given the covariate matrix $X \in \mathbb{R}^{N \times P}$ (i.e., the counts are independent of each other when

accounting for the information provided by the covariate matrix). Though data that is correlated spatially or temporally would violate the assumptions of this model, in practice DivNet performs better than other methods in these cases [393].

5.2.2.2 Estimating diversity

The model is used to estimate the community (and ecosystem) compositions from which the samples originate, which are then used to calculate α- and β-diversity indices. A strength of the model is that its diversity estimates explicitly model the feature-feature network structure present in sampled communities.

One of the parameters that is estimated under the log-ratio model is γ; let $\hat{\gamma}$ be that estimate. The expected value of the random variable $\boldsymbol{Y}_i$ (the log-ratios) is defined as $\hat{Y}_i = X_i^T \hat{\gamma}$, and define the fitted value of the latent composition of the community as $\hat{Z}_i = \phi^{-1}(\boldsymbol{Y}_i)$. Then the following estimate of any α-diversity index $f : \mathbb{S}^{C-1} \to \mathbb{R}$ is used:

$$\hat{\alpha}_i = f(\hat{Z}_i) \tag{5.7}$$

The β-diversity is estimated in a similar way. Given a β-diversity index $g : \mathbb{S}^{C-1} \times \mathbb{S}^{C-1} \to \mathbb{R}$, the estimate of β-diversity is

$$\hat{\beta}_{ij} = g(\hat{Z}_i, \hat{Z}_j), \tag{5.8}$$

where $\hat{Z}_i$ and $\hat{Z}_j$ are the fitted compositions of communities i and j, respectively.

That is, the fitted values of the latent composition of the communities, estimated using the model described above, is used as the input to the α- and β-diversity indices. Thus, diversity is estimated according to a model of community composition that explicitly accounts for (1) the compositional nature of the data, (2) the ecological network of feature-feature co-occurrence, and (3) any accompanying metadata describing the conditions from which the samples were observed. Note that this method gives the estimated diversity of the communities from which the samples originated rather than of the samples themselves, and that observed count information is shared across all

samples to yield better estimates. This is in contrast to the standard plug-in diversity estimates.

5.2.2.3 Parameter estimation

5.2.2.3.1 Estimating model parameters

Many parameters and variables in the model are unknown and must be estimated from the available data. Numerical methods are used to estimate these values as there is currently no analytical method available. In DivNet, and therefore, divnet-rs, a custom implementation of the Expectation-Maximization (EM) algorithm [76] with the Metropolis-Hastings (MH) algorithm [403] is used in the estimation procedure. For mathematical details of parameter estimation, readers are directed to the original DivNet manuscript [393].

5.2.2.3.2 Variance estimation

Another aspect of the DivNet model that requires attention is the estimation of variance for the diversity estimates. Accurate estimates of variance are crucial for hypothesis testing. While DivNet allows both parametric and nonparametric bootstrapping estimates of variance, divnet-rs implements only the parametric bootstrap as it was found to be more effective [393].

Recall that $\hat{\gamma}$ and $\hat{\Sigma}$ are the estimated values of γ and Σ. The parametric bootstrap approach, employed by divnet-rs, estimates $\texttt{Var}(\hat{\alpha}_i)$ and $\texttt{Var}(\hat{\beta}_{ij})$ for any α-diversity and β-diversity indices as follows. Given the fitted log-ratio model with $\mu = X\hat{\gamma}$ and $\Sigma = \hat{\Sigma}$, simulate B datasets. Then, for each of these B datasets, calculate the bootstrap estimates $\{(\hat{\gamma}^{(b)}, \hat{\Sigma}^{(b)})\}_{b=1}^{B}$, using the same estimation procedure described earlier. Next, the diversity index for each community i is calculated for each of the simulated datasets, i.e., $\{\hat{\alpha}^{(b)}\}_{b=1}^{B}$. Finally, the parametric bootstrap estimate of $\texttt{Var}(\hat{\alpha}_i)$ is $\widehat{\texttt{Var}}_b(\hat{\alpha}_i^{(b)})$, where $\widehat{\texttt{Var}}(\cdot)$ is the sample variance.

In plain language, for each bootstrap community estimate, the diversity index in question is calculated. Then the variance of those diversity indices is calculated, and

that value is the bootstrap estimate of the variance of the original diversity estimate. This procedure is applicable to any α-diversity or β-diversity index that is a function of the form given in Equations 5.7 and 5.8.

5.2.2.3.3 Feature covariance estimation

A final consideration of the model used by DivNet and divnet-rs is the parameter Σ, the covariance matrix of the features. As $||\Sigma||_\infty \to 0$ (the infinity norm of the covariance matrix approaches zero, i.e., the variance of the features and the covariance between features approach zero) the distribution of the observed count matrix W converges to a multinomial distribution. In the case of the multinomial distribution, the observed count of each feature is determined by its true proportion in the community. There is no parameter in the multinomial model to account for any variance in the observed counts from the true proportion. In contrast, the overdispersion of features in the log-ratio model as compared to the multinomial model is controlled by Σ, the covariance matrix of the features.

The multivariate normal distribution used to estimate the log-ratios requires inverting Σ, the matrix encoding covariance between features. The inverse (Σ^{-1}) as calculated during the estimation procedure is potentially a poor estimate of the true Σ^{-1} in a typical microbial community setting where the number of features is much greater than the number of samples. The original DivNet manuscript proposes multiple ways to account for this, including using microbial network estimation procedures to estimate Σ^{-1}, or restricting the estimators to diagonal covariance matrices. The second method is the one chosen by divnet-rs, both because it is likely a better estimate under the microbial setting where there are many more features than samples [393], and because it is significantly faster to calculate numerically than than using a full covariance matrix with non-zero off-diagonal entries, or using an external method such as SPIEC-EASI [179]. Restricting the covariance matrix to a diagonal matrix actually ignores any covariance between features (i.e., the inter-feature co-occurrence or network structure), but still allows overdispersion attributed to within-feature interactions as

compared to the standard multinomial model, which does not take this overdispersion into account. In simulation studies, it was shown that the intra-feature overdispersion (i.e., modeling intra-feature interactions) was more important than inter-feature interactions [393]. As divnet-rs prioritized computational efficiency, this tradeoff was deemed acceptable.

5.2.3 Measures of diversity

Estimates of community composition and their bootstrap replicates are generated with divnet-rs. These data are then used as inputs to diversity index functions in order to estimate the diversity of modelled microbial communities, as well as to calculate the variance in those diversity estimates (as described in Section 5.2.2.3.2).

5.2.3.1 Diversity formulas

The proposed framework uses the formulation of similarity aware diversity measures described by Reeve and colleagues [304].[6]

In the following formulas, vectors are shown in bold font, e.g., $\boldsymbol{x} = (x_1, ..., x_N)$. q is the abundance viewpoint parameter (traditionally simply referred to as the viewpoint parameter, but in this work, a similarity viewpoint parameter is introduced, so it is referred to as the abundance viewpoint parameter). It is the same q that determines the order of the Hill number. Hill numbers are weighted power means of order $1 - q$ that average inverses of the relative abundances of the features of a community [139]. p_i is the relative abundance of feature i in a given sample. $\boldsymbol{p} = (p_1, ..., p_S)$ is the relative abundances of all features in a given sample.

There are two notions of relative abundance used in Reeve's diversity measures: raw and normalized [304]. Raw measures take the relative abundance of features with respect to the metacommunity, while normalized measures take the relative abundance

[6] The notation and variable names used in this section match those used by Reeve rather than those used in previous sections for easier reference to the referenced manuscript.

of features with respect to the subcommunity. In this work, "metacommunity" refers to all samples within the count table generated in a study, and the "subcommunities" are samples in that table. The raw relative abundances of all features in the metacommunity, is called $\boldsymbol{P}$. Therefore, $\boldsymbol{P}_{ij}$ represents the abundance of feature i in subcommunity j relative to the total metacommunity. $\boldsymbol{P}_{.j} = (P_{1j}, ..., P_{Sj})$ represents the raw relative abundances of features in subcommunity j (again with respect to the metacommunity). By this formulation, subcommunity j is a fraction w_j of the metacommunity ($\sum_j^J P_{ij} = w_j$), where $\sum_j^J w_j = 1$. In other words, the count of each element in the count table divided by the total count of the metacommunity would yield each value P_{ij}. Normalized relative abundances consider the features of subcommunity j in isolation. In other words, the normalized relative abundances control for the size of the subcommunity via $\hat{\boldsymbol{P}}_{.j} = \boldsymbol{P}_{.j}/w_j$ and are constrained by $\sum_i^I \hat{P}_{ij} = 1$. Raw and normalized relative abundances then can both be used to define measures of diversity.

$\boldsymbol{Z}$ is a similarity matrix where the entries $Z_{ii'}$ represent the similarity between two types or features i and i'. $\boldsymbol{Zp}$ is the matrix-by-column vector multiplication with entries $(\boldsymbol{Zp})_i = \sum_{i'} Z_{ii'} p_{i'}$, and similarly for $\boldsymbol{ZP}_{.j}$.

The diversity measures described below are averages, that is, power means of order $1 - q$ weighted by the relative sizes of the elements (where q is the abundance viewpoint parameter). (For full exposition and interpretation of the measures, see the original work [304].)

The power mean of order r of $\boldsymbol{x}$ weighted by $\boldsymbol{u}$, assuming that $u_i > 0$ for all i (any term equal to zero should be removed prior to its calculation), is defined as:

$$\mathtt{M}_r(\boldsymbol{u}, \boldsymbol{x}) = \begin{cases} \left[\sum_i^I u_i x_i^r\right]^{\frac{1}{r}} & r \neq 0 \\ \prod x_i^{u_i} & r = 0 \end{cases} \tag{5.9}$$

Here, $\boldsymbol{u} = (u_1, \ldots, u_n)$ where $\sum_i^I u_i = 1$, $\boldsymbol{x} = (x_1, \ldots, x_n)$, and r is a real number. Further, at $r = 0$, $\frac{1}{r}$ is undefined and so the given expression comes from the limit as r approaches zero.

The normal notion of Hill number of order q is defined as

$$ {}^{q}D(\boldsymbol{p}) = \mathtt{M}_{1-q}\left(\boldsymbol{p}, \frac{1}{\boldsymbol{p}}\right) \tag{5.10} $$

The similarity-sensitive diversity is similar, though it incorporates the similarity of features via $\frac{1}{\boldsymbol{Zp}}$ rather than using $\frac{1}{\boldsymbol{p}}$:

$$ {}^{q}D^{\boldsymbol{Z}}(\boldsymbol{p}) = \mathtt{M}_{1-q}\left(\boldsymbol{p}, \frac{1}{\boldsymbol{Zp}}\right) \tag{5.11} $$

Note that in the naive-type case where each feature is completely dissimilar from every other feature, $\boldsymbol{Z}$ is the identity matrix and so $\boldsymbol{Zp} = \boldsymbol{p}$ (accounting for any necessary transpositions). Thus, in the naive-type case, the Hill number of order q (Eq. 5.10) is equivalent to the similarity-sensitive diversity of order q parameterized by $\boldsymbol{Z}$.

5.2.3.1.1 α-diversity

For α-diversity, Reeve's measure of subcommunity normalized α-diversity is used (i.e., similarity-sensitive diversity of subcommunity j in isolation) [304]:

$$ {}^{q}\bar{\alpha}_{j}^{\boldsymbol{Z}} = M_{1-q}\left(\bar{\boldsymbol{P}}_{\cdot j}, \frac{1}{\boldsymbol{Z}\bar{\boldsymbol{P}}_{\cdot j}}\right) \tag{5.12} $$

5.2.3.1.2 β-diversity

For β-diversity, Reeve's measure of metacommunity normalized β-diversity (i.e., the effective number of distinct communities) is used:

$$ {}^{q}\bar{B}^{\boldsymbol{Z}} = M_{1-q}\left(\boldsymbol{w}, {}^{q}\bar{\boldsymbol{\beta}}^{\boldsymbol{Z}}\right) \tag{5.13} $$

where ${}^{q}\bar{\boldsymbol{\beta}}^{\boldsymbol{Z}}$ is the subcommunity normalized β-diversity given by

$$ {}^{q}\bar{\boldsymbol{\beta}}_{j}^{\boldsymbol{Z}} = \frac{1}{{}^{q}\hat{\rho}_{j}^{\boldsymbol{Z}}} \tag{5.14} $$

and ${}^{q}\hat{\rho}_{j}^{\boldsymbol{Z}}$ is the normalized ρ (that is the reversed normalized beta–the representativeness of subcommunity j):

$$ {}^{q}\bar{\rho}_j^{\boldsymbol{Z}} = \mathtt{M}_{1-q}\left(\bar{\boldsymbol{P}}_{\cdot j}, \frac{\boldsymbol{Zp}}{\boldsymbol{Z}\bar{\boldsymbol{P}}_{\cdot j}}\right) \tag{5.15} $$

5.2.3.2 Transforming sequence identity

While any notion of feature similarity can be used in the similarity-aware measures described above, this work focuses on using sequence identity. Percent identity between two sequence pairs is transformed into a similarity score via the formula $S = \left(\frac{P}{100}\right)^w$, where P is the percent identity and w is the similarity viewpoint parameter. The effect of this transformation is illustrated in Figure 5.1.

When combined with the similarity-aware measures of diversity described below, this transformation allows varying the "weight" placed on sequence identity. Increasing values of w increasingly deemphasize the similarity of sequence pairs with low percent identity.

5.2.4 Sample distances & ordinations

A distance measure between all pairs of communities can be calculated calculated using the formula for effective number of distinct communities (Eq. 5.13). For each pair, the effective number of distinct communities is calculated. By construction, this measure ranges from one to the total number of subcommunities included in the calculation. Since only two subcommunities are included in the metacommunity (as the measure is calculated for every pair), the value ranges from one when the subcommunities are completely overlapping, to two, when the subcommunities are completely distinct. Finally, 1 is subtracted from the beta diversity measure to transform it into a distance that runs from zero to one. These pairwise distance calculations are done for each abundance-similarity viewpoint pair.

Once distance matrices are obtained, multideminsional scaling (MDS) ordinations are calculated with the base R `cmdscale` function. The variance in the ordinations is calculated in an analogous way to the variance of the diversity indices, i.e.,

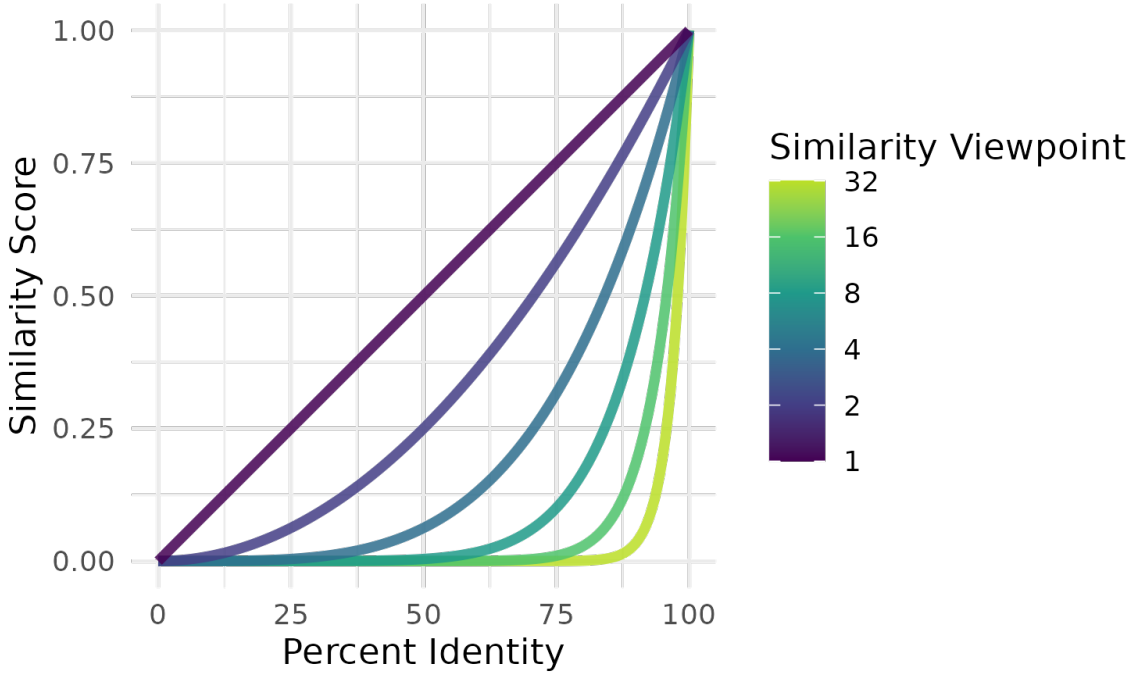

Figure 5.1: Conversion of percent identity to similarity score is determined by the similarity viewpoint parameter. Percent identity is transformed into similarity score, S, via the formula $S = (P/100)^w$, where P is the percent identity and w is the similarity viewpoint. At the minimum similarity viewpoint of 1, the transformation between percent identity and similarity scores is linear: sequence similarity score is directly proportional to the percent identity of the sequence pairs. As the similarity viewpoint increases, the transformation becomes more non-linear, yielding an increasingly more conservative evaluation of similarity by increasingly deemphasizing sequence pairs with lower percent identity (i.e., at higher similarity viewpoints more weight is given to sequence pairs with high percent identity). This is directly analogous to the abundance viewpoint parameter, q, in which increasing values yield more conservative estimates of diversity by deemphasizing less abundant types.

via a bootstrapping procedure. For each bootstrap estimate of community composition generated by divnet-rs, the procedure for calculating pairwise distances based on β-diversity followed by MDS ordination is repeated. The variance in point location in the two-dimensional ordination space is calculated and used as a proxy for the uncertainty in the ordination of community composition. This uncertainty is shown as ellipses centered around each plotted point.[7] Ellipse diameter on the x- and y-axes (PC1 and PC2, respectively) represent four standard deviations in the bootstrap point locations. All abundance viewpoint parameter ordinations are plotted on the ordination, and individual ecosystems/samples are connected via a line and colored by the viewpoint parameter to show the relationship between abundance viewpoint parameter and ordination positioning. This process is repeated for each similarity viewpoint parameter. Thus, for each abundance-similarity viewpoint parameter pair an ordination is calculated for the original community composition estimates as well as for each of the bootstrap estimates.

5.3 Estimating diversity of cattle microbiome communities

To demonstrate the utility of the diversity framework introduced in this chapter, a cattle microbiome dataset was used. Class I α and Class III ribonucleotide reductase (RNR) sequences were used as gene markers for investigating the microbial communities associated with cattle hide and fecal microbiome. Experimental methods for cattle sample collection, Shiga toxigenic *Escherichia coli* (STEC) detection, microbiome sequencing, peptide assemblies, and RNR identification can be found in Appendix A.

5.3.1 Modeling community composition

Community composition estimates were calculated using divnet-rs version 0.3.0. Two divnet-rs runs were conducted, one for the Class I α RNR sequences, and one for

[7] This style of displaying uncertainty on an ordination was influenced by the QIIME script `jackknifed_beta_diversity.py` [52, 178].

the Class III RNR sequences. Options shared between all divnet-rs runs are expectation maximization (EM) iterations: 6, EM burn: 3, Monte-Carlo (MC) iterations: 500, MC burn: 250, step size: 0.01, perturbation: 0.5, and number of replicates: 6. Options specific to Class I α – base taxa 0 (clu_95___seq_30493073), random seed: 9032874. Options specific to Class III – base taxa 9 (clu_95___seq_5546017), random seed: 2398732. Taxa that were both highly abundant and present in a majority of samples were chosen as base taxa for the log-ratio transformation.[8] The covariates included in the model were fraction (cellular/viral), location (fecal/hide), and STEC presence (Yes/No).

5.3.2 Diversity calculations

Abundance viewpoint parameters used were the sequence of numbers from 0 to 10 with a step of 0.5. Similarity viewpoint parameters were 1 and 8. The maximum values of these parameters were chosen as resulting diversity estimates at higher values were essentially indistinguishable from the chosen max values (data not shown).

Diversity values and variance in the estimates were calculated in accordance with the DivNet model of diversity. Groupings for divnet-rs were fraction (cellular/viral), location (fecal/hide), and STEC presence (yes/no). That is, the diversity estimate for each group (i.e., fraction-location-presence) was calculated from "replicate 0" returned by divnet-rs, (i.e., the diversity estimates for the groups themselves). Variance in the diversity estimate was calculated via the bootstrap estimate of variance as described above[9], i.e., diversity was calculated for each bootstrap replicate of community composition, and then the variance was calculated from those bootstrap diversity estimates. Diversity was calculated once for each abundance and viewpoint parameter pair, using equation 5.12 as described above.

Significance testing for all pairs of abundance and viewpoint parameters was done using the `betta` function from the `breakaway` R package [391], with Bonferroni

[8] In accordance with guidelines proposed in [123].

[9] See section 5.2.2.3.2: Variance estimation.

multiple test correction. Alpha diversity results and hypothesis testing was plotted using the R programming language using ggplot2.

A distance measure between all pairs of ecosystems was calculated using the calculation for effective number of distinct communities as described above (Eq. 5.13). To generate the ordinations, these distance matrices were used as input to multidemin-sional scaling (MDS) using the base R `cmdscale` function as described above.

5.4 Results & Discussion

In this chapter, I set out to develop a framework for examining community diversity that address some of the major concerns that arise when dealing with metagenomic studies of microbial community diversity: (1) estimating microbial community diversity of large samples using compositionally aware models in an efficient way, (2) comparing the influence of type- and trait-level diversity when studying microbial communities, (3) measuring the impact of both the abundant community members as well as the influence of the "long tail" of rarer microbes, and (4) attempting to account any noise or negative effects due to the sparsity of metagenomic community structure data.

5.4.1 Accessible compositional models of diversity

A fundamental property of microbial communities as measured by next generation sequencing data is that they are compositional [114]. That is, the observed counts generated from the sequencing procedure are subject to the constant-sum constraint– there is a finite number of sequencing reads, and therefore observations, per sequencing run [297]. This has profound implications for statistical methods and downstream analyses of the data, such as inducing a negative bias between observed features, and generating counts with magnitudes that are decoupled from biological or ecological reality of the ecosystem from which they are sampled [298, 114]. Many of the frequently used analysis toolkits do not have good support for compositional data analysis. However, there has been a growing realization of the importance of explicitly accounting

for the compositional nature of NGS community ecology data, and some tools using a CoDA framework have been developed (e.g., [94, 95, 299, 228, 201]).

Estimating diversity is a particular challenge for microbial communities, both because of the compositional nature of their measurement data, and because of the high levels of diversity coupled with highly uneven community structure, which leads to high-dimensional, sparse count tables of compositional vectors [197, 298]. In addition, microbial communities are highly networked, characterized by many intra- and inter-taxa interactions that have a measurable impact on their structure [92, 96].

Diversity indices are functions that summarize relative abundance and community composition information from one or more communities, with α-diversities summarizing within-community structures, and β-diversities summarizing between-community structure. Microbial ecologists typically use plug-in estimates of diversity that do not always account for important aspects of microbial community data. However, statistically sound alternatives can be found in the literature. Many statistical models of diversity first attempt to model the community structure of the samples, communities, or ecosystems under study, and from there, use that data to estimate the diversity measures under question. While the statistical literature focusing on α-diversity is rich with examples [74, 15, 309, 51], β-diversity is comparatively understudied.

A recent model of community composition and diversity that incorporates many of the important properties of microbial communities is DivNet [393]. DivNet uses a compositional model of community composition that explicitly accounts for the networked structure of microbial communities, while also leveraging information across samples and including covariate information into its model of community composition.

DivNet is highly accurate at both estimating diversity given the modeled community compositions, as well as estimating the variance of those diversity estimates with its bootstrapping procedure [393]. However, the current implementation is limited in the sizes of datasets that it can handle, especially in cases where researchers do not have access to high performance computing. Large microbiome datasets with

hundreds of samples and thousands of features, are common, yet the reference DivNet implementation struggles handling data of this size. As a workaround, taxa or features must be grouped or collapsed based on some external metadata such as taxonomic groups, or through other means like clustering.

To address this limitation in the original DivNet implementation, I introduced divnet-rs, a fast, parallelizable, memory efficient implementation of the DivNet model of community composition. While it only provides a subset of the functionality provided by the original implementation, (e.g., only the parametric bootstrap and using the diagonal rather than full covariance matrices), it makes it possible to apply the DivNet model to datasets with up to hundreds of thousands of features in a high performance computing environment, and to datasets with tens of thousands of features on commodity hardware. This allows researchers to avoid unnecessary grouping of features by taxonomy or with clustering. Divnet-rs has an added advantage of making the bootstrap estimates of community composition available to the researcher. This enables the use of these estimates for downstream tasks, such as estimating variance in ordinations based on the model's output.

These modeled estimates of community composition, as well as the bootstrap replicates, generated by divnet-rs are used as inputs to the described diversity measurements parameterized by the abundance and similarity viewpoint parameters. In this way, the diversity estimates account for critical features of the microbial communities including their compositional nature, networked interactions, and covariate information, as well as patterns of abundance and similarity of features therein.

5.4.2 Microbial community diversity "viewpoints"

5.4.2.1 Type-level vs. trait-level diversity

Some researchers have argued that focusing on the functional potential of a microbial community is more valuable than only using a taxonomic or species based approach [186]. This reflects the importance of the function of microbes in the environment. Functional niche is likely a useful alternative to commonly employed species

classification schemes, as microbial function, physiology, and biochemistry can be more appropriate classifiers of microbes at the community level [188]. While treating functional groupings as the major organizational structure of microbial communities in this way is compelling, it is a break from the more traditional species diversity based approaches commonly employed [188]. Thus, bridging the gap between these two ideas is an open question–one that the work presented in this chapter attempts to partially address.

In this work, a distinction is made between type-level and trait-level measures of diversity. By type-level diversity, I mean the diversity of the types of a community, be they species, operational taxonomic units (OTUs), or some other constructed type. By trait-level diversity, I mean the diversity of entities that have more of a connection to some aspect of the community than the types themselves, e.g., phylogenetic groups, metabolic pathways, sequence similarity, or others. Note that depending on the formulation of trait-level diversity, this distinction is ultimately semantic.

As an example of this, consider the formulation of trait-level diversity using sequence similarity. Homology between sequences, as inferred by statistically significant levels of similarity, likely reflects shared ancestry [162, 280]. Additionally, homology has a long history of being used to infer function of unknown proteins (e.g., [207, 217]). In this way, similarity in the primary structure of protein sequences can be used to infer a sort of trait-level diversity of a community, especially when used with a protein coding marker gene that is connected to important aspects of the community or ecosystem under study, and has a large amount of biochemical characterization (e.g., different classes of ribonucleotide reductase (RNR) [325]). At the same time, individual sequences themselves are "types" and can thus be used to define a type-level view of community diversity. The problem is reconciling these two views of community diversity.

In this study, a smooth transition between type-level and trait-level community diversity is realized using specific protein coding gene markers, combined with estimates of community structure from divnet-rs that are used to calculate diversity

using the formulation of Reeve and colleagues [304]. In this setting, the α-diversity and β-diversity diversity measures have an effective number interpretation: the effective number of distinct types. Though similarity can be any measure (phylogenetic, number of shared traits, sequence identity, etc.), whichever notion of similarity is chosen changes the effective number interpretation of the measure. For example, using a notion of phylogenetic similarity would yield a measure of the effective number of distinct phylogenetic groups. Here, sequence identity is used, so the interpretation is effective number of distinct sequences.

Through the use of the similarity viewpoint parameter, introduced in this work, the notion of "distinct" can be controlled. For example, when the similarity viewpoint parameter is 1 (the minimum value), then the only "distinct" sequence pairs would be those that have 0% identity. Sequences that are 50% identical would measure 0.5 on the similarity scale–exactly halfway between distinct and indistinguishable. At a similarity viewpoint of 2, however, the same 50% identity pair would be 0.25 on the similarity scale. And so on as the similarity viewpoint increases, sequence pairs need ever higher percent identities to not effectively be considered as distinct (Fig. 5.1). "Effectively" is used because while at higher similarity viewpoints sequences with low percent identities have a similarity score very close to zero, it is never zero (rather, the limit approaches zero as the percent identity decreases).

In the context of sequence identity, this is a particularly desirable behavior. The so-called "twilight zone" of sequence identity occurs somewhere between 20-30%–sequences below this level of identity are generally considered to not be related, though exceptions, such as the RNR sequences used in this study, occur [319, 212]. The similarity viewpoint parameter can be adjusted to account for the fact that sequences with lower levels of percent identity are effectively distinct.

As the similarity viewpoint parameter is increased in the similarity transformation function, the differences between sequence pairs with a high percent identity are dilated, whereas differences between sequence pairs with low percent identity are

compressed[10] (Fig. 5.1). In this way, at higher similarity viewpoint parameters, subtle differences between highly similar sequences are more important in the diversity calculations, whereas the difference in similarity between sequences with low sequence identity has less of an effect on the calculations (i.e., they are effectively considered to be distinct.)

Thus, coupling sequence identity of protein coding marker genes with varying values of the similarity viewpoint parameter allows for a smooth bridging of the notions of type- and trait-level diversity.

5.4.2.2 Abundant vs. rare community members

Microbial communities are incredibly complex, and are generally dominated by a small number of abundant members combined with a long tail of of type- and trait-level diversity [350]. This rare biosphere of microbial life contains an enormous amount of taxonomic, phylogenetic, and functional diversity [86, 215], and likely plays an outsized role in the long term maintenance and functional health of microbial communities [350, 339, 331]. Additionally, the long-tail of microbial diversity is likely active in important ecosystem services like nutrient cycling and pollutant degradation, in addition

[10] To understand why this is desirable, consider the following scenario: A hungry graduate student at the University of Delaware is deciding where to go for lunch. Option A is five minutes away, Option B is ten minutes away, and Option C is twenty minutes away (each twice as far as the previous one). There are other options, but they are quite far indeed: Option D is 10 hours away and Option E is 20 hours away. From the student's perspective, and due to the constraints of their one hour lunch break, the initial doublings in time taken to arrive at the restaurant would be perceived as quite important–five minutes feels much quicker than twenty minutes–1/12 vs. 1/3 of the entire lunch break. Option E is also double the time away from the student as compared to Option D. However, the difference in distance and time taken from the student's point of view is essentially meaningless–both options are so far away that they are not even considered. So, even though Option E is twice as far as Option D, from the student's point of view, they may as well be equally as distant. So it goes with the sequence identity: differences between extremely similar proteins pairs by percent identity are magnified by the transformation, whereas differences between very dissimilar protein pairs are compressed.

to having an active effect on health of host organisms through their associated microbiomes [285, 157]. Given the growing acknowledgement of their importance, an increased understanding of their role in shaping community level diversity is crucial–a comprehensive understanding of microbial communities requires an understanding of both the abundant members and those in the long-tail of less abundant microbes [274].

Analogously to the way that varying the similarity viewpoint parameter can give a more nuanced examination of type- vs. trait-level diversity, varying the abundance viewpoint parameter yields a more nuanced examination microbial community diversity by comparing of the effect of abundant vs. rare members on the measured diversity. Comparing diversity calculated at different abundance viewpoint parameter values reveals insights into the structural similarities and differences between communities [376, 169]. Lower values emphasize rarer community members, while higher values place more emphasis on abundant ones. While differences between the more abundant community members likely reflect important ecosystem level differences in niche preference, differences in the long-tail of abundance could be equally as interesting, especially when considering the overall functional potential of a community, or its potential resilience in the face of change.

5.4.2.3 Zero-replacement induced artifacts

Compositional data analysis involves the use of the log-ratio to transform values in the simplex to Euclidean space. One potentially problematic aspect of the log-ratio transformation is that any zero counts must be handled in some way. While there are many options for zero-replacement (e.g., [266]), a commonly used strategy is to replace all zeros with a small constant less than one. Given the sparsity of count tables generated from microbial communities, the zero replacement procedure can lead to noticeable distortions in downstream analyses. For example, the similarity between samples with fewer observations may be artificially inflated due to the samples sharing more features at very low abundance [211].

Table 5.1: Count table for a mock community of four samples and ten taxa.

Sample	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
A1	5	4	3	2	1	0	0	0	0	0
A2	5	4	3	2	1	1	1	1	1	0
B1	0	0	0	0	0	1	2	3	4	5
B2	0	1	1	1	1	1	2	3	4	5

Another benefit of varying the abundance viewpoint parameter is a technical one. Given that the zero-replaced values distort the rare members of the community, one may expect to notice such distortions only when utilizing low abundance viewpoint parameters. For example, selecting an abundance viewpoint parameter of zero is effectively a measure of richness–all species are treated the same regardless of whether they were observed a single time or a million times. In data processed with a zero-replacement procedure prior to analysis, all samples will have a richness value equal to the total number of features, as any feature absent from a sample will be replaced with some non-zero value. The closer to zero that the abundance viewpoint parameter is, the more of an effect the zero-replaced values will have. A smooth increase in the abundance viewpoint parameter can thus be used to examine any possible artifacts of the zero replacement procedure.

As an example, consider the mock community presented in Table 5.1 consisting of four samples and ten taxa. The sample pairs A1-A2 and B1-B2 are structurally similar, though they differ in the amount of taxa with zero counts, with A1 and B1 containing 50% zero count entries, whereas A2 and B2 contain only 10% zero count entries.

Zeros were replaced by a constant value of 0.05 and β-diversity was calculated and converted as described above for sample pairs A1-B1 and A2-B2 for varying values of abundance viewpoint parameter (Fig. 5.2). Intuitively, the β-diversity calculation for A1-B1 should be more affected by the zero replacement as 50% of the values in both samples are replaced, whereas the calculated β-diversity for A2-B2 with only 10% zero

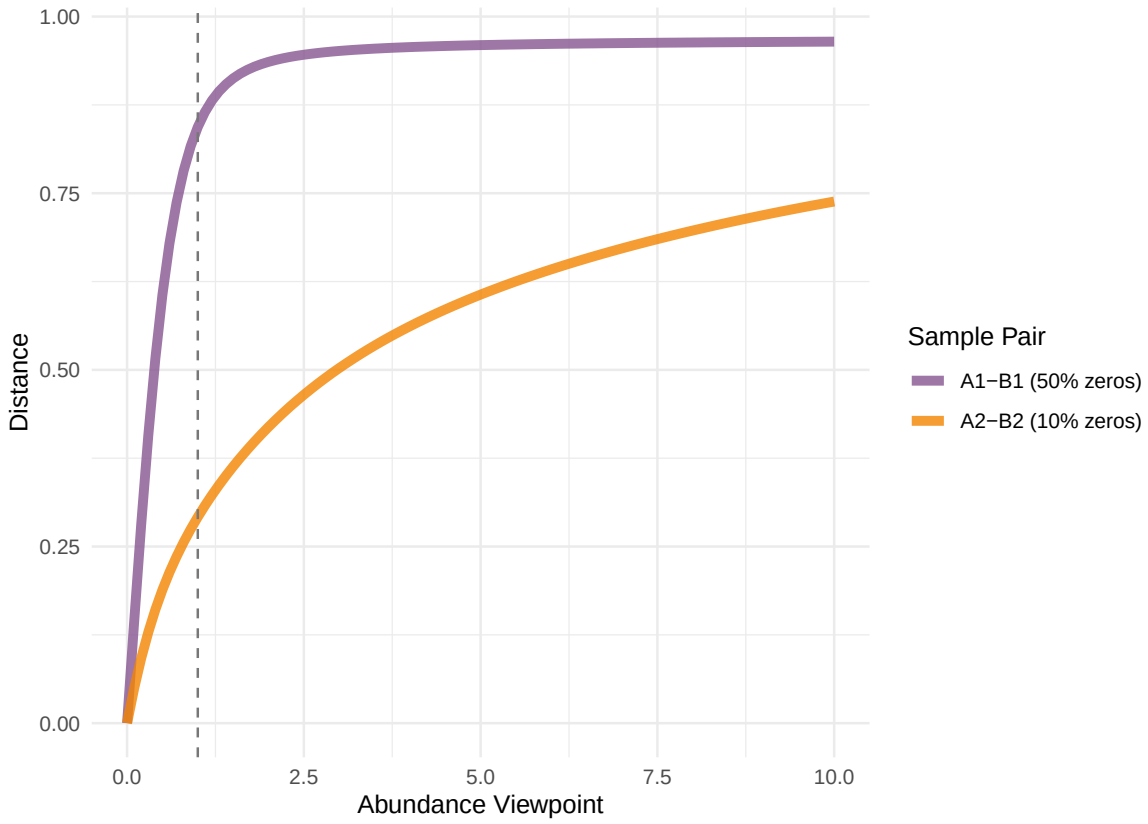

Figure 5.2: β-diversity of a mock community. β-diversity distance (y-axis) calculations for varying abundance viewpoint parameter (x-axis) of the mock community shown in Table 5.1. The lines represent the distance between sample pair A1-B1 (purple), each with 50% zeros, and A2-B2 (orange), each with 10% zeros. Dashed line marks the abundance viewpoint parameter of 1 (equivalent to the Shannon diversity). Here, a distance of zero indicates identical communities while a distance of one represents completely distinct communities.

values should be less affected.

As the abundance viewpoint moves from zero (richness) to one (Shannon) and beyond the rate at which the distance increases is different, with the 50% zero pair (A1-B1) distance increasing more sharply than the 10% zero pair (A2-B2). This indicates that the similarities between the communities represented by A1 and B1 are mostly restricted to the rarer members of those communities. As the rarer taxa are increasingly deemphasized, the sample pair looks increasingly distinct. Samples A1 and B1 share no taxa in the count table, and so the similarity seen between the two samples at lower abundance viewpoints is an artifact of the zero replacement procedure. Compare that to the behavior of sample pair A2-B2, which have considerably more overlap in their taxa, as well as having fewer zero counts. For this pair, the rate of increase of the distance with increasing abundance viewpoint is much less than that of pair A1-B1, indicating that an increasing deemphasis of rare members has less of an effect on the distance between the sample pairs. In the case of the mock community, the sharp increase in distance seen in the A1-B1 calculations can mostly be attributed to the noise introduced by the zero replacement procedure. In real data, that cannot be known for certain; however, such a comparison could potentially guide the researcher to more critically examine samples whose patterns of diversity show rapid or unexpected levels of change in the extreme low range of the abundance viewpoint parameter.

5.4.3 RNR diversity in cattle hide and fecal microbiome

Ribonucleotide reduction is the rate limiting step of DNA synthesis and is catalyzed by the enzyme ribonucleotide reductase (RNR) [174, 5]. RNRs provide the only method of *de novo* deoxyribonucleotide production and are therefore present in virtually all cellular life and common in the genomes of lytic dsDNA viruses [80, 327, 149]. Despite sharing a common ancestor [14, 212], they are biochemically diverse and require different cofactors and environmental conditions [307]. Class I RNRs are O_2-dependent and most require a di-metallic cofactor, with the identity of the cofactor further dividing Class I RNRs into several subclasses. Class II RNRs are O_2-independent and require

adenosylcobalamin (a form of B_{12}). They come in two main sub-types: monomeric and dimeric. In addition to differences in quaternary structure, most dimeric Class II RNRs also require a zinc atom [206]. Class III RNRs are O_2-sensitive and use an iron-sulfur cluster as a cofactor. Because of the large biochemical differences between RNR types, organisms tend to carry the type, or types, best suited to their ecological niche [307, 66], making them interesting targets for microbial ecologists. For example, Class III RNRs are only found in strict or facultative anaerobes.

5.4.3.1 Cattle microbiome RNRs

Putative RNR sequences were defined as any sequence that had significant homology (E-value $< 10^{-1}$) to some sequence in the RNRdb [214]. By this criteria, 344,025 putative RNR sequences were identified from the 52,891,368 OTUs generated from the clustered Plass assembly sequences. Given the lenient criteria acceptance criteria, the putative RNRs were subjected to post-homology search validation with PASV[11] and via manual curation.

Due to the significant variation in oxygen exposure between cattle hide and fecal samples, this analysis focuses on Class I and Class III RNRs. In total, 344,025 sequences of the 52,891,368 OTUs generated from the clustered Plass assembly of cattle hide and fecal metagenomes had significant homology (E-value $< 10^{-1}$) to sequences in the RNRdb [214] (i.e., putative RNR sequences). The post-homology search validation process yielded 1,464 Class I α and 6,224 Class III RNRs. After 95% clustering, 3,474 Class III RNR sequences remained. These RNR sequences were used as gene markers in the diversity framework described above to examine differences between hide and fecal samples, viral and microbial metagenomes, and STEC positive and negative samples (Fig. 5.3).

Class I RNRs showed similar amounts of trait-level alpha diversity, but cellular and viral fecal metagenomes had higher levels of type-level alpha diversity as compared to cellular hide metagenomes (Fig. 5.3). This could potentially be due to higher

[11] See Chapter 2 for a discussion of PASV.

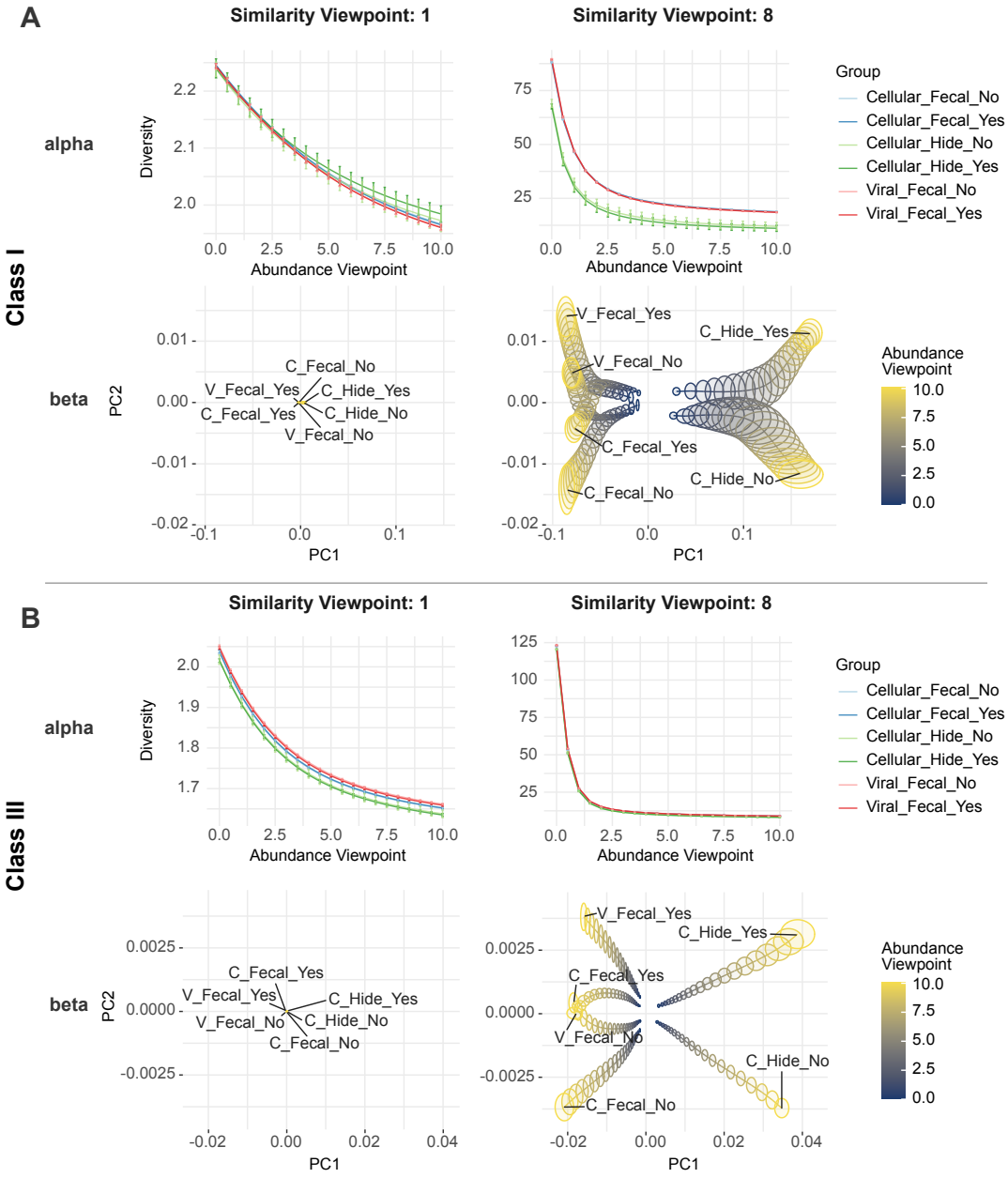

Figure 5.3: Class I and III RNR diversity of the cattle microbiome. α- and β-diversity values with varying abundance and similarity viewpoints for Class I (panel A) and Class III (panel B) ribonucleotide reductase (RNR) from cattle hide and fecal microbiomes. Each column represents a different similarity viewpoint (left column: 1–more emphasis on trait-level diversity, right column: 8–more emphasis on type-level diversity). Within panels, the top row shows α-diversity and the bottom row shows β-diversity ordinations, both for varying abundance and similarity viewpoints. The x-axis for α-diversity plots represents the abundance viewpoint, in which increasing values indicate decreasing emphasis of rare types; the y-axis gives the effective number measure of diversity parameterized by the given abundance and similarity viewpoints; color shows the sample group–cellular fraction/viral fraction, fecal/hide, and STEC positive/STEC negative (Yes/No); and error bars represent two standard deviations. The color in the ordinations represents the abundance viewpoint, in which increasing values (more yellow) indicate decreasing emphasis of rare types. Points and lines indicate estimated diversity values parameterized by the given abundance and similarity viewpoints. Ellipses indicate two standard deviations in the diversity estimates. Distance between points in the ordination space are approximations of the calculated beta diversity between pairs of samples. In both the α- and β-diversity plots, the measure of diversity becomes increasingly more conservative with respect to abundance and similarity as their respective viewpoint parameters are increased.

mutation rates of Class I RNR within the gut as opposed to on the hide. Mutation rates are known to increase due to stress events [93]. One common source of stress in cattle is heat stress, which is known to cause changes in the gut microbial community [63, 271]. Throughout the sampling period, temperatures regularly reached or surpassed temperatures used for the stress treatment in cattle heat stress studies [63, 271]. In one study of heat stress in beef cattle, the relative abundance of protein-coding genes was affected, though the predicted functional profile of the rumen microbial community did not change [271]. While the effect of heat stress on cattle hide microbiota has not been studied, these microbes are routinely exposed to ultraviolet radiation, a known mutagen, so it is possible that heat stress is not significant by comparison for the hide microbial community. In fact, one study found that the percentage of animals suffering from digital dermatitis, a common bacterial infection of bovine hooves [89], actually decreased with increasing temperatures [110].

Additionally, mutation rates have sometimes been measured to be higher for genes with lower rates of transcription [152, 226]. Some facultative anaerobes and their viruses (e.g., *E. coli* and bacteriophage T4) are known to carry both Class I and Class III RNRs and switch between the types depending on environmental conditions [103, 409, 239, 45, 80]. This could be beneficial as most of the cattle gut is not strictly anoxic [202]. Oxygen concentrations decrease sharply after the mouth, meaning that oxygen is quickly depleted and present in very low concentrations throughout most of the gut [227]. Consequently, although microbes may benefit from carrying Class I RNRs, they are likely seldom used within the gut and therefore transcribed infrequently, potentially leading to increased mutation rates.

For Class III RNRs, there were no differences in type-level or trait-level alpha diversity among any of the groups (Fig. 5.3). Cattle hides are exposed to oxygen, making them inhospitable environments for Class III RNRs, so it was initially surprising that the hide samples contained the same trait-level diversity as fecal samples. However, fecal contamination is common on cattle hides [232, 56]. In one survey of microbial diversity in cattle hide and feces, less than 3% of the SSU rRNA OTUs were

specific to the hide [59]. This suggests that microbes carrying Class III RNRs are highly likely to be present on the hides as well, potentially due to cross-contamination, which could result in similar levels of diversity.

In both Class I and Class III RNRs, there were little differences in β-diversity at the type-level, but there were some changes in trait-level β-diversity at a higher similarity viewpoints (Fig. 5.3). β-diversity of each sample type was more similar at low abundance viewpoints and then diverged at higher abundance viewpoints, with PC1 showing an increasing divide between cattle and hide communities. This indicates that the cattle and hide communities share many rare members, but differ in their highly abundant members. This pattern may reflect the Baas Becking hypothesis that "everything is everywhere, but the environment selects" [73], i.e., that while the microbial communities measured in this study are highly similar in terms of their rarer members, their unique niches have allowed different microbes to flourish. It is also important to note that any examination of rare members of the community is more likely to be influenced by zero replacement artifacts, which my contribute to more similar β-diversity measures at lower abundance viewpoints; however, unlike in the mock communities shown in Fig. 5.2, transitions from lower to higher abundance viewpoints were more gradual and so, potentially less influence by zero-replacement induced noise.

Regardless, further investigation is necessary to explore these hypotheses, ideally including other marker genes to include more "views" into the cattle microbiome.

5.5 Conclusions

In this chapter, I have introduced a diversity framework that combines state-of-the-art compositional models of microbial community composition, with similarity-aware measures of diversity. Combining these diversity measures with varying values of similarity and abundance viewpoint parameters, as well as a careful choice of protein coding marker genes, allows for subtle and nuanced queries of microbial community diversity. These include connecting summaries of community diversity to questions of

the differences in type-level and trait-level diversity as well as of the differences between diversity of abundant community members and the "long tail" of the rare biosphere. Finally, particular genome-to-phenome hypotheses are enabled in this framework by the principled choice of appropriate protein coding marker genes.

CONCLUSIONS

Together, the tools and frameworks presented in the previous chapters advance gene-centric approaches for microbial ecology by making specific "pain points" in the sample-to-sequence-to-discovery pipeline more accessible, and increasing the quality of metagenomic analyses. The selected applications span a wide variety of microbial and viral systems, with a particular focus on environmental viruses. PASV gives researchers the ability to validate large peptide datasets rapidly and puts the domain knowledge of individuals or research groups in the hands of non-domain experts, democratizing the study of single genes. InteinFinder lessens the burden of manual identification and curation of inteins in peptide data sets by standardizing the search for inteins, opening the possibility for intein screening to become routine, even in large datasets. Iroki makes powerful visualizations of phylogenetic data with metadata available to researchers without a strong background in programming or command line tools for batch processing. And lastly, the diversity framework makes compositional data models of microbial community structure more accessible to research groups without access to a high performance computing environment, allowing for accurate estimations of diversity on commodity hardware. By reducing the complexity and the amount of domain knowledge traditionally needed for gene-centric study of microbial communities, the work of this dissertation makes a marked improvement in the metagenomic sample-to-sequence-to-discovery pipeline.

BIBLIOGRAPHY

[1] CRediT. https://credit.niso.org/. Accessed: 2023-11-16.

[2] Evelien M Adriaenssens and Don A Cowan. Using signature genes as tools to assess environmental viral ecology and diversity. *Appl. Environ. Microbiol.*, 80(15):4470–4480, August 2014.

[3] Charles Affourtit, Mary S Albury, Paul G Crichton, and Anthony L Moore. Exploring the molecular nature of alternative oxidase regulation and catalysis. *FEBS Lett.*, 510(3):121–126, January 2002.

[4] Md. Faiz Ahmad, Prem Singh Kaushal, Qun Wan, Sanath R. Wijerathna, Xiuxiang An, Mingxia Huang, and Chris Godfrey Dealwis. Role of Arginine 293 and Glutamine 288 in Communication between Catalytic and Allosteric Sites in Yeast Ribonucleotide Reductase. *Journal of Molecular Biology*, 419(5):315–329, June 2012.

[5] Md Faiz Ahmad, Prem Singh Kaushal, Qun Wan, Sanath R Wijerathna, Xiuxiang An, Mingxia Huang, and Chris Godfrey Dealwis. Role of arginine 293 and glutamine 288 in communication between catalytic and allosteric sites in yeast ribonucleotide reductase. *J. Mol. Biol.*, 419(5):315–329, June 2012.

[6] Pakorn Aiewsakun, Evelien M Adriaenssens, Rob Lavigne, Andrew M Kropinski, and Peter Simmonds. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *The Journal of General Virology*, 99(9):1331–1343, September 2018.

[7] J Aitchison. The statistical analysis of compositional data. *J. R. Stat. Soc.*, 44(2):139–160, January 1982.

[8] J Aitchison, C Barceló-Vidal, J A Martín-Fernández, and V Pawlowsky-Glahn. Logratio analysis and compositional distance. *Math. Geol.*, 32(3):271–275, April 2000.

[9] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990.

[10] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, September 1997.

[11] Christopher L Anderson, Matthew B Sullivan, and Samodha C Fernando. Dietary energy drives the dynamic response of bovine rumen viral communities. *Microbiome*, 5(1):155, November 2017.

[12] Rika E Anderson, Mitchell L Sogin, and John A Baross. Evolutionary strategies of viruses, bacteria and archaea in hydrothermal vent ecosystems revealed through metagenomics. *PLoS One*, 9(10):e109696, October 2014.

[13] Simon C Andrews. The ferritin-like superfamily: Evolution of the biological iron storeman from a rubrerythrin-like ancestor. *Biochim. Biophys. Acta*, 1800(8):691–705, August 2010.

[14] L Aravind, Y I Wolf, and E V Koonin. The ATP-cone: an evolutionarily mobile, ATP-binding regulatory domain. *J. Mol. Microbiol. Biotechnol.*, 2(2):191–194, April 2000.

[15] Julyan Arbel, Kerrie Mengersen, and Judith Rousseau. Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity. *aoas*, 10(3):1496–1516, September 2016.

[16] David W Armitage and Stuart E Jones. How sample heterogeneity can obscure the signal of microbial interactions. *ISME J.*, 13(11):2639–2646, November 2019.

[17] Francesco Asnicar, George Weingart, Timothy L. Tickle, Curtis Huttenhower, and Nicola Segata. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, 3:e1029, June 2015.

[18] Christian Bachmaier, Ulrik Brandes, and Barbara Schlieper. Drawing phylogenetic trees. (Extended abstract). In Xiaotie Deng and Ding-Zhu Du, editors, *ISAAC: 16th International Symposium on Algorithms and Computation*, volume 3827 of *Lecture Notes in Computer Science*, pages 1110–1121. Springer, 2005.

[19] Charles Bachy and Alexandra Z Worden. Microbial ecology: finding structure in the rare biosphere. *Curr. Biol.*, 24(8):R315–7, April 2014.

[20] Mohammad Bahram, Sten Anslan, Falk Hildebrand, Peer Bork, and Leho Tedersoo. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. *Environ. Microbiol. Rep.*, 11(4):487–494, August 2019.

[21] Jianfa Bai, Zachary D Paddock, Xiaorong Shi, Shubo Li, Baoyan An, and Tiruvoor G Nagaraja. Applicability of a multiplex PCR to detect the seven major

shiga toxin-producing escherichia coli based on genes that code for serogroup-specific o-antigens and major virulence factors in cattle feces. *Foodborne Pathog. Dis.*, 9(6):541–548, June 2012.

[22] Jianfa Bai, Xiaorong Shi, and T G Nagaraja. A multiplex PCR procedure for the detection of six major virulence genes in escherichia coli O157:H7. *J. Microbiol. Methods*, 82(1):85–89, July 2010.

[23] Marie Ballif, Paul Harino, Serej Ley, Mireia Coscolla, Stefan Niemann, Robyn Carter, Christopher Coulter, Sonia Borrell, Peter Siba, Suparat Phuanukoonnon, Sebastien Gagneux, and Hans-Peter Beck. Drug resistance-conferring mutations in mycobacterium tuberculosis from madang, papua new guinea. *BMC Microbiol.*, 12:191, September 2012.

[24] David A Baltrus. Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.*, 28(8):489–495, August 2013.

[25] O Béjà, L Aravind, E V Koonin, M T Suzuki, A Hadd, L P Nguyen, S B Jovanovich, C M Gates, R A Feldman, J L Spudich, E N Spudich, and E F DeLong. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–1906, September 2000.

[26] M Belfort and R J Roberts. Homing endonucleases: keeping the house in order. *Nucleic Acids Res.*, 25(17):3379–3388, September 1997.

[27] Christopher M. Bellas, Alexandre M. Anesio, and Gary Barker. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Frontiers in Microbiology*, 6(JUL):656, 2015.

[28] Archana Belle, Markus Landthaler, and David A Shub. Intronless homing: site-specific endonuclease SegF of bacteriophage T4 mediates localized marker exclusion analogous to homing endonucleases of group I introns. *Genes Dev.*, 16(3):351–362, February 2002.

[29] Shellie R Bench, Thomas E Hanson, Kurt E Williamson, Dhritiman Ghosh, Mark Radosovich, Kui Wang, and K Eric Wommack. Metagenomic characterization of chesapeake bay virioplankton. *Appl. Environ. Microbiol.*, 73(23):7629–7641, December 2007.

[30] Dominic J. Bennett, Mark D. Sutton, and Samuel T. Turvey. treeman: an R package for efficient and intuitive manipulation of phylogenetic trees. *BMC Research Notes*, 10(1):30, January 2017.

[31] Michelle A Berry, Jeffrey D White, Timothy W Davis, Sunit Jain, Thomas H Johengen, Gregory J Dick, Orlando Sarnelle, and Vincent J Denef. Are oligotypes meaningful ecological and phylogenetic units? a case study of microcystis in freshwater lakes. *Front. Microbiol.*, 8:365, March 2017.

[32] Deborah A Berthold and Pål Stenmark. Membrane-bound diiron carboxylate proteins. *Annu. Rev. Plant Biol.*, 54:497–517, 2003.

[33] Yvan Bettarel, Thierry Bouvier, Corinne Bouvier, Claire Carré, Anne Desnues, Isabelle Domaizon, Stéphan Jacquet, Agnès Robin, and Télesphore Sime-Ngando. Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. *FEMS Microbiology Ecology*, 76(2):360–372, May 2011.

[34] Richa Bharti and Dominik G Grimm. Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform.*, 22(1):178–193, January 2021.

[35] Rie Dybboe Bjerre, Luisa Warchavchik Hugerth, Fredrik Boulund, Maike Seifert, Jeanne Duus Johansen, and Lars Engstrand. Effects of sampling strategy and DNA extraction on human skin microbiome investigations. *Sci. Rep.*, 9(1):17287, November 2019.

[36] K Björklöf, E L Nurmiaho-Lassila, N Klinger, K Haahtela, and M Romantschuk. Colonization strategies and conjugal gene transfer of inoculated pseudomonas syringae on the leaf surface. *J. Appl. Microbiol.*, 89(3):423–432, September 2000.

[37] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[38] 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, Yang Bai, Jordan E Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J Brislawn, C Titus Brown, Benjamin J Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K Cope, Ricardo Da Silva, Christian Diener, Pieter C Dorrestein, Gavin M Douglas, Daniel M Durall, Claire Duvallet, Christian F Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M Gauglitz, Sean M Gibbons, Deanna L Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A Huttley, Stefan Janssen, Alan K Jarmusch, Lingjing Jiang, Benjamin D Kaehler, Kyo Bin Kang, Christopher R Keefe, Paul Keim, Scott T Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G I Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D Martin, Daniel McDonald, Lauren J McIver, Alexey V Melnik, Jessica L Metcalf, Sydney C Morgan, Jamie T Morton, Ahmad Turan Naimey, Jose A Navas-Molina, Louis Felix Nothias, Stephanie B Orchanian, Talima Pearson, Samuel L Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S Robeson, 2nd, Patrick Rosenthal, Nicola Segata, Michael

Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R Spear, Austin D Swafford, Luke R Thompson, Pedro J Torres, Pauline Trinh, Anupriya Tripathi, Peter J Turnbaugh, Sabah Ul-Hasan, Justin J J van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C Weber, Charles H D Williamson, Amy D Willis, Zhenjiang Zech Xu, Jesse R Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J Gregory Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, 37(8):852–857, August 2019.

[39] P. Bork, C. Bowler, C. De Vargas, G. Gorsky, E. Karsenti, and P. Wincker. *Tara* Oceans studies plankton at Planetary scale. *Science*, 348(6237):873, 2015.

[40] P Bork and E V Koonin. Predicting functions from protein sequences–where are the bottlenecks? *Nat. Genet.*, 18(4):313–318, April 1998.

[41] M. Bostock, V. Ogievetsky, and J. Heer. D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[42] Patrick H Bradley, Stephen Nayfach, and Katherine S Pollard. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS Comput. Biol.*, 14(8):e1006242, August 2018.

[43] Lucas P P Braga, Chloé Orland, Erik J S Emilson, Amelia A Fitch, Helena Osterholz, Thorsten Dittmar, Nathan Basiliko, Nadia C S Mykytczuk, and Andrew J Tanentzap. Viruses direct carbon cycling in lake sediments under global change. *Proc. Natl. Acad. Sci. U. S. A.*, 119(41):e2202261119, October 2022.

[44] Cynthia Brewer, Mark Harrower, and The Pennsylvania State University. ColorBrewer2, 2013.

[45] Edward J Brignole, Nozomi Ando, Christina M Zimanyi, and Catherine L Drennan. The prototypic class ia ribonucleotide reductase from escherichia coli: still surprising after all these years. *Biochem. Soc. Trans.*, 40(3):523–530, June 2012.

[46] 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*, 11:2639, July 2017.

[47] Brian P Callahan, Matthew Stanger, and Marlene Belfort. A redox trap to augment the intein toolbox. *Biotechnol. Bioeng.*, 110(6):1565–1573, June 2013.

[48] Brian P Callahan, Natalya I Topilina, Matthew J Stanger, Patrick Van Roey, and Marlene Belfort. Structure of catalytically competent intein caught in a redox trap with functional and evolutionary implications. *Nat. Struct. Mol. Biol.*, 18(5):630–633, May 2011.

[49] Antonio Pedro Camargo, Stephen Nayfach, I-Min A Chen, Krishnaveni Palaniappan, Anna Ratner, Ken Chu, Stephan J Ritter, T B K Reddy, Supratim Mukherjee, Frederik Schulz, Lee Call, Russell Y Neches, Tanja Woyke, Natalia N Ivanova, Emiley A Eloe-Fadrosh, Nikos C Kyrpides, and Simon Roux. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.*, 51(D1):D733–D743, January 2023.

[50] Yuanpei Cao, Anru Zhang, and Hongzhe Li. Multi-sample estimation of bacterial composition matrix in metagenomics data. June 2017.

[51] Yuanpei Cao, Anru Zhang, and Hongzhe Li. Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*, 107(1):75–92, March 2020.

[52] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336, May 2010.

[53] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May 2010.

[54] Alfredo Caprioli, Stefano Morabito, Hubert Brugère, and Eric Oswald. Enterohaemorrhagic Escherichia coli: emerging issues on virulence and modes of transmission. *Veterinary Research*, 36(3):289–311, May 2005.

[55] P Carol and M Kuntz. A plastid terminal oxidase comes to light: implications for carotenoid biosynthesis and chlororespiration. *Trends Plant Sci.*, 6(1):31–36, January 2001.

[56] Natalia Cernicchiaro, Ana R S Oliveira, Allison Hoehn, Charley A Cull, Lance W Noll, Pragathi Belagola Shridhar, Tiruvoor G Nagaraja, Samuel E Ives, David G Renter, and Michael W Sanderson. Quantification of bacteria indicative of fecal and environmental contamination from hides to carcasses. *Foodborne Pathog. Dis.*, 16(12):844–855, December 2019.

[57] Wei-Hua Chen and Martin J. Lercher. ColorTree: a batch customization tool for phylogenic trees. *BMC Research Notes*, 2(1):155, 2009.

[58] Yingzhi Chen, Meng Zhang, Hongyue Jin, Yisi Tang, Huiyuan Wang, Qin Xu, Yaping Li, Feng Li, and Yongzhuo Huang. Intein-mediated site-specific synthesis of tumor-targeting protein delivery system: Turning PEG dilemma into prodrug-like feature. *Biomaterials*, 116:57–68, February 2017.

[59] Jessica Chopyk. *The influence of viral and bacterial community diversity on pathogenic E. coli prevalence in pre-harvest cattle*. PhD thesis, University of Delaware, Ann Arbor, United States, 2015.

[60] Jessica Chopyk, Sarah Allard, Daniel J Nasko, Anthony Bui, Emmanuel F Mongodin, and Amy R Sapkota. Agricultural freshwater pond supports diverse and dynamic bacterial and viral populations. *Front. Microbiol.*, 9:3489, April 2018.

[61] Jessica Chopyk, Ryan M. Moore, Zachary DiSpirito, Zachary R. Stromberg, Gentry L. Lewis, David G. Renter, Natalia Cernicchiaro, Rodney A. Moxley, and K. Eric Wommack. Presence of pathogenic Escherichia coli is correlated with bacterial community diversity and composition on pre-harvest cattle hides. *Microbiome*, 4(1):9, March 2016.

[62] Jessica Chopyk, Ryan M Moore, Zachary DiSpirito, Zachary R Stromberg, Gentry L Lewis, David G Renter, Natalia Cernicchiaro, Rodney A Moxley, and K Eric Wommack. Presence of pathogenic escherichia coli is correlated with bacterial community diversity and composition on pre-harvest cattle hides. *Microbiome*, 4:9, March 2016.

[63] Gustavo Felipe Correia Sales, Beatriz Ferreira Carvalho, Rosane Freitas Schwan, Leonardo de Figueiredo Vilela, Javier Andrés Moreno Meneses, Mateus Pies Gionbelli, and Carla Luiza da Silva Ávila. Heat stress influence the microbiota and organic acids concentration in beef cattle rumen. *J. Therm. Biol.*, 97:102897, April 2021.

[64] Paul I Costea, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, Marja Driessen, Rajna Hercog, Ferris-Elias Jung, Jens Roat Kultima, Matthew R Hayward, Luis Pedro Coelho, Emma Allen-Vercoe, Laurie Bertrand, Michael Blaut, Jillian R M Brown, Thomas Carton, Stéphanie Cools-Portier, Michelle Daigneault, Muriel Derrien, Anne Druesne, Willem M de Vos, B Brett Finlay, Harry J Flint, Francisco Guarner, Masahira Hattori, Hans Heilig, Ruth Ann Luna, Johan van Hylckama Vlieg, Jana Junick, Ingeborg Klymiuk, Philippe Langella, Emmanuelle Le Chatelier, Volker Mai, Chaysavanh Manichanh, Jennifer C Martin, Clémentine Mery, Hidetoshi Morita, Paul W O'Toole, Céline Orvain, Kiran Raosaheb Patil, John Penders, Søren Persson, Nicolas Pons, Milena Popova, Anne Salonen, Delphine Saulnier,

Karen P Scott, Bhagirath Singh, Kathleen Slezak, Patrick Veiga, James Versalovic, Liping Zhao, Erwin G Zoetendal, S Dusko Ehrlich, Joel Dore, and Peer Bork. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.*, 35(11):1069–1076, November 2017.

[65] Joseph A. Cotruvo and JoAnne Stubbe. Class I Ribonucleotide Reductases: Metallocofactor Assembly and Repair In Vitro and In Vivo. *Annual Review of Biochemistry*, 80(1):733–767, June 2011.

[66] Joseph A Cotruvo and Joanne Stubbe. Class I ribonucleotide reductases: metallocofactor assembly and repair in vitro and in vivo. *Annu. Rev. Biochem.*, 80:733–767, 2011.

[67] Felipe H. Coutinho, Cynthia B. Silveira, Gustavo B. Gregoracci, Cristiane C. Thompson, Robert A. Edwards, Corina P. D. Brussaard, Bas E. Dutilh, and Fabiano L. Thompson. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*, 8(May):1–12, 2017.

[68] Alexander I Culley, Brenda F Asuncion, and Grieg F Steward. Detection of inteins among diverse DNA polymerase genes of uncultivated members of the phycodnaviridae. *ISME J.*, 3(4):409–418, April 2009.

[69] Lucas Czech, Jaime Huerta-Cepas, and Alexandros Stamatakis. A Critical Review on the Use of Support Values in Tree Viewers and Bioinformatics Toolkits. *Molecular Biology and Evolution*, 34(6):1535–1542, March 2017.

[70] Lucas Czech, Jaime Huerta-Cepas, and Alexandros Stamatakis. A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Mol. Biol. Evol.*, 34(6):1535–1542, June 2017.

[71] Aguirre de Cárcer Daniel, Carlos Pedrós-Alió, David A Pearce, and Antonio Alcamí. Composition and Interactions among Bacterial, Microeukaryotic, and T4-like Viral Assemblages in Lakes from Both Polar Zones. *Frontiers in microbiology*, 7:337–337, March 2016.

[72] Koen A L De Smet, Karen E Kempsell, Alex Gallagher, Ken Duncan, and Douglas B Young. Alteration of a single amino acid residue reverses fosfomycin resistance of recombinant MurA from mycobacterium tuberculosis the EMBL accession number for the sequence in this paper is X96711. *Microbiology*, 145(11):3177–3184, November 1999.

[73] Rutger de Wit and Thierry Bouvier. 'everything is everywhere, but, the environment selects'; what did baas becking and beijerinck really say? *Environ. Microbiol.*, 8(4):755–758, April 2006.

[74] Glenn De'ath. The multinomial diversity model: linking shannon diversity to multiple predictors. *Ecology*, 93(10):2286–2296, October 2012.

[75] Christoph M Deeg, Cheryl-Emiliane T Chow, and Curtis A Suttle. The kinetoplastid-infecting bodo saltans virus (BsV), a window into the most abundant giant viruses in the sea. *Elife*, 7, March 2018.

[76] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via theEMAlgorithm. *J. R. Stat. Soc.*, 39(1):1–22, September 1977.

[77] Diana M A Dewsbury, David G Renter, Pragathi B Shridhar, Lance W Noll, Xiaorong Shi, Tiruvoor G Nagaraja, and Natalia Cernicchiaro. Summer and winter prevalence of shiga Toxin-Producing escherichia coli (STEC) o26, o45, o103, o111, o121, o145, and O157 in feces of feedlot cattle. *Foodborne Pathog. Dis.*, 12(8):726–732, August 2015.

[78] Gregory J Dick. The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. *Nat. Rev. Microbiol.*, 17(5):271–283, May 2019.

[79] Bas E Dutilh, Arvind Varsani, Yigang Tong, Peter Simmonds, Sead Sabanadzovic, Luisa Rubino, Simon Roux, Alejandro Reyes Muñoz, Cédric Lood, Elliot J Lefkowitz, Jens H Kuhn, Mart Krupovic, Robert A Edwards, J Rodney Brister, Evelien M Adriaenssens, and Matthew B Sullivan. Perspective on taxonomic classification of uncultivated viruses. *Curr. Opin. Virol.*, 51:207–215, December 2021.

[80] Bhakti Dwivedi, Bingjie Xue, Daniel Lundin, Robert A Edwards, and Mya Breitbart. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evol. Biol.*, 13:33, February 2013.

[81] Bhakti Dwivedi, Bingjie Xue, Daniel Lundin, Robert A. Edwards, and Mya Breitbart. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evolutionary Biology*, 13(1):33, 2013.

[82] Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, October 2011.

[83] Robert A. Edwards, Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E. Dutilh. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews*, 40(2):258–272, 2016.

[84] H Eklund, U Uhlin, M Färnegårdh, D T Logan, and P Nordlund. Structure and function of the radical enzyme ribonucleotide reductase. *Prog. Biophys. Mol. Biol.*, 77(3):177–268, November 2001.

[85] Skander Elleuche and Stefanie Pöggeler. Inteins, valuable genetic elements in molecular biology and biotechnology. *Appl. Microbiol. Biotechnol.*, 87(2):479–489, June 2010.

[86] Mostafa S Elshahed, Noha H Youssef, Anne M Spain, Cody Sheik, Fares Z Najar, Leonid O Sukharnikov, Bruce A Roe, James P Davis, Patrick D Schloss, Vanessa L Bailey, and Lee R Krumholz. Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl. Environ. Microbiol.*, 74(17):5422–5428, September 2008.

[87] Joanne B. Emerson, Brian C. Thomas, Karen Andrade, Karla B. Heidelberg, and Jillian F. Banfield. New Approaches Indicate Constant Viral Diversity despite Shifts in Assemblage Structure in an Australian Hypersaline Lake. *Applied and Environmental Microbiology*, 79(21):6755, November 2013.

[88] Ertan Eryilmaz, Neel H Shah, Tom W Muir, and David Cowburn. Structural and dynamical features of inteins and implications on protein splicing. *J. Biol. Chem.*, 289(21):14506–14511, May 2014.

[89] N J Evans, R D Murray, and S D Carter. Bovine digital dermatitis: Current concepts from laboratory to farm. *Vet. J.*, 211:3–13, May 2016.

[90] T C Evans, Jr, D Martin, R Kolly, D Panne, L Sun, I Ghosh, L Chen, J Benner, X Q Liu, and M Q Xu. Protein trans-splicing and cyclization by a naturally split intein from the dnae gene of synechocystis species PCC6803. *J. Biol. Chem.*, 275(13):9091–9094, March 2000.

[91] Douglas W. Fadrosh, Bing Ma, Pawel Gajer, Naomi Sengamalay, Sandra Ott, Rebecca M. Brotman, and Jacques Ravel. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, 2(1):6, 2014.

[92] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.*, 10(8):538–550, July 2012.

[93] Thomas Ferenci. Irregularities in genetic variation and mutation rates with environmental stresses. *Environ. Microbiol.*, 21(11):3979–3988, November 2019.

[94] Andrew D Fernandes, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One*, 8(7):e67019, July 2013.

[95] Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, May 2014.

[96] Lucía Fernández, Ana Rodríguez, and Pilar García. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J.*, 12(5):1171–1179, May 2018.

[97] Jacquelyn S Fetrow. Active site profiling to identify protein functional sites in sequences and structures using the deacon active site profiler (DASP). *Curr. Protoc. Bioinformatics*, 14(1):8.10.1–8.10.16, 2006.

[98] Jonathan Filée. Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. *Curr. Opin. Virol.*, 33:81–88, December 2018.

[99] Jan F. Finke and Curtis A. Suttle. The Environment and Cyanophage Diversity: Insights From Environmental Sequencing of DNA Polymerase. *Frontiers in Microbiology*, 10:167, 2019.

[100] Omri M Finkel, Adrien Y Burch, Steven E Lindow, Anton F Post, and Shimshon Belkin. Geographical location determines the population structure in phyllosphere microbial communities of a salt-excreting desert tree. *Appl. Environ. Microbiol.*, 77(21):7647–7655, November 2011.

[101] Robert D Finn, Teresa K Attwood, Patricia C Babbitt, Alex Bateman, Peer Bork, Alan J Bridge, Hsin-Yu Chang, Zsuzsanna Dosztányi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A Natale, Marco Necci, Gift Nuka, Christine A Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C Potter, Neil D Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Ioannis Xenarios, Lai-Su Yeh, Siew-Yit Young, and Alex L Mitchell. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, 45(D1):D190–D199, January 2017.

[102] Fatima Foflonker, Dana C Price, Huan Qiu, Brian Palenik, Shuyi Wang, and Debashish Bhattacharya. Genome of the halotolerant green alga picochlorum sp. reveals strategies for thriving under fluctuating environmental conditions. *Environ. Microbiol.*, 17(2):412–426, February 2015.

[103] M Fontecave, R Eliasson, and P Reichard. Oxygen-sensitive ribonucleoside triphosphate reductase is present in anaerobic escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, 86(7):2147–2151, April 1989.

[104] Kristina Friedel, Monika A Popp, Julian C J Matern, Emerich M Gazdag, Ilka V Thiel, Gerrit Volkmann, Wulf Blankenfeldt, and Henning D Mootz. A functional interplay between intein and extein sequences in protein splicing compensates for the essential block B histidine. *Chem. Sci.*, 10(1):239–251, January 2019.

[105] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, 8(9):e1002687, September 2012.

[106] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, December 2012.

[107] Clara A Fuchsman, Roy Eric Collins, Gabrielle Rocap, and William J Brazelton. Effect of the environment on horizontal gene transfer between bacteria and archaea. *PeerJ*, 5:e3865, September 2017.

[108] M Stanley Fujimoto, Anton Suvorov, Nicholas O Jensen, Mark J Clement, and Seth M Bybee. Detecting false positive sequence homology: a machine learning approach. *BMC Bioinformatics*, 17:101, February 2016.

[109] M Stanley Fujimoto, Anton Suvorov, Nicholas O Jensen, Mark J Clement, Quinn Snell, and Seth M Bybee. The OGCleaner: filtering false-positive homology clusters. *Bioinformatics*, 33(1):125–127, January 2017.

[110] E Gernand, S König, and C Kipp. Influence of on-farm measurements for heat stress indicators on dairy cow productivity, female fertility, and health. *J. Dairy Sci.*, 102(7):6660–6671, July 2019.

[111] Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S Cooper, Daniel S Brewer, Mihaela Pertea, and Steven L Salzberg. Major data analysis errors invalidate cancer microbiome findings. *MBio*, page e0160723, October 2023.

[112] Walter R Gilks, Benjamin Audit, Daniela De Angelis, Sophia Tsoka, and Christos A Ouzounis. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18(12):1641–1649, December 2002.

[113] Stephen J Giovannoni, H James Tripp, Scott Givan, Mircea Podar, Kevin L Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H Richardson, Michiel Noordewier, Michael S Rappé, Jay M Short, James C Carrington, and Eric J Mathur. Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245, August 2005.

[114] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.*, 8:2224, November 2017.

[115] Gregory Brian Gloor, Jean M Macklaim, Michael Vu, and Andrew D Fernandes. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *AJS*, 45(4):73–87, July 2016.

[116] J Peter Gogarten and Elena Hilario. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol. Biol.*, 6:94, November 2006.

[117] J Peter Gogarten, Alireza G Senejani, Olga Zhaxybayeva, Lorraine Olendzenski, and Elena Hilario. Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.*, 56:263–287, January 2002.

[118] H Goodrich-Blair and D A Shub. Beyond homing: competition between intron endonucleases confers a selective advantage on flanking genetic markers. *Cell*, 84(2):211–221, January 1996.

[119] Uri Gophna and Neta Altman-Price. Horizontal gene transfer in Archaea-From mechanisms to genome evolution. *Annu. Rev. Microbiol.*, 76:481–502, September 2022.

[120] Matti Gralka. Searching for principles of microbial ecology across levels of biological organization. *Integr. Comp. Biol.*, June 2023.

[121] Cathleen M Green, Olga Novikova, and Marlene Belfort. The dynamic intein landscape of eukaryotes. *Mob. DNA*, 9:4, January 2018.

[122] D. A. Green. A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India*, 39(2):289–295, 2011.

[123] Michael Greenacre, Marina Martínez-Álvaro, and Agustín Blasco. Compositional data analysis of microbiome and Any-Omics datasets: A validation of the additive logratio transformation. *Front. Microbiol.*, 12:727398, October 2021.

[124] Ann C Gregory, Ahmed A Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, Ksenia Arkhipova, Margaux Carmichael, Corinne Cruaud, Céline Dimier, Guillermo Domínguez-Huerta, Joannie Ferland, Stefanie Kandels, Yunxiao Liu, Claudie Marec, Stéphane Pesant, Marc Picheral, Sergey Pisarev, Julie Poulain, Jean-Éric Tremblay, Dean Vik, Tara Oceans Coordinators, Marcel Babin, Chris Bowler, Alexander I Culley, Colomban de Vargas, Bas E Dutilh, Daniele Iudicone, Lee Karp-Boss, Simon Roux, Shinichi Sunagawa, Patrick Wincker, and Matthew B Sullivan. Marine DNA viral macro- and microdiversity from pole to pole. *Cell*, 177(5):1109–1123.e14, May 2019.

[125] Hans-Peter Grossart, Ramon Massana, Katherine D McMahon, and David A Walsh. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnol. Oceanogr.*, 65(S1), January 2020.

[126] Mathieu Groussin, Mathilde Poyet, Ainara Sistiaga, Sean M Kearney, Katya Moniz, Mary Noel, Jeff Hooker, Sean M Gibbons, Laure Segurel, Alain Froment,

Rihlat Said Mohamed, Alain Fezeu, Vanessa A Juimo, Sophie Lafosse, Francis E Tabe, Catherine Girard, Deborah Iqaluk, Le Thanh Tu Nguyen, B Jesse Shapiro, Jenni Lehtimäki, Lasse Ruokolainen, Pinja P Kettunen, Tommi Vatanen, Shani Sigwazi, Audax Mabulla, Manuel Domínguez-Rodrigo, Yvonne A Nartey, Adwoa Agyei-Nkansah, Amoako Duah, Yaw A Awuku, Kenneth A Valles, Shadrack O Asibey, Mary Y Afihene, Lewis R Roberts, Amelie Plymoth, Charles A Onyekwere, Roger E Summons, Ramnik J Xavier, and Eric J Alm. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*, 184(8):2053–2067.e18, April 2021.

[127] Fengfei Gu, Senlin Zhu, Jinxiu Hou, Yifan Tang, Jian-Xin Liu, Qingbiao Xu, and Hui-Zeng Sun. The hindgut microbiome contributes to host oxidative stress in postpartum dairy cows by affecting glutathione synthesis process. *Microbiome*, 11(1):87, April 2023.

[128] Lei Han, Junsheng Chen, Kai Ding, Huifang Zong, Yueqing Xie, Hua Jiang, Baohong Zhang, Huili Lu, Weihan Yin, John Gilly, and Jianwei Zhu. Efficient generation of bispecific IgG antibodies by split intein mediated protein trans-splicing system. *Sci. Rep.*, 7(1):8360, August 2017.

[129] Mira V. Han and Christian M. Zmasek. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10(1):356, October 2009.

[130] D. D. Hancock, T. E. Besser, M. L. Kinsel, D. H. Tarr, P. I .and Rice, and M. G. Paros. The prevalence of Escherichia coli O157.H7 in dairy and beef cattle in Washington State. *Epidemiology and Infection*, 113(2):199–207, October 1994.

[131] J Handelsman, M R Rondon, S F Brady, J Clardy, and R M Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.*, 5(10):R245–9, October 1998.

[132] Victor Hanson-Smith and Alexander Johnson. PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories. *PLOS Computational Biology*, 12(7):1–10, October 2016.

[133] Angela F Harper, Janelle B Leuthaeuser, Patricia C Babbitt, John H Morris, Thomas E Ferrin, Leslie B Poole, and Jacquelyn S Fetrow. An atlas of peroxiredoxins created using an active site Profile-Based approach to functionally relevant clustering of proteins. *PLoS Comput. Biol.*, 13(2):e1005284, February 2017.

[134] Amelia O Harrison. Ribonucleotide reductase genes influence the biology and ecology of marine viruses. Master's thesis, University of Delaware, 2019.

[135] Amelia O Harrison, Ryan M Moore, Shawn W Polson, and K Eric Wommack. Reannotation of the ribonucleotide reductase in a cyanophage reveals life history strategies within the virioplankton. *Front. Microbiol.*, 10:134, February 2019.

[136] Amelia O. Harrison, Ryan M. Moore, Shawn W. Polson, and K. Eric Wommack. Reannotation of the Ribonucleotide Reductase in a Cyanophage Reveals Life History Strategies Within the Virioplankton. *Frontiers in Microbiology*, 10:134, 2019.

[137] Shaomei He, Maximilian P Lau, Alexandra M Linz, Eric E Roden, and Katherine D McMahon. Extracellular electron transfer may be an overlooked contribution to pelagic respiration in Humic-Rich freshwater lakes. *mSphere*, 4(1), January 2019.

[138] Zilong He, Huangkai Zhang, Shenghan Gao, Martin J. Lercher, Wei Hua Chen, and Songnian Hu. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research*, 44(W1):W236–W241, July 2016.

[139] M O Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[140] S S Hirano and C D Upper. Bacteria in the leaf ecosystem with emphasis on pseudomonas syringae-a pathogen, ice nucleus, and epiphyte. *Microbiol. Mol. Biol. Rev.*, 64(3):624–653, September 2000.

[141] Simon Hoffmann, Tobias M E Terhorst, Rohit K Singh, Daniel Kümmel, Shmuel Pietrokovski, and Henning D Mootz. Biochemical and structural characterization of an unusual and naturally split class 3 intein. *Chembiochem*, 22(2):364–373, January 2021.

[142] T C Hsieh, K H Ma, and Anne Chao. iNEXT: an R package for rarefaction and extrapolation of species diversity (H ill numbers). *Methods Ecol. Evol.*, 7(12):1451–1456, December 2016.

[143] Sijun Huang, Steven W Wilhelm, Nianzhi Jiao, and Feng Chen. Ubiquitous cyanobacterial podoviruses in the global oceans unveiled through viral DNA polymerase gene sequences. *ISME J.*, 4(10):1243–1251, October 2010.

[144] Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, 33(6):1635–1638, June 2016.

[145] Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638, 2016.

[146] Michelle T Hulin, Mojgan Rabiey, Ziyue Zeng, Andrea Vadillo Dieguez, Sophia Bellamy, Phoebe Swift, John W Mansfield, Robert W Jackson, and Richard J Harrison. Genomic and functional analysis of phage-mediated horizontal gene transfer in pseudomonas syringae on the plant surface. *New Phytol.*, 237(3):959–973, February 2023.

[147] Daniel H. Huson, Daniel C. Richter, Christian Rausch, Tobias Dezulian, Markus Franz, and Regula Rupp. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8(1):460, November 2007.

[148] Marion Hutinel, Jerker Fick, D G Joakim Larsson, and Carl-Fredrik Flach. Investigating the effects of municipal and hospital wastewaters on horizontal gene transfer. *Environ. Pollut.*, 276:116733, May 2021.

[149] Jaime Iranzo, Mart Krupovic, and Eugene V Koonin. The Double-Stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio*, 7(4), August 2016.

[150] Thomas A Isenbarger, Christopher E Carr, Sarah Stewart Johnson, Michael Finney, George M Church, Walter Gilbert, Maria T Zuber, and Gary Ruvkun. The most conserved genome segments for life detection on earth and other planets. *Orig. Life Evol. Biosph.*, 38(6):517–533, December 2008.

[151] Natalia Ivanova, Susannah G Tringe, Konstantinos Liolios, Wen-Tso Liu, Norman Morrison, Philip Hugenholtz, and Nikos C Kyrpides. A call for standardized classification of metagenome projects. *Environ. Microbiol.*, 12(7):1803–1805, July 2010.

[152] Justin Jee, Aviram Rasouly, Ilya Shamovsky, Yonatan Akivis, Susan R Steinman, Bud Mishra, and Evgeny Nudler. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature*, 534(7609):693–696, June 2016.

[153] Seth G John, Carolina B Mendez, Li Deng, Bonnie Poulos, Anne Kathryn M Kauffman, Suzanne Kern, Jennifer Brum, Martin F Polz, Edward A Boyle, and Matthew B Sullivan. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.*, 3(2):195–202, April 2011.

[154] Erica S Johnson. Protein modification by SUMO. *Annu. Rev. Biochem.*, 73:355–382, 2004.

[155] Craig E Jones, Alfred L Brown, and Ute Baumann. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8:170, May 2007.

[156] A Jordan and P Reichard. Ribonucleotide reductases. *Annu. Rev. Biochem.*, 67:71–98, 1998.

[157] Alexandre Jousset, Christina Bienhold, Antonis Chatzinotas, Laure Gallien, Angélique Gobet, Viola Kurm, Kirsten Küsel, Matthias C Rillig, Damian W Rivett, Joana F Salles, Marcel G A van der Heijden, Noha H Youssef, Xiaowei Zhang, Zhong Wei, and W H Gera Hol. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J.*, 11(4):853–862, April 2017.

[158] Pedro C. Junger, André M. Amado, Rodolfo Paranhos, Anderson S. Cabral, Saulo M. S. Jacques, and Vinicius F. Farjalla. Salinity Drives the Virioplankton Abundance but Not Production in Tropical Coastal Lagoons. *Microbial Ecology*, 75(1):52–63, January 2018.

[159] Minoru Kanehisa, Yoko Sato, and Kanae Morishima. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.*, 428(4):726–731, February 2016.

[160] Jonathan Kans. *Entrez Direct: E-utilities on the UNIX Command Line*. National Center for Biotechnology Information (US), May 2020.

[161] Antti Karkman, Thi Thuy Do, Fiona Walsh, and Marko P J Virta. Antibiotic-Resistance genes in waste water. *Trends Microbiol.*, 26(3):220–228, March 2018.

[162] S Karlin and S F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.*, 87(6):2264–2268, March 1990.

[163] Alex Kasrayan, Annika L Persson, Margareta Sahlin, and Britt-Marie Sjoberg. The conserved active site asparagine in class I ribonucleotide reductase is essential for catalysis. *J. Biol. Chem.*, 277(8):5749–5755, February 2002.

[164] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, April 2013.

[165] 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.

[166] Swati Kaushik, Anu G Nair, Eshita Mutt, Hari Prasanna Subramanian, and Ramanathan Sowdhamini. Rapid and enhanced remote homology detection by cascading hidden markov model searches in sequence space. *Bioinformatics*, 32(3):338–344, February 2016.

[167] Danielle S Kelley, Christopher W Lennon, SEA-PHAGES, Marlene Belfort, and Olga Novikova. Mycobacteriophages as incubators for intein dissemination and evolution. *MBio*, 7(5), October 2016.

[168] Rachel A Keown, Jacob T Dums, Phillip J Brumm, Joyanne MacDonald, David A Mead, Barbra D Ferrell, Ryan M Moore, Amelia O Harrison, Shawn W Polson, and K Eric Wommack. Novel viral DNA polymerases from metagenomes suggest genomic sources of Strand-Displacing biochemical phenotypes. *Front. Microbiol.*, 13:858366, April 2022.

[169] R Kindt, P Van Damme, and A J Simons. Tree diversity in western kenya: Using profiles to characterise richness and evenness. *Biodiversity & Conservation*, 15(4):1253–1270, April 2006.

[170] Claudia Knief, Nathanaël Delmotte, Samuel Chaffron, Manuel Stark, Gerd Innerebner, Reiner Wassmann, Christian von Mering, and Julia A Vorholt. Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J.*, 6(7):1378–1390, July 2012.

[171] Stacy T Knutson, Brian M Westwood, Janelle B Leuthaeuser, Brandon E Turner, Don Nguyendac, Gabrielle Shea, Kiran Kumar, Julia D Hayden, Angela F Harper, Shoshana D Brown, John H Morris, Thomas E Ferrin, Patricia C Babbitt, and Jacquelyn S Fetrow. An approach to functionally relevant clustering of the protein universe: Active site profile-based clustering of protein structures and sequences: Functionally relevant clustering of protein superfamilies. *Protein Sci.*, 26(4):677–699, April 2017.

[172] Alexander Koeppel, Elizabeth B Perry, Johannes Sikorski, Danny Krizanc, Andrew Warner, David M Ward, Alejandro P Rooney, Evelyne Brambilla, Nora Connor, Rodney M Ratcliff, Eviatar Nevo, and Frederick M Cohan. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc. Natl. Acad. Sci. U. S. A.*, 105(7):2504–2509, February 2008.

[173] Matthias Kolberg, Kari R Strand, Pål Graff, and K Kristoffer Andersson. Structure, function, and mechanism of ribonucleotide reductases. *Biochim. Biophys. Acta*, 1699(1-2):1–34, June 2004.

[174] Matthias Kolberg, Kari R Strand, Pål Graff, and K Kristoffer Andersson. Structure, function, and mechanism of ribonucleotide reductases. *Biochim. Biophys. Acta*, 1699(1-2):1–34, June 2004.

[175] Matthias Kolberg, Kari R Strand, Pål Graff, and K Kristoffer Andersson. Structure, function, and mechanism of ribonucleotide reductases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1699(1):1–34, June 2004.

[176] Johannes Köster. Rust-Bio: a fast and safe bioinformatics library. *Bioinformatics*, 32(3):444–446, February 2016.

[177] Lukasz Kreft, Alexander Botzki, Frederik Coppens, Klaas Vandepoele, and Michiel Van Bel. PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, 33(18):2946–2947, 2017.

[178] Justin Kuczynski, Jesse Stombaugh, William Anton Walters, Antonio González, J Gregory Caporaso, and Rob Knight. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr. Protoc. Microbiol.*, Chapter 1:Unit 1E.5., November 2012.

[179] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.*, 11(5):e1004226, May 2015.

[180] Joshua Ladau, Thomas J Sharpton, Mariel M Finucane, Guillaume Jospin, Steven W Kembel, James O'Dwyer, Alexander F Koeppel, Jessica L Green, and Katherine S Pollard. Global marine bacterial diversity peaks at high latitudes in winter. *The ISME Journal*, 7:1669, March 2013.

[181] Yemin Lan, Gail Rosen, and Ruth Hershberg. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome*, 4(1):18, 2016.

[182] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

[183] Alyse A. Larkin, Sara K. Blinebry, Caroline Howes, Yajuan Lin, Sarah E. Loftus, Carrie A. Schmaus, Erik R. Zinser, and Zackary I. Johnson. Niche partitioning and biogeography of high light adapted Prochlorococcus across taxonomic ranks in the North Pacific. *The ISME Journal*, 10:1555–1567, January 2016.

[184] Fabien Lauer. MLweb: A toolkit for machine learning on the web. *Neurocomputing*, 282:74–77, 2017.

[185] Pierre Legendre and Louis Legendre. *Numerical ecology*, volume 24 of *Developments in environmental modelling*. Elsevier, 2012.

[186] Philippe Lemanceau, Manuel Blouin, Daniel Muller, and Yvan Moënne-Loccoz. Let the core microbiota be functional. *Trends Plant Sci.*, 22(7):583–595, July 2017.

[187] David Lembo, Manuela Donalisio, Anders Hofer, Maura Cornaglia, Wolfram Brune, Ulrich Koszinowski, Lars Thelander, and Santo Landolfo. The ribonucleotide reductase R1 homolog of murine cytomegalovirus is not a functional enzyme subunit but is required for pathogenesis. *J. Virol.*, 78(8):4278–4288, April 2004.

[188] Michael Lemke and Rob DeSalle. The next generation of microbial ecology and its importance in environmental sustainability. *Microb. Ecol.*, 85(3):781–795, April 2023.

[189] Christopher W Lennon and Marlene Belfort. Inteins. *Curr. Biol.*, 27(6):R204–R206, March 2017.

[190] Christopher W Lennon, Matthew Stanger, Nilesh K Banavali, and Marlene Belfort. Conditional protein splicing switch in hyperthermophiles through an Intein-Extein partnership. *MBio*, 9(1), January 2018.

[191] Christopher W Lennon, Matthew Stanger, and Marlene Belfort. Protein splicing of a recombinase intein induced by ssDNA and DNA damage. *Genes Dev.*, 30(24):2663–2668, December 2016.

[192] Christopher W Lennon, Matthew J Stanger, and Marlene Belfort. Mechanism of Single-Stranded DNA activation of recombinase intein splicing. *Biochemistry*, 58(31):3335–3339, August 2019.

[193] Xavier Leroy, Damien Doligez, Alain Frisch, Jacques Garrigue, Didier Rémy, and Jérôme Vouillon. The OCaml system, release 4.14. https://v2.ocaml.org/manual/, 2022. Accessed: 2022-5-25.

[194] Ivica Letunic and Peer Bork. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 44(W1):W242–W245, 2016.

[195] Ivica Letunic and Peer Bork. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, 47(W1):W256–W259, July 2019.

[196] Janelle B Leuthaeuser, John H Morris, Angela F Harper, Thomas E Ferrin, Patricia C Babbitt, and Jacquelyn S Fetrow. DASP3: identification of protein sequences belonging to functionally relevant groups. *BMC Bioinformatics*, 17(1):458, November 2016.

[197] Hongzhe Li. Microbiome, metagenomics, and High-Dimensional compositional data analysis. *Annu. Rev. Stat. Appl.*, 2(1):73–94, April 2015.

[198] Liangzhi Li, Shuguang Peng, Zhenhua Wang, Teng Zhang, Hongguang Li, Yansong Xiao, Jingjun Li, Yongjun Liu, and Huaqun Yin. Genome mining reveals abiotic stress resistance genes in plant genomes acquired from microbes via HGT. *Front. Plant Sci.*, 13:1025122, November 2022.

[199] Meina Li, Lijun Cao, Musoki Mwimba, Yan Zhou, Ling Li, Mian Zhou, Patrick S Schnable, Jamie A O'Rourke, Xinnian Dong, and Wei Wang. Comprehensive mapping of abiotic stress inputs into the soybean circadian clock. *Proc. Natl. Acad. Sci. U. S. A.*, 116(47):23840–23849, November 2019.

[200] Meng Li, Jinjie Zhao, Nianwu Tang, Hang Sun, and Jinling Huang. Horizontal gene transfer from bacteria and plants to the arbuscular mycorrhizal fungus rhizophagus irregularis. *Front. Plant Sci.*, 9:701, May 2018.

[201] Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.*, 11(1):3514, July 2020.

[202] David Lloyd, Alan Chapman, Jayne E Ellis, Kevin Hillman, Timothy A Paget, Nigel Yarlett, and Alan G Williams. Chapter five - oxygen levels are key to understanding "anaerobic" protozoan pathogens with micro-aerophilic lifestyles. In Robert K Poole and David J Kelly, editors, *Advances in Microbial Physiology*, volume 79, pages 163–240. Academic Press, January 2021.

[203] Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G Graeber, A Brantley Hall, Kathleen Lake, Carol J Landers, Himel Mallick, Damian R Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A White, 3rd, IBDMDB Investigators, Jonathan Braun, Lee A Denson, Janet K Jansson, Rob Knight, Subra Kugathasan, Dermot P B McGovern, Joseph F Petrosino, Thaddeus S Stappenbeck, Harland S Winter, Clary B Clish, Eric A Franzosa, Hera Vlamakis, Ramnik J Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019.

[204] Briallen Lobb, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C Doxey. An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom*, 6(3), March 2020.

[205] Christoph Loderer, Venkateswara Rao Jonna, Mikael Crona, Inna Rozman Grinberg, Margareta Sahlin, Anders Hofer, Daniel Lundin, and Britt-Marie Sjöberg. A unique cysteine-rich zinc finger domain present in a majority of class II ribonucleotide reductases mediates catalytic turnover. *J. Biol. Chem.*, 292(46):19044–19054, November 2017.

[206] Christoph Loderer, Venkateswara Rao Jonna, Mikael Crona, Inna Rozman Grinberg, Margareta Sahlin, Anders Hofer, Daniel Lundin, and Britt-Marie Sjöberg. A unique cysteine-rich zinc finger domain present in a majority of class II ribonucleotide reductases mediates catalytic turnover. *J. Biol. Chem.*, 292(46):19044–19054, November 2017.

[207] Yaniv Loewenstein, Domenico Raimondo, Oliver C Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, and Anna Tramontano. Protein function annotation by homology-based inference. *Genome Biol.*, 10(2):207, February 2009.

[208] Rocío López-Igual, Joaquín Bernal-Bayard, Alfonso Rodríguez-Patón, Jean-Marc Ghigo, and Didier Mazel. Engineered toxin-intein antimicrobials can selectively target and kill antibiotic-resistant bacteria in mixed populations. *Nat. Biotechnol.*, 37(7):755–760, July 2019.

[209] Mario López-Pérez, Jose M Haro-Moreno, Rafael Gonzalez-Serrano, Marcos Parras-Moltó, and Francisco Rodriguez-Valera. Genome diversity of marine phages recovered from mediterranean metagenomes: Size matters. *PLoS Genet.*, 13(9):e1007018, September 2017.

[210] C. Lozupone and R. Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, December 2005.

[211] Sugnet Lubbe, Peter Filzmoser, and Matthias Templ. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics Intellig. Lab. Syst.*, 210:104248, March 2021.

[212] Daniel Lundin, Gustav Berggren, Derek T Logan, and Britt-Marie Sjöberg. The origin and evolution of ribonucleotide reduction. *Life*, 5(1):604–636, February 2015.

[213] Daniel Lundin, Gustav Berggren, Derek T Logan, and Britt-Marie Sjöberg. The origin and evolution of ribonucleotide reduction. *Life*, 5(1):604–636, February 2015.

[214] Daniel Lundin, Eduard Torrents, Anthony M Poole, and Britt-Marie Sjöberg. RNRdb, a curated database of the universal enzyme family ribonucleotide reductase, reveals a high level of misannotation in sequences deposited to genbank. *BMC Genomics*, 10(1):589, 2009.

[215] Michael D J Lynch and Josh D Neufeld. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.*, 13(4):217–229, April 2015.

[216] Bin Ma, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, Xiaofei Lv, Ronghui Pan, Jiabao Zhang, Yongguan Zhu, and Jianming Xu. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.*, 14(1):7318, November 2023.

[217] Yannick Mahlich, Martin Steinegger, Burkhard Rost, and Yana Bromberg. HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics*, 34(13):i304–i312, July 2018.

[218] S S Mao, T P Holler, G X Yu, J M Bollinger, Jr, S Booker, M I Johnston, and J Stubbe. A model for the role of multiple cysteine residues involved in ribonucleotide reduction: amazing and still confusing. *Biochemistry*, 31(40):9733–9743, October 1992.

[219] S S Mao, G X Yu, D Chalfoun, and J Stubbe. Characterization of C439SR1, a mutant of escherichia coli ribonucleotide diphosphate reductase: evidence that C439 is a residue essential for nucleotide reduction and C439SR1 is a protein possessing novel thioredoxin-like activity. *Biochemistry*, 31(40):9752–9759, October 1992.

[220] Shengyong Mao, Mengling Zhang, Junhua Liu, and Weiyun Zhu. Characterising the bacterial microbiota across the gastrointestinal tracts of dairy cattle: membership and potential function. *Scientific Reports*, 5:16116, November 2015.

[221] Aron Marchler-Bauer, Yu Bo, Lianyi Han, Jane He, Christopher J Lanczycki, Shennan Lu, Farideh Chitsaz, Myra K Derbyshire, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, Fu Lu, Gabriele H Marchler, James S Song, Narmada Thanki, Zhouxi Wang, Roxanne A Yamashita, Dachuan Zhang, Chanjuan Zheng, Lewis Y Geer, and Stephen H Bryant. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.*, 45(D1):D200–D203, January 2017.

[222] Aron Marchler-Bauer, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Marina V Omelchenko, Cynthia L Robertson, James S Song, Narmada Thanki, Roxanne A Yamashita, Dachuan Zhang, Naigong Zhang, Chanjuan Zheng, and Stephen H Bryant. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.*, 39(Database issue):D225–9, January 2011.

[223] Rachel Marine, Coleen McCarren, Vansay Vorrasane, Dan Nasko, Erin Crowgey, Shawn W Polson, and K Eric Wommack. Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome*, 2(1):3, January 2014.

[224] Rachel L Marine, Daniel J Nasko, Jeffrey Wray, Shawn W Polson, and K Eric Wommack. Novel chaperonins are prevalent in the virioplankton and demonstrate links to viral biology and ecology. *ISME J.*, 11(11):2479–2491, November 2017.

[225] Rebeca Martín, Sylvie Miquel, Philippe Langella, and Luis G Bermúdez-Humarán. The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, 5(3):413–423, April 2014.

[226] Iñigo Martincorena, Aswin S N Seshasayee, and Nicholas M Luscombe. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature*, 485(7396):95–98, May 2012.

[227] Kristina Martinez-Guryn, Vanessa Leone, and Eugene B Chang. Regional diversity of the gastrointestinal microbiome. *Cell Host Microbe*, 26(3):314–324, September 2019.

[228] Cameron Martino, James T Morton, Clarisse A Marotz, Luke R Thompson, Anupriya Tripathi, Rob Knight, and Karsten Zengler. A novel sparse compositional technique reveals microbial perturbations. *mSystems*, 4(1), February 2019.

[229] Lauren D McDaniel, Elizabeth Young, Jennifer Delaney, Fabian Ruhnau, Kim B Ritchie, and John H Paul. High frequency of horizontal gene transfer in the oceans. *Science*, 330(6000):50, October 2010.

[230] Allison McDonald and Greg Vanlerberghe. Branched mitochondrial electron transport in the animalia: presence of alternative oxidase in several animal phyla. *IUBMB Life*, 56(6):333–341, June 2004.

[231] Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J Caporaso. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7, December 2012.

[232] J M McEvoy, A M Doherty, M Finnerty, J J Sheridan, L McGuire, I S Blair, D A McDowell, and D Harrington. The relationship between hide cleanliness and bacterial numbers on beef carcasses at a commercial abattoir. *Lett. Appl. Microbiol.*, 30(5):390–395, May 2001.

[233] Alexa B R McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert J Prill, Elizabeth Hénaff, Noah Alexander, Samuel S Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita R Colwell, Gail L Rosen, and Christopher E Mason. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.*, 18(1):182, September 2017.

[234] Michael R McLaren, Amy D Willis, and Benjamin J Callahan. Consistent and correctable bias in metagenomic sequencing experiments. *Elife*, 8, September 2019.

[235] Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.*, 10(4):e1003531, April 2014.

[236] A S Md Mukarram Hossain, Benjamin P Blackburne, Abhijeet Shah, and Simon Whelan. Evidence of statistical inconsistency of phylogenetic methods in the presence of multiple sequence alignment uncertainty. *Genome Biol. Evol.*, 7(8):2102–2116, July 2015.

[237] Yunjun Mei, Congcong He, Yongchi Huang, Ying Liu, Ziqian Zhang, Xiangdong Chen, and Ping Shen. Salinity Regulation of the Interaction of Halovirus SNJ1 with Its Host and Alteration of the Halovirus Replication Strategy to Adapt to the Variable Ecosystem. *PLOS ONE*, 10(4):e0123874, April 2015.

[238] Tomoko Mihara, Yosuke Nishimura, Yugo Shimizu, Hiroki Nishiyama, Genki Yoshikawa, Hideya Uehara, Pascal Hingamp, Susumu Goto, and Hiroyuki Ogata. Linking Virus Genomes with Host Taxonomy. *Viruses*, 8(3):66–66, March 2016.

[239] Eric S Miller, Elizabeth Kutter, Gisela Mosig, Fumio Arisaka, Takashi Kunisawa, and Wolfgang Rüger. Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.*, 67(1):86–156, table of contents, March 2003.

[240] Carolina M Mizuno, Charlotte Guyomar, Simon Roux, Régis Lavigne, Francisco Rodriguez-Valera, Matthew B Sullivan, Reynald Gillet, Patrick Forterre, and Mart Krupovic. Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat. Commun.*, 10(1):752, February 2019.

[241] Carolina Megumi Mizuno, Francisco Rodriguez-Valera, Nikole E. Kimes, and Rohit Ghai. Expanding the marine virosphere using metagenomics. *PLOS Genetics*, 9(12):1–13, December 2013.

[242] Shuming Mo, Bing Yan, Tingwei Gao, Jinhui Li, Muhammad Kashif, Jingjing Song, Lirong Bai, Dahui Yu, Jianping Liao, and Chengjian Jiang. Sulfur metabolism in subtropical marine mangrove sediments fundamentally differs from other habitats as revealed by SMDB. *Sci. Rep.*, 13(1):8126, May 2023.

[243] Adam Monier, Sebastian Sudek, Naomi M Fast, and Alexandra Z Worden. Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.*, 7(9):1764–1774, September 2013.

[244] Ryan M Moore, Amelia O Harrison, Sean M McAllister, Shawn W Polson, and K Eric Wommack. Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ*, 8:e8584, February 2020.

[245] Ryan M Moore, Amelia O Harrison, Daniel J Nasko, Jessica Chopyk, Metehan Cebeci, Barbra D Ferrell, Shawn W Polson, and K Eric Wommack. PASV: Automatic protein partitioning and validation using conserved residues. February 2021.

[246] Henning D Mootz, Elyse S Blum, Amy B Tyszkiewicz, and Tom W Muir. Conditional protein splicing: a new tool to control protein structure and function in vitro and in vivo. *J. Am. Chem. Soc.*, 125(35):10561–10569, September 2003.

[247] Jacob H Munson-McGee, Shengyun Peng, Samantha Dewerff, Ramunas Stepanauskas, Rachel J Whitaker, Joshua S Weitz, and Mark J Young. A virus

or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *The ISME Journal*, 12(7):1706–1714, 2018.

[248] Albert Leopold Müller, Kasper Urup Kjeldsen, Thomas Rattei, Michael Pester, and Alexander Loy. Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *The ISME journal*, 9(5):1152–1165, 2015.

[249] Vamsi Nallapareddy, Nicola Bordin, Ian Sillitoe, Michael Heinzinger, Maria Littmann, Vaishali P Waman, Neeladri Sen, Burkhard Rost, and Christine Orengo. CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models. *Bioinformatics*, 39(1), January 2023.

[250] Adit Naor, Neta Altman-Price, Shannon M Soucy, Anna G Green, Yulia Mitiagin, Israela Turgeman-Grott, Noam Davidovich, Johann Peter Gogarten, and Uri Gophna. Impact of a homing intein on recombination frequency and organismal fitness. *Proc. Natl. Acad. Sci. U. S. A.*, 113(32):E4654–61, August 2016.

[251] Daniel J Nasko, Jessica Chopyk, Eric G Sakowski, Barbra D Ferrell, Shawn W Polson, and K Eric Wommack. Family a DNA polymerase phylogeny uncovers diversity and replication gene organization in the virioplankton. *Front. Microbiol.*, 9:3053, December 2018.

[252] Stephen Nayfach and Katherine S Pollard. Toward accurate and quantitative comparative metagenomics. *Cell*, 166(5):1103–1116, August 2016.

[253] Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, Krishna Palaniappan, Joshua Ladau, Supratim Mukherjee, T B K Reddy, Torben Nielsen, Edward Kirton, José P Faria, Janaka N Edirisinghe, Christopher S Henry, Sean P Jungbluth, Dylan Chivian, Paramvir Dehal, Elisha M Wood-Charlson, Adam P Arkin, Susannah G Tringe, Axel Visel, IMG/M Data Consortium, Tanja Woyke, Nigel J Mouncey, Natalia N Ivanova, Nikos C Kyrpides, and Emiley A Eloe-Fadrosh. A genomic catalog of earth's microbiomes. *Nat. Biotechnol.*, 39(4):499–509, April 2021.

[254] Jacob T Nearing, André M Comeau, and Morgan G I Langille. Identifying biases and their potential solutions in human microbiome studies. *Microbiome*, 9(1):113, May 2021.

[255] Daniel Nelson. Phage taxonomy: we agree to disagree. *Journal of bacteriology*, 186(21):7029–7031, November 2004.

[256] Yosuke Nishimura, Hiroyasu Watai, Takashi Honda, Tomoko Mihara, Kimiho Omae, Simon Roux, Romain Blanc-Mathieu, Keigo Yamamoto, Pascal Hingamp, Yoshihiko Sako, Matthew B Sullivan, Susumu Goto, Hiroyuki Ogata, Takashi Yoshida, Environmental Viral, Genomes Shed, Yosuke Nishimura, Hiroyasu

Watai, Takashi Honda, Tomoko Mihara, Kimiho Omae, Simon Roux, Romain Blanc-Mathieu, Keigo Yamamoto, Pascal Hingamp, Yoshihiko Sako, Matthew B Sullivan, Susumu Goto, Hiroyuki Ogata, and Takashi Yoshida. Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere*, 2(2), 2017.

[257] Yosuke Nishimura, Takashi Yoshida, Megumi Kuronishi, Hideya Uehara, Hiroyuki Ogata, and Susumu Goto. ViPTree: the viral proteomic tree server. *Bioinformatics*, 33(15):2379–2380, March 2017.

[258] Tania Nobre, M Doroteia Campos, Eva Lucic-Mercy, and Birgit Arnholdt-Schmitt. Misannotation awareness: A tale of two Gene-Groups. *Front. Plant Sci.*, 7:868, June 2016.

[259] Pär Nordlund and Peter Reichard. Ribonucleotide reductases. *Annu. Rev. Biochem.*, 75:681–706, 2006.

[260] Pär Nordlund and Peter Reichard. Ribonucleotide Reductases. *Annual Review of Biochemistry*, 75(1):681–706, June 2006.

[261] Olga Novikova, Pradeepa Jayachandran, Danielle S Kelley, Zachary Morton, Samantha Merwin, Natalya I Topilina, and Marlene Belfort. Intein clustering suggests functional importance in different domains of life. *Mol. Biol. Evol.*, 33(3):783–799, March 2016.

[262] Olga Novikova, Natalya Topilina, and Marlene Belfort. Enigmatic distribution, evolution, and function of inteins. *J. Biol. Chem.*, 289(21):14490–14497, May 2014.

[263] T Heath Ogden and Michael S Rosenberg. Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.*, 55(2):314–328, April 2006.

[264] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt.

Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, January 2016.

[265] Zachary Paddock, Xiaorong Shi, Jianfa Bai, and T G Nagaraja. Applicability of a multiplex PCR to detect o26, o45, o103, o111, o121, o145, and O157 serogroups of escherichia coli in cattle feces. *Vet. Microbiol.*, 156(3-4):381–388, May 2012.

[266] Javier Palarea-Albaladejo and Josep Antoni Martín-Fernández. zcompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics Intellig. Lab. Syst.*, 143:85–96, April 2015.

[267] Amy Y Pan. Statistical analysis of microbiome data: The challenge of sparsity. *Current Opinion in Endocrine and Metabolic Research*, 19:35–40, August 2021.

[268] Sunita Panda, Ananya Nanda, Nilanjan Sahu, Deepak K Ojha, Biswaranjan Pradhan, Anjali Rai, Amol R Suryawanshi, Nilesh Banavali, and Sasmita Nayak. SufB intein splicing in mycobacterium tuberculosis is influenced by two remote conserved n-extein histidines. *Biosci. Rep.*, 42(3), March 2022.

[269] 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. *Environ. Microbiol.*, 18(5):1403–1414, May 2016.

[270] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

[271] Tansol Park, Lu Ma, Shengtao Gao, Dengpan Bu, and Zhongtang Yu. Heat stress impacts the multi-domain ruminal microbiota and some of the functional features independent of its effect on feed intake in lactating dairy cows. *J. Anim. Sci. Biotechnol.*, 13(1):71, June 2022.

[272] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2(11):1533–1542, November 2017.

[273] Frédéric Partensky and Laurence Garczarek. Prochlorococcus: advantages and limits of minimalism. *Ann. Rev. Mar. Sci.*, 2:305–331, 2010.

[274] Francisco Pascoal, Rodrigo Costa, and Catarina Magalhães. The microbial rare biosphere: current concepts, methods and ecological principles. *FEMS Microbiol. Ecol.*, 97(1), January 2021.

[275] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20, January 2019.

[276] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*, 10(12):1200–1202, December 2013.

[277] Theetha L Pavankumar. Inteins: Localized distribution, gene regulation, and protein engineering for biological applications. *Microorganisms*, 6(1), February 2018.

[278] Rich Pawlowicz. The electrical conductivity of seawater at high temperatures and salinities. *Desalination*, 300:32–39, August 2012.

[279] Rich Pawlowicz. Electrical Properties of Sea Water: Theory and Applications. In J. Kirk Cochran, Henry J. Bokuniewicz, and Patricia L. Yager, editors, *Encyclopedia of Ocean Sciences (Third Edition)*, pages 71–80. Academic Press, Oxford, January 2019.

[280] William R Pearson. An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics*, Chapter 3:3.1.1–3.1.8, June 2013.

[281] F B Perler, G J Olsen, and E Adam. Compilation and analysis of intein sequences. *Nucleic Acids Res.*, 25(6):1087–1093, March 1997.

[282] Francine B Perler. InBase: the intein database. *Nucleic Acids Res.*, 30(1):383–384, January 2002.

[283] Annika L Persson, Mathias Eriksson, Bettina Katterle, Stephan Pötsch, Margareta Sahlin, and Britt-Marie Sjöberg. A new mechanism-based radical intermediate in a mutant R1 protein affecting the catalytically essential glu441 inescherichia coli ribonucleotide reductase. *J. Biol. Chem.*, 272(50):31533–31541, December 1997.

[284] Stéphane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Sabrina Speich, Romain Troublé, Céline Dimier, Sarah Searson, Tara Oceans Consortium Coordinators, Silvia G Acinas, Peer Bork, Emmanuel Boss, Chris Bowler, Colomban De Vargas, Michael Follows, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels-Lewis, Lee Karp-Boss, Eric Karsenti, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Jeroen

Raes, Emmanuel G Reynaud, Christian Sardet, Mike Sieracki, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Shinichi Sunagawa, Didier Velayoudon, Jean Weissenbach, and Patrick Wincker. Open science resources for the discovery and analysis of *Tara* Oceans data. *Scientific Data*, 2, May 2015.

[285] Michael Pester, Norbert Bittner, Pinsurang Deevong, Michael Wagner, and Alexander Loy. A 'rare biosphere' microorganism contributes to sulfate reduction in a peatland. *ISME J.*, 4(12):1591–1602, December 2010.

[286] E. C. Pielou. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13(C):131–144, 1966.

[287] S Pietrokovski. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci.*, 3(12):2340–2350, December 1994.

[288] S Pietrokovski. Intein spread and extinction in evolution. *Trends Genet.*, 17(8):465–472, August 2001.

[289] Shmuel Pietrokovski. Modular organization of inteins and c-terminal autocatalytic domains: Modular organization of inteins and CADs. *Protein Sci.*, 7(1):64–71, January 1998.

[290] Alessandra Pontiroli, Aurora Rizzi, Pascal Simonet, Daniele Daffonchio, Timothy M Vogel, and Jean-Michel Monier. Visual evidence of horizontal gene transfer between plants and bacteria in the phytosphere of transplastomic tobacco. *Appl. Environ. Microbiol.*, 75(10):3314–3322, May 2009.

[291] Welkin H. Pope, Deborah Jacobs-Sera, Daniel A. Russell, Craig L. Peebles, Zein Al-Atrache, Turi A. Alcoser, Lisa M. Alexander, Matthew B. Alfano, Samantha T. Alford, Nichols E. Amy, Marie D. Anderson, Alexander G. Anderson, Andrew A. S. Ang, Manuel Ares, Jr., Amanda J. Barber, Lucia P. Barker, Jonathan M. Barrett, William D. Barshop, Cynthia M. Bauerle, Ian M. Bayles, Katherine L. Belfield, Aaron A. Best, Agustin Borjon, Jr., Charles A. Bowman, Christine A. Boyer, Kevin W. Bradley, Victoria A. Bradley, Lauren N. Broadway, Keshav Budwal, Kayla N. Busby, Ian W. Campbell, Anne M. Campbell, Alyssa Carey, Steven M. Caruso, Rebekah D. Chew, Chelsea L. Cockburn, Lianne B. Cohen, Jeffrey M. Corajod, Steven G. Cresawn, Kimberly R. Davis, Lisa Deng, Dee R. Denver, Breyon R. Dixon, Sahrish Ekram, Sarah C. R. Elgin, Angela E. Engelsen, Belle E. V. English, Marcella L. Erb, Crystal Estrada, Laura Z. Filliger, Ann M. Findley, Lauren Forbes, Mark H. Forsyth, Tyler M. Fox, Melissa J. Fritz, Roberto Garcia, Zindzi D. George, Anne E. Georges, Christopher R. Gissendanner, Shannon Goff, Rebecca Goldstein, Kobie C. Gordon, Russell D. Green, Stephanie L. Guerra, Krysta R. Guiney-Olsen, Bridget G. Guiza, Leila Haghighat, Garrett V. Hagopian, Catherine J. Harmon, Jeremy S. Harmson, Grant A. Hartzog, Samuel E. Harvey, Siping He, Kevin J. He, Kaitlin E.

Healy, Ellen R. Higinbotham, Erin N. Hildebrandt, Jason H. Ho, Gina M. Hogan, Victoria G. Hohenstein, Nathan A. Holz, Vincent J. Huang, Ericka L. Hufford, Peter M. Hynes, Arrykka S. Jackson, Erica C. Jansen, Jonathan Jarvik, Paul G. Jasinto, Tuajuanda C. Jordan, Tomas Kasza, Murray A. Katelyn, Jessica S. Kelsey, Larisa A. Kerrigan, Daryl Khaw, Junghee Kim, Justin Z. Knutter, Ching-Chung Ko, Gail V. Larkin, Jennifer R. Laroche, Asma Latif, Kohana D. Leuba, Sequoia I. Leuba, Lynn O. Lewis, Kathryn E. Loesser-Casey, Courtney A. Long, A. Javier Lopez, Nicholas Lowery, Tina Q. Lu, Victor Mac, Isaac R. Masters, Jazmyn J. McCloud, Molly J. McDonough, Andrew J. Medenbach, Anjali Menon, Rachel Miller, Brandon K. Morgan, Patrick C. Ng, Elvis Nguyen, Katrina T. Nguyen, Emilie T. Nguyen, Kaylee M. Nicholson, Lindsay A. Parnell, Caitlin E. Peirce, Allison M. Perz, Luke J. Peterson, Rachel E. Pferdehirt, Seegren V. Philip, Kit Pogliano, Joe Pogliano, Tamsen Polley, Erica J. Puopolo, Hannah S. Rabinowitz, Michael J. Resiss, Corwin N. Rhyan, Yetta M. Robinson, Lauren L. Rodriguez, Andrew C. Rose, Jeffrey D. Rubin, Jessica A. Ruby, Margaret S. Saha, James W. Sandoz, Judith Savitskaya, Dale J. Schipper, Christine E. Schnitzler, Amanda R. Schott, J. Bradley Segal, Christopher D. Shaffer, Kathryn E. Sheldon, Erica M. Shepard, Jonathan W. Shepardson, Madav K. Shroff, Jessica M. Simmons, Erika F. Simms, Brandy M. Simpson, Kathryn M. Sinclair, Robert L. Sjoholm, Ingrid J. Slette, Blaire C. Spaulding, Clark L. Straub, Joseph Stukey, Trevor Sughrue, Tin-Yun Tang, Lyons M. Tatyana, Stephen B. Taylor, Barbara J. Taylor, Louise M. Temple, Jasper V. Thompson, Michael P. Tokarz, Stephanie E. Trapani, Alexander P. Troum, Jonathan Tsay, Anthony T. Tubbs, Jillian M. Walton, Danielle H. Wang, Hannah Wang, John R. Warner, Emilie G. Weisser, Samantha C. Wendler, Kathleen A. Weston-Hafer, Hilary M. Whelan, Kurt E. Williamson, Angelica N. Willis, Hannah S. Wirtshafter, Theresa W. Wong, Phillip Wu, Yun jeong Yang, Brandon C. Yee, David A. Zaidins, Bo Zhang, Melina Y. Zúniga, Roger W. Hendrix, and Graham F. Hatfull. Expanding the Diversity of Mycobacteriophages: Insights into Genome Architecture and Evolution. *PLOS ONE*, 6(1):e16329, January 2011.

[292] Björn Possé, Lieven De Zutter, Marc Heyndrickx, and Lieve Herman. Novel differential and confirmation plating media for shiga toxin-producing escherichia coli serotypes o26, o103, o111, O145 and sorbitol-positive and -negative O157. *FEMS Microbiol. Lett.*, 282(1):124–131, May 2008.

[293] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3), 2010.

[294] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, March 2010.

[295] Elmar Pruesse, Frank Oliver Glöckner, and Jörg Peplies. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829, May 2012.

[296] 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, November 2012.

[297] Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A field guide for the compositional analysis of any-omics data. *Gigascience*, 8(9), September 2019.

[298] Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, August 2018.

[299] Thomas P Quinn, Mark F Richardson, David Lovell, and Tamsyn M Crowley. propr: An r-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.*, 7(1):16252, November 2017.

[300] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[301] Eric J. Raes, Levente Bodrossy, Jodie van de Kamp, Andrew Bissett, and Anya M. Waite. Marine bacterial richness increases towards higher latitudes in the eastern Indian Ocean. *Limnology and Oceanography Letters*, 3(1):10–19, February 2018.

[302] Andrew Rambaut. FigTree, 2006.

[303] Yoav Raz and Emmanuel Tannenbaum. The influence of horizontal gene transfer on the mean fitness of unicellular populations in static environments. *Genetics*, 185(1):327–337, May 2010.

[304] Richard Reeve, Tom Leinster, Christina A Cobbold, Jill Thompson, Neil Brummitt, Sonia N Mitchell, and Louise Matthews. How to partition diversity. April 2014.

[305] P Reichard. From RNA to DNA, why so many ribonucleotide reductases? *Science*, 260(5115):1773–1777, June 1993.

[306] P Reichard. From RNA to DNA, why so many ribonucleotide reductases? *Science*, 260(5115):1773, June 1993.

[307] P Reichard. From RNA to DNA, why so many ribonucleotide reductases? *Science*, 260(5115):1773–1777, June 1993.

[308] Linta Reji, Emily L Cardarelli, Kristin Boye, John R Bargar, and Christopher A Francis. Diverse ecophysiological adaptations of subsurface thaumarchaeota in floodplain sediments revealed through genome-resolved metagenomics. *ISME J.*, 16(4):1140–1152, April 2022.

[309] Boyu Ren, Sergio Bacallado, Stefano Favaro, Susan Holmes, and Lorenzo Trippa. Bayesian nonparametric ordination for the analysis of microbial communities. *J. Am. Stat. Assoc.*, 112(520):1430–1442, February 2017.

[310] Liam J. Revell. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[311] Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K Swan, Esther A Gies, Jeremy A Dodsworth, Brian P Hedlund, George Tsiamis, Stefan M Sievert, Wen-Tso Liu, Jonathan A Eisen, Steven J Hallam, Nikos C Kyrpides, Ramunas Stepanauskas, Edward M Rubin, Philip Hugenholtz, and Tanja Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, July 2013.

[312] J Rivera-Pinto, J J Egozcue, V Pawlowsky-Glahn, R Paredes, M Noguera-Julian, and M L Calle. Balances: a new perspective for microbiome analysis. *mSystems*, 3(4), July 2018.

[313] J Rivera-Pinto, J J Egozcue, V Pawlowsky-Glahn, R Paredes, M Noguera-Julian, and M L Calle. Balances: a new perspective for microbiome analysis. *mSystems*, 3(4), July 2018.

[314] Oscar Robinson, David Dylus, and Christophe Dessimoz. Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Molecular Biology and Evolution*, 33(8):2163–2166, 2016.

[315] Forest Rohwer and Rob Edwards. The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16):4529–4535, August 2002.

[316] Forest Rohwer and Rebecca Vega Thurber. Viruses manipulate the marine environment. *Nature*, 459(7244):207–212, May 2009.

[317] B Rost. Twilight zone of protein sequence alignments. *Protein Eng.*, 12(2):85–94, February 1999.

[318] B Rost. Twilight zone of protein sequence alignments. *Protein Eng.*, 12(2):85–94, February 1999.

[319] Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.*, 12(2):85–94, February 1999.

[320] Simon Roux, Jennifer R Brum, Bas E Dutilh, Shinichi Sunagawa, Melissa B Duhaime, Alexander Loy, Bonnie T Poulos, Natalie Solonenko, Elena Lara, Julie Poulain, Stéphane Pesant, Stefanie Kandels-Lewis, Céline Dimier, Marc Picheral, Sarah Searson, Corinne Cruaud, Adriana Alberti, Carlos M Duarte, Josep M Gasol, Dolors Vaqué, Tara Oceans Coordinators, Peer Bork, Silvia G Acinas, Patrick Wincker, and Matthew B Sullivan. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693, September 2016.

[321] Simon Roux, Jennifer R. Brum, Bas E. Dutilh, Shinichi Sunagawa, Melissa B. Duhaime, Alexander Loy, Bonnie T. Poulos, Natalie Solonenko, Elena Lara, Julie Poulain, Stéphane Pesant, Stefanie Kandels-Lewis, Céline Dimier, Marc Picheral, Sarah Searson, Corinne Cruaud, Adriana Alberti, Carlos M. Duarte, Josep M. Gasol, Dolors Vaqué, Peer Bork, Silvia G. Acinas, Patrick Wincker, and Matthew B. Sullivan. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693, 2016.

[322] Simon Roux, Steven J. Hallam, Tanja Woyke, and Matthew B. Sullivan. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, 4:1–20, 2015.

[323] Inna Rozman Grinberg, Daniel Lundin, Margareta Sahlin, Mikael Crona, Gustav Berggren, Anders Hofer, and Britt-Marie Sjöberg. A glutaredoxin domain fused to the radical-generating subunit of ribonucleotide reductase (RNR) functions as an efficient RNR reductant. *J. Biol. Chem.*, 293(41):15889–15900, October 2018.

[324] Anna Sajduda, Anna Brzostek, Marta Poplawska, Ewa Augustynowicz-Kopec, Zofia Zwolska, Stefan Niemann, Jaroslaw Dziadek, and Doris Hillemann. Molecular characterization of rifampin- and isoniazid-resistant mycobacterium tuberculosis strains isolated in poland. *J. Clin. Microbiol.*, 42(6):2425–2431, June 2004.

[325] Eric G Sakowski, Erik V Munsell, Mara Hyatt, William Kress, Shannon J Williamson, Daniel J Nasko, Shawn W Polson, and K Eric Wommack. Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proc. Natl. Acad. Sci. U. S. A.*, 111(44):15786–15791, November 2014.

[326] Eric G. Sakowski, Erik V. Munsell, Mara Hyatt, William Kress, Shannon J. Williamson, Daniel J. Nasko, Shawn W. Polson, and K. Eric Wommack. Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44):15786–15791, November 2014.

[327] Eric G Sakowski, Erik V Munsell, Mara Hyatt, William Kress, Shannon J Williamson, Daniel J Nasko, Shawn W Polson, and K Eric Wommack. Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proc. Natl. Acad. Sci. U. S. A.*, 111(44):15786–15791, November 2014.

[328] Lana Saleh and Francine B Perler. Protein splicingin cis andin trans. *Chem. Rec.*, 6(4):183–193, 2006.

[329] Rodrigo Santamaría and Roberto Therón. Treevolution: Visual analysis of phylogenetic trees. *Bioinformatics*, 25(15):1970–1971, 2009.

[330] Tasha M Santiago-Rodriguez and Emily B Hollister. Unraveling the viral dark matter through viral metagenomics. *Front. Immunol.*, 13:1005107, September 2022.

[331] Jimmy H W Saw. Characterizing the uncultivated microbial minority: towards understanding the roles of the rare biosphere in microbial communities. *mSystems*, 6(4):e0077321, August 2021.

[332] Gary Sawers. Biochemistry, physiology and molecular biology of glycyl radical enzymes. *FEMS Microbiol. Rev.*, 22(5):543–551, December 1998.

[333] Helen F Schmidt, Eric G Sakowski, Shannon J Williamson, Shawn W Polson, and K Eric Wommack. Shotgun metagenomics indicates novel family a DNA polymerases predominate within marine virioplankton. *ISME J.*, 8(1):103–114, January 2014.

[334] Helen F Schmidt, Eric G Sakowski, Shannon J Williamson, Shawn W Polson, and K Eric Wommack. Shotgun metagenomics indicates novel family a DNA polymerases predominate within marine virioplankton. *ISME J.*, 8(1):103–114, January 2014.

[335] Alexandra M Schnoes, Shoshana D Brown, Igor Dodevski, and Patricia C Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, 5(12):e1000605, December 2009.

[336] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, July 2014.

[337] João C Setubal. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys. Rev.*, 13(6):905–909, December 2021.

[338] James Seward, Michael A Carson, L J Lamit, Nathan Basiliko, Joseph B Yavitt, Erik Lilleskov, Christopher W Schadt, Dave Solance Smith, Jim Mclaughlin, Nadia Mykytczuk, Shanay Willims-Johnson, Nigel Roulet, Tim Moore, Lorna Harris, and Suzanna Bräuer. Peatland microbial community composition is driven by a natural climate gradient. *Microb. Ecol.*, 80(3):593–602, October 2020.

[339] Ashley Shade, Stuart E Jones, J Gregory Caporaso, Jo Handelsman, Rob Knight, Noah Fierer, and Jack A Gilbert. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio*, 5(4):e01371–14, July 2014.

[340] Michael Shaffer, Mikayla A Borton, Bridget B McGivern, Ahmed A Zayed, Sabina Leanti La Rosa, Lindsey M Solden, Pengfei Liu, Adrienne B Narrowe, Josué Rodríguez-Ramos, Benjamin Bolduc, M Consuelo Gazitúa, Rebecca A Daly, Garrett J Smith, Dean R Vik, Phil B Pope, Matthew B Sullivan, Simon Roux, and Kelly C Wrighton. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.*, 48(16):8883–8900, September 2020.

[341] Neel H Shah and Tom W Muir. Inteins: Nature's gift to protein chemists. *Chem. Sci.*, 5(1):446–461, 2014.

[342] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, November 2003.

[343] Maria A Sierra, Qianhao Li, Smruti Pushalkar, Bidisha Paul, Tito A Sandoval, Angela R Kamer, Patricia Corby, Yuqi Guo, Ryan Richard Ruff, Alexander V Alekseyenko, Xin Li, and Deepak Saxena. The influences of bioinformatics tools and reference databases in analyzing the human oral microbial community. *Genes*, 11(8), August 2020.

[344] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, 7:539, October 2011.

[345] Rachel L. Simister, Peter Deines, Emmanuelle S. Botté, Nicole S. Webster, and Michael W. Taylor. Sponge-specific clusters revisited: A comprehensive phylogeny of sponge-associated microorganisms. *Environmental Microbiology*, 14(2):517–524, 2012.

[346] Peter Simmonds. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology*, 96(6):1193–1206, 2015.

[347] Peter Simmonds, Mike J. Adams, Mária Benkő, Mya Breitbart, J. Rodney Brister, Eric B. Carstens, Andrew J. Davison, Eric Delwart, Alexander E. Gorbalenya, Balázs Harrach, Roger Hull, Andrew M.Q. King, Eugene V. Koonin,

Mart Krupovic, Jens H. Kuhn, Elliot J. Lefkowitz, Max L. Nibert, Richard Orton, Marilyn J. Roossinck, Sead Sabanadzovic, Matthew B. Sullivan, Curtis A. Suttle, Robert B. Tesh, René A. van der Vlugt, Arvind Varsani, and F. Murilo Zerbini. Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, 15:161, January 2017.

[348] Chris S Smillie, Mark B Smith, Jonathan Friedman, Otto X Cordero, Lawrence A David, and Eric J Alm. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241–244, October 2011.

[349] Patricia A Sobecky and Jonna M Coombs. Horizontal gene transfer in metal and radionuclide contaminated soils. *Methods Mol. Biol.*, 532:455–472, 2009.

[350] Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103(32):12115–12120, 2006.

[351] Lindsey Solden, Karen Lloyd, and Kelly Wrighton. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr. Opin. Microbiol.*, 31:217–226, June 2016.

[352] Shannon M Soucy, Matthew S Fullmer, R Thane Papke, and Johann Peter Gogarten. Inteins as indicators of gene flow in the halobacteria. *Front. Microbiol.*, 5:299, June 2014.

[353] Shannon M Soucy and J Peter Gogarten. Inteins as indicators of Bio-Communication. In Guenther Witzany, editor, *Biocommunication of Archaea*, pages 265–275. Springer International Publishing, Cham, 2017.

[354] M W Southworth, J Benner, and F B Perler. An alternative protein splicing mechanism for inteins lacking an n-terminal nucleophile. *EMBO J.*, 19(18):5019–5026, September 2000.

[355] Vivek Srinivas, Hugo Lebrette, Daniel Lundin, Yuri Kutin, Margareta Sahlin, Michael Lerche, Jürgen Eirich, Rui M. M. Branca, Nicholas Cox, Britt-Marie Sjöberg, and Martin Högbom. Metal-free ribonucleotide reduction powered by a DOPA radical in Mycoplasma pathogens. *Nature*, 563(7731):416–420, November 2018.

[356] J L Stein, T L Marsh, K Y Wu, H Shizuya, and E F DeLong. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.*, 178(3):591–599, February 1996.

[357] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, 16(7):603–606, July 2019.

[358] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028, November 2017.

[359] Zachary R Stromberg, Nicholas W Baumann, Gentry L Lewis, Nicholas J Sevart, Natalia Cernicchiaro, David G Renter, David B Marx, Randall K Phebus, and Rodney A Moxley. Prevalence of enterohemorrhagic escherichia coli o26, o45, o103, o111, o121, o145, and O157 on hides and preintervention carcass surfaces of feedlot cattle at harvest. *Foodborne Pathog. Dis.*, 12(7):631–638, July 2015.

[360] Iris J E Stulemeijer and Matthieu H A J Joosten. Post-translational modification of host proteins in pathogen-triggered defence signalling in plants. *Mol. Plant Pathol.*, 9(4):545–560, July 2008.

[361] Ben C Stöver and Kai F Müller. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, 11:7, January 2010.

[362] Ben C Stöver, Sarah Wiechers, and Kai F Müller. JPhyloIO — A Java library for event-based reading and writing of different alignment and tree formats through one common interface Aims and concept Event based document reading Writing events using data adapters, 2016.

[363] Matthew B Sullivan. Viromes, not gene markers, for studying double-stranded DNA virus communities. *J. Virol.*, 89(5):2459–2461, March 2015.

[364] Matthew B Sullivan, Joshua S Weitz, and Steven Wilhelm. Viral ecology comes of age. *Environ. Microbiol. Rep.*, 9(1):33–35, February 2017.

[365] Curtis A. Suttle. Marine viruses – major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812, October 2007.

[366] Kristen S Swithers, Alireza G Senejani, Gregory P Fournier, and J Peter Gogarten. Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evol. Biol.*, 9:303, December 2009.

[367] Rubén Sánchez, François Serra, Joaquín Tárraga, Ignacio Medina, José Carbonell, Luis Pulido, Alejandro De María, Salvador Capella-Gutíerrez, Jaime Huerta-Cepas, Toni Gabaldón, Joaquín Dopazo, and Hernán Dopazo. Phylemon 2.0: A suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Research*, 39:470–474, 2011.

[368] S Tabor and C C Richardson. A single residue in DNA polymerases of the escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.*, 92(14):6339–6343, July 1995.

[369] Eric Talevich, Brandon M. Invergo, Peter J.A. Cock, and Brad A. Chapman. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13:209, 2012.

[370] Javier Tamames, Marta Cobo-Simón, and Fernando Puente-Sánchez. Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics*, 20(1):960, December 2019.

[371] Amalio Telenti, Levi C T Pierce, William H Biggs, Julia di Iulio, Emily H M Wong, Martin M Fabani, Ewen F Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A Perkins, Franz J Och, Yaron Turpaz, and J Craig Venter. Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U. S. A.*, 113(42):11901–11906, October 2016.

[372] Natalya I Topilina, Cathleen M Green, Pradeepa Jayachandran, Danielle S Kelley, Matthew J Stanger, Carol Lyn Piazza, Sasmita Nayak, and Marlene Belfort. SufB intein of mycobacterium tuberculosis as a sensor for oxidative and nitrosative stresses. *Proc. Natl. Acad. Sci. U. S. A.*, 112(33):10348–10353, August 2015.

[373] Natalya I Topilina, Olga Novikova, Matthew Stanger, Nilesh K Banavali, and Marlene Belfort. Post-translational environmental switch of RadA activity by extein-intein interactions in protein splicing. *Nucleic Acids Res.*, 43(13):6631–6648, July 2015.

[374] Eduard Torrents. Ribonucleotide reductases: essential enzymes for bacterial life. *Front. Cell. Infect. Microbiol.*, 4:52, April 2014.

[375] Béla Tóthmérész. Comparison of different methods for diversity ordering. *J. Veg. Sci.*, 6(2):283–290, 1995.

[376] Béla Tóthmérész. Comparison of different methods for diversity ordering. *J. Veg. Sci.*, 6(2):283–290, 1995.

[377] Patricia Q Tran, Samantha C Bachand, Peter B McIntyre, Benjamin M Kraemer, Yvonne Vadeboncoeur, Ismael A Kimirei, Rashid Tamatamah, Katherine D McMahon, and Karthik Anantharaman. Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater lake tanganyika. *ISME J.*, 15(7):1971–1986, February 2021.

[378] H James Tripp, Ian Hewson, Sam Boyarsky, Joshua M Stuart, and Jonathan P Zehr. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res.*, 39(20):8792–8802, November 2011.

[379] Jan Dirk Van Elsas, Sarah Turner, and Mark J Bailey. Horizontal gene transfer in the phytosphere. *New Phytol.*, 157(3):525–537, March 2003.

[380] Timothy G. Vaughan. IcyTree: Rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, 33(15):2392–2394, 2017.

[381] Catherine Vilchèze, Torin R Weisbrod, Bing Chen, Laurent Kremer, Manzour H Hazbón, Feng Wang, David Alland, James C Sacchettini, and William R Jacobs, Jr. Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria. *Antimicrob. Agents Chemother.*, 49(2):708–720, February 2005.

[382] Julia Villarroel, A Kortine Kleinheinz, I Vanessa Jurtz, Henrike Zschach, Ole Lund, Morten Nielsen, V Mette Larsen, Kortine Annina Kleinheinz, Vanessa Isabell Jurtz, Henrike Zschach, Ole Lund, Morten Nielsen, and Mette Voldby Larsen. HostPhinder: A Phage Host Prediction Tool. *Viruses*, 8(5):1–22, 2016.

[383] Rutger A. Vos, Jason Caravas, Klaas Hartmann, Mark A. Jensen, and Chase Miller. BIO::Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics*, 12:63, February 2011.

[384] Han Wang, Jonathan Liu, Kai P Yuet, Andrew J Hill, and Paul W Sternberg. Split cGAL, an intersectional strategy using a split intein for refined spatiotemporal transgene control in caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.*, 115(15):3900–3905, April 2018.

[385] Hao Wang, Lin Wang, Baihua Zhong, and Zhuojun Dai. Protein splicing of inteins: A powerful tool in synthetic biology. *Front Bioeng Biotechnol*, 10:810180, February 2022.

[386] Li-San Wang, Jim Leebens-Mack, P Kerr Wall, Kevin Beckmann, Claude W dePamphilis, and Tandy Warnow. The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8(4):1108–1119, July 2011.

[387] Wei-Lin Wang, Shao-Yan Xu, Zhi-Gang Ren, Liang Tao, Jian-Wen Jiang, and Shu-Sen Zheng. Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.*, 21(3):803–814, January 2015.

[388] Xinzhen Wang, Junjie Liu, Zhenhua Yu, Jian Jin, Xiaobing Liu, and Guanghua Wang. Novel groups of cyanobacterial podovirus DNA polymerase (pol) genes exist in paddy waters in northeast china. *FEMS Microbiol. Ecol.*, 92(12), December 2016.

[389] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[390] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[391] Amy Willis, John Bunge, and Thea Whitman. Improved detection of changes in species richness in high diversity microbial communities. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 66(5):963–977, 2017.

[392] Amy D Willis. Rarefaction, alpha diversity, and statistics. *Front. Microbiol.*, 10:2407, October 2019.

[393] Amy D Willis and Bryan D Martin. Estimating diversity in networked ecological communities. *Biostatistics*, 23(1):207–222, January 2022.

[394] Christian Winter, Blake Matthews, and Curtis A Suttle. Effects of environmental variation and spatial distance on Bacteria, Archaea and viruses in sub-polar and arctic waters. *The ISME Journal*, 7:1507, April 2013.

[395] K Eri Wommack, Daniel J Nasko, Jessica Chopyk, and Eric G Sakowski. Counts and sequences, observations that continue to change our understanding of viruses in nature. *J. Microbiol.*, 53(3):181–192, March 2015.

[396] K Eric Wommack, Jaysheel Bhavsar, Shawn W Polson, Jing Chen, Michael Dumas, Sharath Srinivasiah, Megan Furman, Sanchita Jamindar, and Daniel J Nasko. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, 6(3):421–433, 2012.

[397] K. Eric Wommack, Daniel J. Nasko, Jessica Chopyk, and Eric G. Sakowski. Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology*, 53(3):181–192, 2015.

[398] Karen M Wong, Marc A Suchard, and John P Huelsenbeck. Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476, January 2008.

[399] John C Wooley and Yuzhen Ye. Metagenomics: Facts and artifacts, and computational challenges*. *J. Comput. Sci. Technol.*, 25(1):71–81, January 2009.

[400] H Wu, Z Hu, and X Q Liu. Protein trans-splicing by a split intein encoded in a split DnaE gene of synechocystis sp. PCC6803. *Proc. Natl. Acad. Sci. U. S. A.*, 95(16):9226–9231, August 1998.

[401] Ling-Yi Wu, Gonçalo J Piedade, Ryan M Moore, Amelia O Harrison, Ana M Martins, Kay D Bidle, Shawn W Polson, Eric G Sakowski, Jozef I Nissimov, Jacob T Dums, Barbra D Ferrell, and K Eric Wommack. Ubiquitous, b12-dependent virioplankton utilizing ribonucleotide-triphosphate reductase demonstrate interseasonal dynamics and associate with a diverse range of bacterial hosts in the pelagic ocean. *ISME Commun*, 3(1):108, October 2023.

[402] Zhiqiang Wu, Li Yang, Xianwen Ren, Guimei He, Junpeng Zhang, Jian Yang, Zhaohui Qian, Jie Dong, Lilian Sun, Yafang Zhu, Jiang Du, Fan Yang, Shuyi Zhang, and Qi Jin. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *The ISME Journal*, 10(3):609–620, 2016.

[403] Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, December 2013.

[404] Yong Xu, Eric Dugat-Bony, Rahat Zaheer, Lorna Selinger, Ruth Barbieri, Krysty Munns, Tim A. McAllister, and L. Brent Selinger. Escherichia coli O157:H7 Super-Shedder and Non-Shedder Feedlot Steers Harbour Distinct Fecal Bacterial Communities. *PLOS ONE*, 9(5):e98115, May 2014.

[405] Ming-Yuan Xue, Yun-Yi Xie, Yifan Zhong, Xiao-Jiao Ma, Hui-Zeng Sun, and Jian-Xin Liu. Integrated meta-omics reveals new ruminal microbial features associated with feed efficiency in dairy cattle. *Microbiome*, 10(1):32, February 2022.

[406] Ming Yan, Akbar Adjie Pratama, Sripoorna Somasundaram, Zongjun Li, Yu Jiang, Matthew B Sullivan, and Zhongtang Yu. Interrogating the viral dark matter of the rumen ecosystem with a global virome database. *Nat. Commun.*, 14(1):5254, August 2023.

[407] Kuan Yang and Liqing Zhang. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Res.*, 36(5):e33, March 2008.

[408] Mingzhang Yang, Myra K Derbyshire, Roxanne A Yamashita, and Aron Marchler-Bauer. NCBI's conserved domain database and tools for protein domain analysis. *Curr. Protoc. Bioinformatics*, 69(1):e90, March 2020.

[409] P Young, M Ohman, M Q Xu, D A Shub, and B M Sjöberg. Intron-containing T4 bacteriophage gene suny encodes an anaerobic ribonucleotide reductase. *J. Biol. Chem.*, 269(32):20229–20232, August 1994.

[410] Zhuofeng Yu, Pinjing He, Liming Shao, Hua Zhang, and Fan Lü. Co-occurrence of mobile genetic elements and antibiotic resistance genes in municipal solid waste landfill leachates: A preliminary insight into the role of landfill age. *Water Res.*, 106:583–592, December 2016.

[411] Qianmu Yuan, Junjie Xie, Jiancong Xie, Huiying Zhao, and Yuedong Yang. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Brief. Bioinform.*, 24(3), May 2023.

[412] Rahat Zaheer, Eric Dugat-Bony, Devon Holman, Elodie Cousteix, Yong Xu, Krysty Munns, Lorna J Selinger, Rutn Barbieri, Trevor Alexander, Tim A McAllister, and L Brent Selinger. Changes in bacterial community composition of Escherichia coli O157:H7 super-shedder cattle occur in the lower intestine. *PLOS ONE*, 12(1):e0170050–e0170050, January 2017.

[413] Qinglu Zeng, Richard P Bonocora, and David A Shub. A free-standing homing endonuclease targets an intron insertion site in the psba gene of cyanophages. *Curr. Biol.*, 19(3):218–222, February 2009.

[414] Marie Lisandra Zepeda Mendoza, Thomas Sicheritz-Pontén, and M Thomas P Gilbert. Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses. *Brief. Bioinform.*, 16(5):745–758, September 2015.

[415] Jiachao Zhang, Chuanbiao Xu, Dongxue Huo, Qisong Hu, and Qiannan Peng. Comparative study of the gut microbiome potentially related to milk protein in murrah buffaloes (bubalus bubalis) and chinese holstein cattle. *Sci. Rep.*, 7:42189, February 2017.

[416] Ru-Yi Zhang, Bin Zou, Yong-Wei Yan, Che Ok Jeon, Meng Li, Mingwei Cai, and Zhe-Xue Quan. Design of targeted primers based on 16S rRNA sequences in meta-transcriptomic datasets and identification of a novel taxonomic group in the asgard archaea. *BMC Microbiol.*, 20(1):25, February 2020.

[417] Yong-Zhen Zhang, Mang Shi, and Edward C Holmes. Using metagenomics to characterize an expanding virosphere. *Cell*, 172(6):1168–1172, March 2018.

[418] Zhiyi Zhang and Jun Zhou. Re-parameterization of multinomial distributions and diversity indices. *J. Stat. Plan. Inference*, 140(7):1731–1738, July 2010.

[419] L. Zhao, P.J. Tyler, J. Starnes, C.L. Bratcher, D. Rankins, T.A. McCaskey, and L. Wang. Correlation analysis of Shiga toxin–producing Escherichia coli shedding and faecal bacterial composition in beef cattle. *Journal of Applied Microbiology*, 115(2):591–603, August 2013.

[420] A M Zuurmond, L N Olsthoorn-Tieleman, J Martien de Graaf, A Parmeggiani, and B Kraal. Mutant EF-Tu species reveal novel features of the enacyloxin IIa inhibition mechanism on the ribosome. *J. Mol. Biol.*, 294(3):627–637, December 1999.

Appendix A

CATTLE MICROBIOME EXPERIMENTAL METHODS

A.1 Sample collection, STEC detection, and microbiome sequencing

A.1.1 Sample collection

Cattle population characteristics, sample collection, and STEC detection in fecal and hide samples is described in detail elsewhere [77, 62]. A brief overview will be provided here; for full details, see the original cited works. Over a twelve week period in from June to August 2013, fecal and hide samples were collected. Fecal samples were taken at the feedlot and hide-on carcass surface sponge samples were taken at the abattoir as described in [77]. Twenty-four pen-floor fecal samples were collected each week from two pens 12 to 24 hours prior to transport to the harvesting location, where 24 hide-on carcass samples (12 samples from two separate pens (pens changed each week)) were collected using 11.5 * 23.0-cm sponges (Speci-Sponge®; Nasco, Fort Atkinson, WI) pre-moistened with 35 mL of 0.1% sterile buffered peptone water (BPW) [77]. After cattle were stunned and bled, but prior to hide removal, 1000 square centimeters of the hide was swabbed 15 cm from the midline at the level of the diaphragm. Five gram and 5 mL aliquots from each fecal and hide sample, respectively, were reserved for microbiome analysis. Aliquots were snap-frozen in liquid nitrogen (LN2) within 1 hour of collection for fecal samples and 2 hours for hide samples, then stored at -80 degrees C until completion of a molecular detection assay for EHEC. Detailed descriptions of the sampling procedure and characteristics of the study population are presented in [77].

A.1.2 STEC detection in hide samples

The prevalence of EHEC in hide samples (the 35 mL sample-BPW suspension plus 90 mL *E. coli* broth (EC; Oxoid Lt., Hampshire, UK)) was tested in a previous study [359] using the NeoSEEK™ STEC Detection and Identification test (NS; Neogen Corp., Lansing, MI). The NeoSEEK™ test determines the presence or absence of EHEC O26, O45, O103, O111, O121, O145, and O157 using PCR and mass spectrometry to test for more than 70 markers including O-group, Shiga toxin, and intimin. NeoSEEK™ testing was conducted at GeneSEEK® Inc. (Lincoln, NE).

A.1.3 STEC detection in fecal samples

Pen floor samples were collected from the feedlot, snap frozen, and processed for extraction of microbial nucleic acid extraction and detection of STEC prevalence. STEC prevalence and serogroup identification was conducted by a collaborating lab by PCR for various genes associated with STEC: *hly*, *eae*, *stx1*, *stx2*, and O-group [77]. Specifically, two grams from each fecal sample was enriched in *E. coli* broth followed by 980 µL aliquot added to serogroup-specific IMS beads (Abraxis®, Warminister, PA) for specific STEC serogroups (O26, O45, O103, O111, O121, O145, and O157). These IMS suspensions were cultured on MacConkey agar with cefixime and potassium tellurite for O157 and/or modified Poss [292] for the non-O157 serogroups. After 24h incubation, six colonies from each plate were selected and streaked on blood agar plates for an additional incubation of 24 hours at 37 degrees C. On the O157 plates, any colonies that were positive for latex agglutination and indole production were used in a multiplex PCR to identify key O157 genes: *fliC* (encodes the *E. coli* flagellum), *rfbE* (encodes the *E. coli* O157 serotype), *ehxA* (enterohemolysin), *stx1*, *stx2* [22]. Non-O157 serogroups were tested for a variety of serogroup specific genes (O26, O45, O103, O111, O121, and O145) [265]. If positive, these were further subjected to tests for a suite of virulence genes including *eae*, *stx1*, and *stx2* [21].

A.1.4 Microbiome sequencing

Microbial nucleic acid was extracted with the MO BIO PowerViral® DNA/RNA isolation kit (MO BIO Laboratories, Inc.). DNA concentrates were sent to the sequencing center at the University of Delaware for standard library preparation for Illumina HiSeq 2500 SBS 2x251 sequencing.

Cell-free DNA extracts (viromes) were constructed from the microbial extractions using an adapted $FeCl_3$ method [153]. Fecal samples were diluted to 50 mL with phosphate buffered saline and gently shaken for one hour, then filtered with a 0.22 µm polycarbonate filter and spiked with 50 µL of $FeCl_3$, and left for incubation at room temperature for one hour. $FeCl_3$ flocculate was then filtered onto a 1.0 µm polycarbonate filter. Phages in Fe precipitates were resuspended in 500 µL of oxalic acid buffer. Remaining free cellular DNA was digested with a two hour DNAse incubation, then 0.22 µm filtered again. Finally, a SSU rRNA PCR was performed to ensure phage concentrates were free from cellular DNA contamination [59]. Viromes underwent library preparation for and sequencing on the Illumina HiSeq 2500 1x151.

A.2 Bioinformatics methods

A.2.1 Read quality control

A quality control pipeline constructed from standard read quality control programs was used. First, adapters were trimmed from the forward and reverse reads using Trimmomatic version 0.35 [37] (default settings except – seed mismatches: 2, palindrome clip threshold: 30, simple clip threshold: 10). Then, single-end quality trimming was performed on any read pairs that were broken up by the adapter trimming. Next, the read pairs that remained after adapter trimming were merged using FLASH version 1.2.11 (default settings except – max overlap: 250). Then, reads were subjected to quality trimming with Trimmomatic version 0.35 [37] (default settings except – headcrop: 0, sliding window size: 10 with minimum quality score of 15, minimum length: 50). After adapter trimming, paired-end merging, and quality trimming, reads were mapped against *Homo sapiens* (genome assembly GRCh38.p13, NCBI RefSeq assembly

accession GCF_000001405.39) and *Bos taurus* (genome assembly ARS-UCD1.2, NCBI RefSeq assembly accession number GCF_002263795.1) to remove contaminant reads (Bowtie2 version 2.3.5.1 [182]; default settings except – sensitive, end-to-end, random seed 123123). Many of these steps can lead to broken read pairs. When this occurred, broken read pairs were repaired with the FixPairs program[1] called as part of the QC pipeline script). The qc pipeline script was version 0.8.2 and is available on GitHub https://github.com/mooreryan/qc, and was run with Ruby version 2.6.5p114 and Java OpenJDK version 11.0.5.

A.2.2 Generating peptide data

De novo peptide assembly directly from the reads was performed with Plass [357] version `c4f7b` with the default settings. Each sample was assembled individually. In addition to individual samples, the eleven fecal viromes were co-assembled.

Next, MMseqs2 version `5ae55` was used to cluster all the peptides generated in the Plass assemblies (including each individual sample, plus the viral co-assembly) (default settings except–coverage mode: 1, minimum sequence identity 0.95, alignment coverage 0.70) [358]. Cluster centroids were treated as operational taxonomic units (OTUs) going forward.

To assign a count to each of the OTUs, QC reads from each sample were mapped to OTUs using MMseqs2's map submodule (default settings except – coverage: 0.95, coverage mode: 2, minimum sequence identity: 0.9, maximum sequences: 50, split memory limit: 85g). This mapping was converted to btab format with `mmseqs convertalis` command. All btab tables were collated into a final count table, with special care taken to avoid double counting and to select only the best mappings for counting. First, "best" hits were selected from the hit tables according the following sort order: bit score (higher is better), percent identity (higher is better), E-value (lower is better), target sequence coverage (higher is better), alignment length (higher is better). Any ties are broken by proceeding to the next lowest level. Bit score is used

[1] Available on GitHub: https://github.com/mooreryan/FixPairs

as a determiner of "best" hit in most cases except when two hits have the same bit score, then percent identity is used, and so on down the line of sorting criteria until the tie is broken. Finally, the count is corrected for any double counting induced by both reads from a single fragment mapping to the same OTU. These last two steps were done with custom scripts written in the Crystal language version 0.31.1.

A.2.3 RNRs from Plass assemblies

MMseqs2 version `e1a1c` was used to search the Plass assembly OTUs against ribonucleotide reductase (RNR) sequences from the RNRdb [214] (default settings except – max sequences: 300, number of iterations: 3, starting sensitivity: 1; sensitivity steps: 3, sensitivity: 7, max accept: 1, format mode: 2). The query sequences were the Plass assembly OTUs and the target sequences were the sequences from the RNRdb. Due to the set up of this homology search, significant hits are treated as the putative RNR sequences.

Post-homology search filtering of putative RNRs was done with PASV version 1.3.0 (library version 0.5.0). Class I α RNR PASV options include using Clustal Omega [344] as the aligner, key residues of 437, 439, 441, 462, and 621 with respect to the *E. coli* reference (Escherichia coli str. K-12 substr. W3110, ribonucleoside diphosphate reductase 1, alpha subunit) and start and end positions of 437 and 625, respectively. Class III RNRs use a different PASV profile with key residues, 79, 290, 543, 546, 561, 564, 580, and no start and end region. The Class III reference was NRDD_BPT4 Anaerobic ribonucleoside-triphosphate reductase. Each sequence was assigned a "signature" based on the PASV results for each of the searches. Sequences that had `NCECP` were considered putative Class I α RNRs. Sequences that had `CCCCCCG` were considered putative Class III RNRs. These sequences went through an additional round of manual validation.

For any Plass OTU identified as an RNR sequence, the abundance of that OTU as calculated above is used for the abundance of that RNR.

Table A.1: Cattle microbiome sequencing yield

Fraction	Location	Samples	Reads (M)		Bases (Gb)	
			Raw	QC	Raw	QC
cellular	fecal	17	413	410	207	137
cellular	hide	17	445	409	223	101
viral	fecal	11	163	155	49	34
Total		*45*	*1021*	*974*	*480*	*272*

To generate similarity scores between RNR sequences, multiple sequence alignments were constructed. Class III RNRs were clustered at 95% identity over 80% of the alignment length using MMseqs2 version `45111`. All classes of RNR were aligned (each class aligned separately) using the MAFFT plugin in Geneious, then 95% gap columns were masked. Sequence identity was calculated with a custom program, `msa_pairwise`[2] version 0.2.0 using option "identity".

A.3 Sequencing yield

In total, cattle microbiome samples were sequenced: 17 each of cellular fraction fecal and hide samples, and 11 fecal viromes. Total sequencing yield was 480 gigabases (Gb) (after quality control and read merging, 272 Gb) (Table A.1).

[2] Available on GitHub: https://github.com/mooreryan/bio_ballyhoo