THE UNIVERSITY OF CHICAGO

COMBINING METAGENOMICS WITH STRUCTURAL BIOINFORMATICS REVEALS THE SELECTIVE PRESSURES DRIVING PROTEIN EVOLUTION IN GLOBALLY-PREVALENT MICROBIAL POPULATIONS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
EVAN KIEFL

CHICAGO, ILLINOIS
JUNE 2022

I dedicate this work to my grandfather, Kenneth Laine.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Tables from are available for download at `https://figshare.com/articles/SAR11_SAAVs_Tables_and_Figures/5254537` (Chapter 3) and `https://doi.org/10.6084/m9.figshare.19363997` (Chapter 4). Page numbers refer to table captions.

ACKNOWLEDGMENTS

I am extremely thankful to my family for providing their nurturing support that has made this journey possible. Thank you Dad, for piquing my interest in science from a young age. Thank you Mom, for your dedication to my success and always believing in me. And thank you Jana, for being my biggest fan.

I am also very grateful for my supervisor. Meren, you gave me the space I needed to explore and discovery independently, but more importantly you provided accurate guidance whenever I needed it. You also taught me about open-source software and the power of unorthodoxy. Thank you for your efforts in my development as a scientist and software developer.

ABSTRACT

Microbes play important roles in disease, human health, and climate change. Understanding how environmental selective forces shape their evolution underpins our ability to prevent, promote, and engineer their behavior. The genetic diversity of microbial populations can be quantified with metagenomics, however, such diversity represents the outcome of both stochastic and selective forces, making it difficult to identify whether variants are maintained by adaptive, neutral, or purifying processes. This is partly due to the reliance on gene sequences to interpret variants, which disregards the physical properties of three-dimensional gene products that define the functional landscape on which selection acts. Although it is understood that the accuracy of sequence-based evolutionary models improves by integrating structural information of the encoded protein, including structural bioinformatics into metagenomic analyses is hampered by the absence of computational tools that allow researchers to seamlessly integrate these traditionally distinct data types. In my dissertation, I bridge this disconnect by developing *anvi'o structure*, a computational tool for the analysis and visualization of metagenomic sequence variants in the context of predicted protein structures and binding sites. Taking a marine microbial population as a model system, I illustrate how structure-informed analyses yield insight into the evolutionary relationship between microbes and their environments that can only be learned by combining metagenomics with structural biology. Overall, my work sheds light on how environments induce selective pressures that in turn impact the genetic diversity of populations, and provides a software tool that enables the community to employ similar analyses on different microbial systems.

CHAPTER 1

INTRODUCTION

1.1 DNA is the lens with which we study microbial life

Microbes are so abundant on earth that it is estimated they outweigh animals by 154:1 [Bar-On et al., 2018]. Their prevalence places them center stage in many of the critical biogeochemical processes on earth. Indeed, even the pleasure of an oxygen-filled atmosphere is owed to prototypical photosynthetic microbes that oxygenated our atmosphere 2.4 billion years ago [Lyons et al., 2014], and still today marine phytoplankton contribute around half of our atmospheric oxygen [Field et al., 1998]. Microbes are perhaps even more diverse than they are abundant, comprising the vast majority of the tree of life [Hug et al., 2016], and persisting in an incredible diversity of environmental niches. To put it plainly, microbes have colonized every niche Earth has to offer: from the extreme heat and pressures found in deep-sea thermal vents [Jannasch and Mottl, 1985, Dick et al., 2013], to the intestinal tract of humans [Fan and Pedersen, 2021, Kho and Lal, 2018, Sender et al., 2016], microbes are unreasonably proficient at finding a place to call home, which they manage to do by evolving lifestyles that suit their environment. The scientific discipline that investigates the relationship between microbes and their environment is broadly referred to as *microbial ecology*. However, it has traditionally been very difficult to quantify microbial ecosystems with any depth of resolution since most cells look the same under a microscope. For this reason, even basic questions such as “who is where and with what abundance?” had remained unresolved until the advent of DNA sequencing.

A draft of the first human genome was sequenced in 2001, signifying a landmark milestone in DNA sequencing technology [Lander et al., 2001]. Yet this achievement was just the start of a sequencing revolution. Ever since, the cost of sequencing has decreased at a rate that far exceeds Moore’s law (Figure 1.1). In 2001, sequencing a megabase pair of DNA sequence cost $10,000 USD. In 2020 the cost was just ten cents, representing a six order of magnitude reduction in costs over a two decade period [NIH]. The affordability of sequencing has had pervading implications across the biological sciences, and some fields, such as microbial ecology, have been profoundly shaped by this technological progress. Indeed, sequencing the DNA of microbes in their natural habitats is the primary means of measuring natural microbial ecosystems, effectively providing a lens into the evolution and lifestyles of microbes.

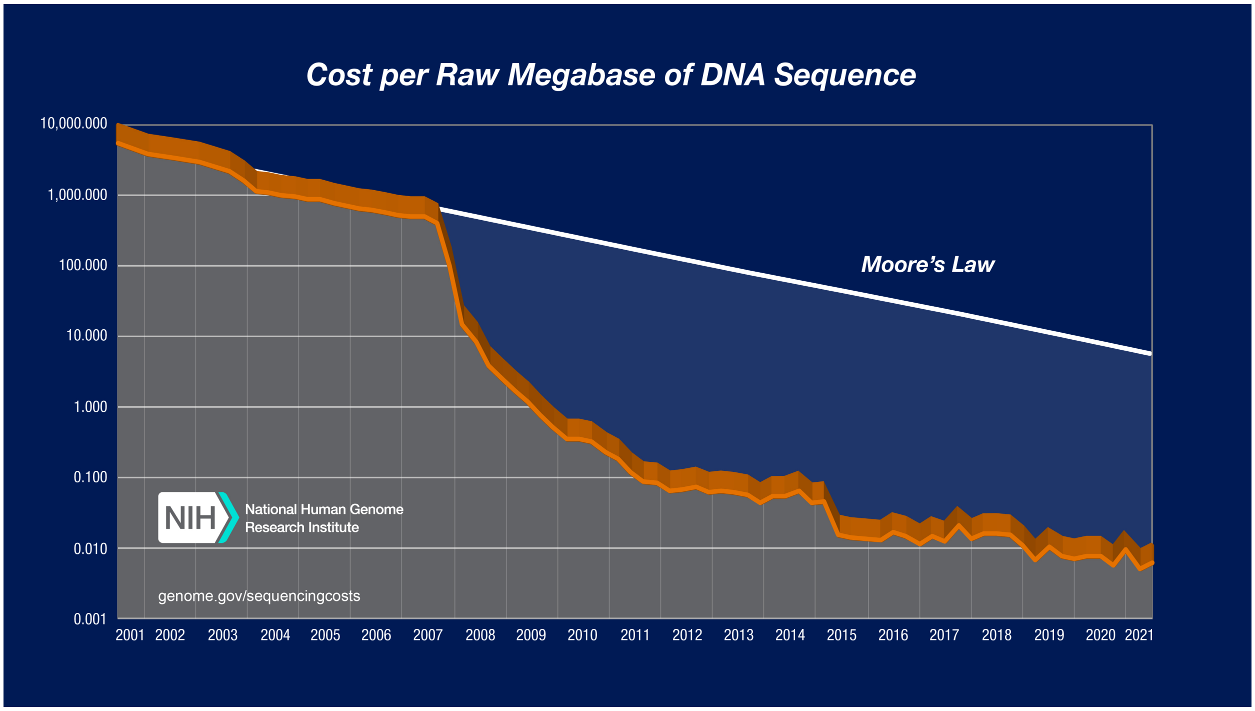


Figure 1.1: The cost of sequencing a raw megabase of DNA sequence. Figure taken from `https://www.genome.gov/sequencingcostsdata` [NIH].

1.2 Studying within-population diversity with metagenomics

Generating large sequencing datasets in microbial ecology has become increasingly practical as sequence costs drop. Once unaffordable, it is now commonplace to sequence the totality of DNA in a complex microbial sample, which is broadly known as metagenomics [Quince et al., 2017]. The widespread use of sequencing of environmental samples with metagenomics has catalyzed a revolution in the identification and characterization of uncultured clades of life [Sunagawa et al., 2015, Human Microbiome Project Consortium, 2012a, Hug et al., 2016, Delmont et al., 2018], yet it also offers unique opportunities to study the genetic diversity of naturally occurring populations at unparalleled resolution [Denef, 2019]. In particular, the alignment of short reads derived from one or more environments to a reference genome provides direct access to environment-specific single nucleotide variants (SNVs), and analysis of these variants has the potential to uncover evolutionary processes occurring within natural populations that can be directly associated to environmental conditions across space and time [Denef, 2019]. This approach has distinct advantages compared to identifying polymorphisms via alignment between two or more isolated genomes. This is because the sparsity of sequenced genomes typically demands inter-species comparisons, except for highly studied model organisms such as *E. coli*, in which thousands of genomes have been sequenced. Even then, such genomes are sampled in a culture-biased manner and therefore under-represent natural diversity. In contrast, reads aligned from metagenomic read recruitment experiments typically share greater than 95% average nucleotide identity (ANI) throughout the genome [Bendall et al., 2016, Kashtan et al., 2014, Konstantinidis and DeLong, 2008, Oh et al., 2011b, Caro-Quintero and Konstantinidis, 2012], which matches current species boundary definitions [Konstantinidis and Tiedje, 2005, Varghese et al., 2015, Jain et al., 2018]. Aligned reads are therefore likely sampled from members of a single sequence-discrete population (see notable exception [Delmont et al., 2019]), and depending on the population's environmental abundance, an immense number of its members may contribute reads in a culture-unbiased manner. This makes metagenomic read recruitment well-suited to densely sample natural populations, providing a lens into their genomic heterogeneity.

1.3 Protein structures are a missing link in metagenomic analysis

Genetic diversity is commonly quantified through single nucleotide variants (SNVs), *i.e.* nucleotide positions in the reference genome where the aligned reads exhibit minor allele frequencies large enough to be distinguished from sequencing error [Denef, 2019]. SNVs can be leveraged in a number of ways, including to quantify evolutionary processes through statistics such as nucleotide diversity, fixation index, and rates of synonymous and nonsynonymous polymorphism (for a review see [Garud and Pollard, 2020]).

Given the critical role that structure plays in the sequence-structure-function paradigm [Anfinsen, 1973], integrating protein structure into classically sequence-based frameworks is now commonplace in fields outside of microbial ecology, such as protein evolution, where it is now broadly appreciated that "bringing molecules back into molecular evolution" creates a unified view of protein evolution that increases the accuracy of evolutionary models and data-driven inference [Wilke, 2012, Harms and Thornton, 2013, Sikosek and Chan, 2014]. Furthermore, with the advent of modern de novo protein structure prediction capabilities, such as DeepMind's AlphaFold [Jumper et al., 2021], it will soon be a reality that the majority of novel coding sequences uncovered with metagenomics will have high quality structure predictions that are either readily available or can be easily calculated. Yet in general, methods for studying genetic variation in metagenomics continue to be dominated by purely sequenced-based approaches, where as a matter of practical convenience, we tend to "treat molecular sequences as mere strings of letters, the patterns of which carry the traces of historical processes, rather than as functioning objects for which the physical properties determine their behavior" [Harms and Thornton, 2013].

Although it is clear that structure-informed analyses yield insight into evolutionary processes that cannot be learned with sequence alone, including structural bioinformatics into metagenomic analyses is hampered by the absence of computational tools that allow users to seamlessly integrate these traditionally distinct data types. In particular, there exists a need for tools that allow users to interactively analyze and visualize patterns of polymorphism in environmental populations in the context of protein structures. Currently, researchers interested in doing so must face the challenges of interfacing the inputs and outputs of myriad programs that do one or more of: sequence quality filtering, metagenomic read recruitment, open reading frame (ORF) prediction, sequence variant calling and filtering, protein structure prediction, and interactive protein visualization. The totality of this effort inhibits researchers from interactively browsing, filtering, and visualizing genetic diversity of populations in the context of protein structure in a high throughput manner.

1.4 A data-rich field demands powerful computational tools

With the advent of affordable sequencing, many exceedingly large, terabyte scale datasets are now publicly available [Sunagawa et al., 2015, Salazar et al., 2019, Human Microbiome Project Consortium, 2012b,a, Yachida et al., 2019, Zhernakova et al., 2016, Zeevi et al., 2015], which has consequently has led to a dramatic shift in the needs of the microbial ecologist, who now operates in a data-rich landscape. Microbial ecology is now in the middle of an inflection point where more data exists than we have scientists to analyze it. So fertile are the soils that every investigation yields new discovery. And the generation of datasets is accelerating, creating exponentially more opportunity [Abdill et al., 2022].

How are microbial ecologists expected to keep up with this rate of data? In my opinion, we are doing a poor job. As an analogy, if each dataset is a gold mine, we barely enter the mine shaft before moving our operations to a bigger and more recently discovered mine. For example, the sequencing data used in this dissertation comes almost entirely from a public dataset known as the Tara Oceans Project metagenomes [Sunagawa et al., 2015], which contains the DNA of 40 million open reading frames from thousands of microbial species, data ultimately deriving from around 2.4 trillion raw nucleotide base pairs. I conjecture this dataset alone could yield 1,000 more dissertations before its scientific yield would feel saturated. Yet this is unlikely to ever happen, because datasets ten times larger will likely be generated in the next decade.

Certainly it is not a bad situation to be surrounded by 'too much' data. Indeed, a surplus of data is desired: until recently [Farley et al., 2018], macroecology has been considered a data-sparse field, where it is not uncommon for researchers to spend years acquiring datasets that are infinitesimally smaller than the big data seen in microbial ecology [Knapp et al., 2012]. Yet the complexity of analyses, pipelines, and algorithms required to make sense of big data grows accordingly, and as a field, microbial ecologists have not kept pace.

How do we address this growing gap between availability of data and proper and efficient use of it? No single answer exists, however what is certainly true is that developing computational tools that enable researchers aids tremendously in our navigation of this data-rich landscape.

1.5 Anvi'o

Anvi'o is a multi-'omics software platform that helps microbial ecologists draw scientific insight from the swaths of data that now surround them. It is primarily used for genomics, genome-resolved metagenomics and metatranscriptomics, pangenomics, metapangenomics, phylogenomics and microbial population genetics [Eren et al., 2021]. Its codebase currently stands at >140,000 lines of code and is used by hundreds of biologists in more than 40 countries around the world.

Anvi'o operates via a network of around 160 command-line programs that can be chained together, and which all operate on shared database structures that are passed from program to program. This network of programs means users can create their own workflows traversing a linear path within this network. By providing full interactive access to the underlying data at each intermediate step, users can explore their data to determine where the analysis should go next. This is in contrast to traditional data pipelines that take in raw input and spit out summary tables and interactive figures. On one hand this approach is good because it requires very little computational expertise to use, which is appreciated in a field where computational sciences have not been taught as part of its formal curriculum, but it suffers from rigidity because pipelines constrain the kinds of questions that can be asked to those envisioned by the developer. The modular workflow style of anvi'o is also in contrast to creating custom workflows, which requires substantial data wrangling skills in matching the output of one program to the input of the next and requiring in-depth knowledge about each program. Anvi'o circumvents this requirement by passing around intermediate data objects that are created, modified, accessed, and generally operated upon by anvi'o programs.

1.6 Thesis topics

This body of work details my contributions to microbial ecology as both a software developer and a researcher. My research has focused on studying how changing environments dynamically shape the selective pressures that drive genetic diversity within natural microbial populations. While there exists a large repertoire of bioinformatic tools upon which my research is based, my specific research questions have demanded tools that did not exist priorly. For this reason, my graduate studies have revolved around developing computational tools in tandem with my research. As such, I have contributed extensively to anvi'o as a primary developer in an attempt to address what I see as a lack of computational solutions that enable effective data analysis and integration of distinct data types within the field. I've also invested extensive effort into increasing the accessibility of anvi'o to its users, who often do not have computational expertise, by documenting code, writing tutorials online, and creating extensive reproducible workflows for my publications, essentially paving way for others to apply similar analyses with their own data. The totality of these contributions are formally recognized in Chapter 2, which contains a recent commentary of anvi'o that summarizes the progress that the many developers of anvi'o have made over the last 5 years.

Throughout my studies I have focused on a model system known as the 1a.3.V subclade of SAR11, an abundant and widespread bacteria found in surface oceans. Chapter 3 introduces the first characterization of this subclade, which we defined based on geographical co-occurrence patterns, phylogenomics, and pangenomics. By analyzing how single-amino acid variants (SAAVs) in 1a.3.V distribute across environments, we partitioned this subclade into distinct 'proteotypes' that display distribution patterns that match ocean current temperatures. These proteotype distribution patterns favor the hypothesis that temperature and/or its covariables critically affect the maintenance of genetic variation in 1a.3.V.

The results of Chapter 3 indicate that environmentally-mediated selection induces selective pressures that shape the genetic diversity of 1a.3.V. However, which variants are under the influence of which selective pressures? This is not easily answered with analyses that rely solely on sequences. To better address this question, Chapter 4 introduces *anvi'o structure*, a new tool that provides an automated, scalable, and interactive work environment for analyzing and visualizing metagenomic variants with respect to predicted protein structures and ligand binding sites. We applied anvi'o structure to the 1a.3.V model system, and explored patterns of synonymous and nonsynonymous variation with respect to the structural parameters relative solvent accessibility (RSA) and distance-to-ligand (DTL). We revealed that as much as 59% of nonsynonymous genetic variance can be explained by these two parameters alone, identified an instance where nitrogen availability dictated how close nonsynonymous variants were 'allowed' to get to the active site of a protein regulated by nitrogen availability, and identified that synonymous polymorphism, though not affecting amino acid sequences, distributes according to structural protein features.

Overall, this work trailblazes new analyses that blend metagenomics with structural biology, demonstrating how such an integration increases the interpretability of evolutionary processes that drive polymorphism in their natural environments, and provides new computational tools so that the community can perform similar analyses with their own microecological systems with a fraction of the effort.

CHAPTER 2

COMMUNITY-LED, INTEGRATED, REPRODUCIBLE MULTI-OMICS WITH ANVI'O

This chapter is derived from the following publication:

Eren, A. Murat, **Evan Kiefl**, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, et al. 2021. "Community-Led, Integrated, Reproducible Multi-Omics with Anvi'o." Nature Microbiology 6 (1): 3–6.

2.1 Author contributions

This is a perspective piece describing the progress that has been made in the development of anvi'o over the last 5 years, which is represented by >140,000 lines of code that has been contributed by 34 authors from around the world. This code has directly or indirectly influenced the analysis of hundreds of papers, and the codebase itself contains substantial contributions for several PhD dissertations, mine being just one. With that in mind, it is impossible to accurately attribute each author's contributions. Nevertheless, what follows is a specific summary of my personal contributions that this paper recognizes.

My specific contributions include collaborating with microbial ecologists and computational biologists to implement feature requests, diagnose errors, and fix bugs in ≈150 GitHub issues and pull requests. I have redesigned mission-critical code, decreasing compute times from >100 hours to <1 hour for terabyte-scale analyses. I have spearheaded the development of protein structure prediction, analysis, and visualization capabilities within anvi'o. I have made significant contributions to the storage, processing, and analysis strategies related to single nucleotide variants (SNVs), single codon variants (SCVs), single amino acid variants (SAAVs), and insertions/deletions (INDELs). Overall, I have played a primary role in the overall health and maintenance of the codebase in what has totaled >1,600 GitHub commits. I have established docstring conventions for the codebase, contributing >3,000 lines of docstrings, and authored >20,000 words of user tutorials ((1) https://merenlab.org/2015/07/20/analyzing-variability, (2) https://merenlab.org/tutorials/infant-gut, (3) https://merenlab.org/2018/09/04/getting-started-with-anvio-structure, (4) https://merenlab.org/2020/07/22/interacdome). I've also mentored junior developers through the program development life-cycle via code reviews and pair programming.

2.2 Discussion

Generating hundreds of millions of sequences from a microbial habitat is now commonplace for many microbiologists [White et al., 2016]. While the massive data streams offer detailed snapshots of the lifestyles of microorganisms, this data revolution in microbiology means that a new generation of computational tools is needed to empower life scientists in the era of multi-omics.

To meet the growing computational needs of the life sciences, computer scientists and bioinformaticians have created thousands of software tools [Callahan et al., 2018]. These software fall into two general categories: 'essential tools' that implement functions fundamental to most bioinformatics tasks, and 'workflows' that make specific analytic strategies accessible.

If a comprehensive microbial 'omics investigation is a sophisticated dish, then essential tools are the kitchenware needed to cook. A chef can combine them in unique ways to answer any question, yet such freedom in data analysis not only requires the mastery of each essential tool but also demands experience in data wrangling and fluency in the command line environment to match the output format of one tool to the input requirements of the next. This barrier is overcome by workflows, which implement popular analysis strategies and make them accessible to those who have limited training in computation. If a comprehensive microbial ‘omics investigation is a sophisticated dish, then each ‘omics workflow is a recipe that turns raw material into a specific meal. For instance, a workflow for ‘pangenomics’ would typically take in a set of genomes and (1) identify open reading frames in all input genomes, (2) reciprocally align all translated amino acid sequences, (3) identify gene clusters by resolving pairwise sequence homology across all genes, and (4) report the distribution of gene clusters across genomes. By doing so, a software that implements pangenomics, such as Roary [Page et al., 2015], would seamlessly run multiple essential tools consecutively, resolve input/output requirements of each, and address various ad hoc computational challenges to concoct a pangenome. Popular efforts to make accessible workflows that form the backbone of ‘omics-based microbiological studies include the Galaxy platform [Jalili et al., 2020], bioBakery software collection [McIver et al., 2018], M-Tools (*i.e.*, GroopM [Imelfort et al., 2014], CheckM [Parks et al., 2015]), and KBase [Arkin et al., 2018]. While ‘omics workflows conveniently summarise raw data into tables and figures, the ability to analyse data beyond pre-defined strategies they implement continues to be largely limited to master chefs, presenting the developers of ‘omics workflows with a substantial responsibility: pre-determining the investigative routes their software enables users to traverse, which can influence how researchers interact with their data, conceivably affecting biological interpretations.

We introduced *anvi’o* (an analysis and visualisation platform for ‘omics data) as an alternative solution for microbiologists who wanted more freedom in research questions they could ask of their data [Murat Eren et al., 2015]. We started with what we regarded as the most pressing need at the time: a platform that enabled the reconstruction and interactive refinement of microbial genomes from environmental metagenomes. Fundamentals of this strategy were already established by those who pioneered genome-resolved metagenomics [Tyson et al., 2004], but interactive visualisation and editing software that would enable microbiologists to intimately work with metagenome-assembled genomes was lacking. During the past five years anvi’o has become a community-driven software platform that currently stands upon more than 90,000 lines of open-source code and supports interactive and fully integrated access to state-of-the-art ‘omics strategies including genomics, genome-resolved metagenomics and metatranscriptomics, pangenomics, metapangenomics, phylogenomics, and microbial population genetics (Figure 2.1).

Anvi’o differs from existing bioinformatics software due to its modular architecture, which enables flexibility, interactivity, reproducibility, and extensibility. To achieve this, the platform contains more than 100 interoperable programs, each of which performs individual tasks that can be combined to build new and unique analytical workflows. Anvi’o programs generate, modify, query, split, and merge anvi’o projects, which are really a set of extensible, self-contained SQLite databases. The interconnected nature of anvi’o programs which are glued together by these common data structures yields a network (`http://merenlab.org/nt`), rather than predetermined, linear paths for analysis. Through this modularity, anvi’o empowers its users to navigate through ‘omics data without imposing rigid workflows.

Integrated interactive visualisation is at the center of anvi’o and helps researchers to engage with their data in all stages of analysis. Within the same interface, an anvi’o user can visualise amino acid sequence alignments between homologous genes across multiple genomes, investigate nucleotide-level coverage patterns and variants on the same DNA segment across metagenomes, interrogate associations between the genomic abundance and transcriptomic activity of environmental microbes, display phylogenetic trees and clustering dendrograms, and more. Furthermore, users can extend anvi’o displays with project-specific external data, increasing the utility of interactive interfaces for holistic descriptions of complex systems. The anvi’o interactive interface also provides its users with the artistic freedom to change colours, sizes, and drawing styles of display objects, add annotations, or reorder data layers for detailed communication of intricate observations. Because each anvi’o project is self-contained, researchers can easily make their analyses available online as a whole or in part, thereby enabling the integration, reusability, and reproducibility of their findings beyond static figures or tables. This strategy promotes transparency by permitting community validation and scrutiny through full access to data that underlie final conclusions.

Several key studies that used anvi'o during the past few years have demonstrated the integrative capabilities of the platform by implementing a combination of 'omics strategies to facilitate in-depth analysis of naturally occurring microbial habitats. For instance, Reveillaud and Bordenstein et al. reconstructed new genomes of Wolbachia, a fastidious endosymbiont [Werren et al., 2008], from individual insect ovary metagenomes, and computed a pangenome to compare these novel genomes to an existing reference [Reveillaud et al., 2019]. They were then able to characterise the ecology of gene clusters in the environment by effectively combining metagenomics and pangenomics, discovering new members of the Wolbachia mobilome [Reveillaud et al., 2019]. Yeoman et al. combined phylogenomics and pangenomics to infer ancestral relationships between a set of cultivar and metagenome-assembled genomes through a de novo identified set of single-copy core genes [Yeoman et al., 2019]. They demonstrated the correspondence among these genomes based on gene cluster membership patterns, phylogenomic inference, and average nucleotide identity in a single display [Yeoman et al., 2019]. Delmont and Kiefl et al. characterised the population structure of a subclade of SAR11, one of the most abundant microbial populations on Earth, by describing the environmental core genes of a single genome across surface ocean metagenomes [Delmont et al., 2019]. By linking single-amino acid variants in the environment to the predicted tertiary structures of these genes, they combined microbial population genetics with protein biochemistry to shed light on distinct evolutionary processes shaping the population structures of these bacteria [Delmont et al., 2019]. Each of these studies employs unique approaches beyond well-established 'omics workflows to create rich, reproducible, and shareable data products (see `http://merenlab.org/data`).

Anvi'o does not implement strategies that take in raw data and produce summary tables or figures via a single command. As a result, anvi'o has a relatively steep learning curve.

To address this, we have written extensive online tutorials that currently exceed 120,000 words, organised free workshops for hands-on anvi'o training, and created open educational resources to learn microbial ‘omics. To interact with anvi'o users we set up an online forum and messaging service. During the past two years, more than 750 registered members of these services have engaged in technical and scientific discussions via more than 9,000 messages. But even when resources for learning are available, the journey from raw ‘omics data to biological insights often takes a significant number of atomic steps of computation. To ameliorate the burden of scale and reproducibility in big data analyses we have also introduced anvi'o workflows, which automate routine computational steps of commonly used analytical strategies in microbial ‘omics (`http://merenlab.org/anvio-workflows`). The anvi'o workflows are powered by Snakemake [Köster and Rahmann, 2012], which ensures relatively easy deployment to any computer system and automatic parallelisation of independent analysis steps. By turning raw input into data products to be analysed in the anvi'o software ecosystem, anvi'o workflows reduce the barriers for advanced use of computational resources and processing of large data streams for microbial ‘omics.

As the developers of anvi'o who strive to create an open community resource, our next big challenge is to attract bioinformaticians to consider anvi'o as a software development ecosystem they can use for their own science. Any program that reads from or writes to anvi'o projects either directly (in any modern programming language) or through anvi'o application programmer interfaces (in Python) will immediately become accessible to anvi'o users, and such applications will benefit from the data integration, interactive data visualisation, and error checking assurances anvi'o offers. As an open-source platform that empowers microbiologists by offering them integrated yet uncharted means to steer through complex ‘omics data, anvi'o welcomes its new users and contributors.

2.3 Acknowledgements

The URL `https://github.com/merenlab/anvio/blob/master/AUTHORS.txt` serves a complete list of anvi'o developers. We thank the creators of other open-source software tools for their generosity, anvi'o users for their patience with us, and Karen Lolans (0000-0003-1903-756X) for her critical reading of the manuscript and suggestions. The authors gratefully acknowledge support for anvi'o from the Simons Foundation and Alfred P. Sloan Foundation.

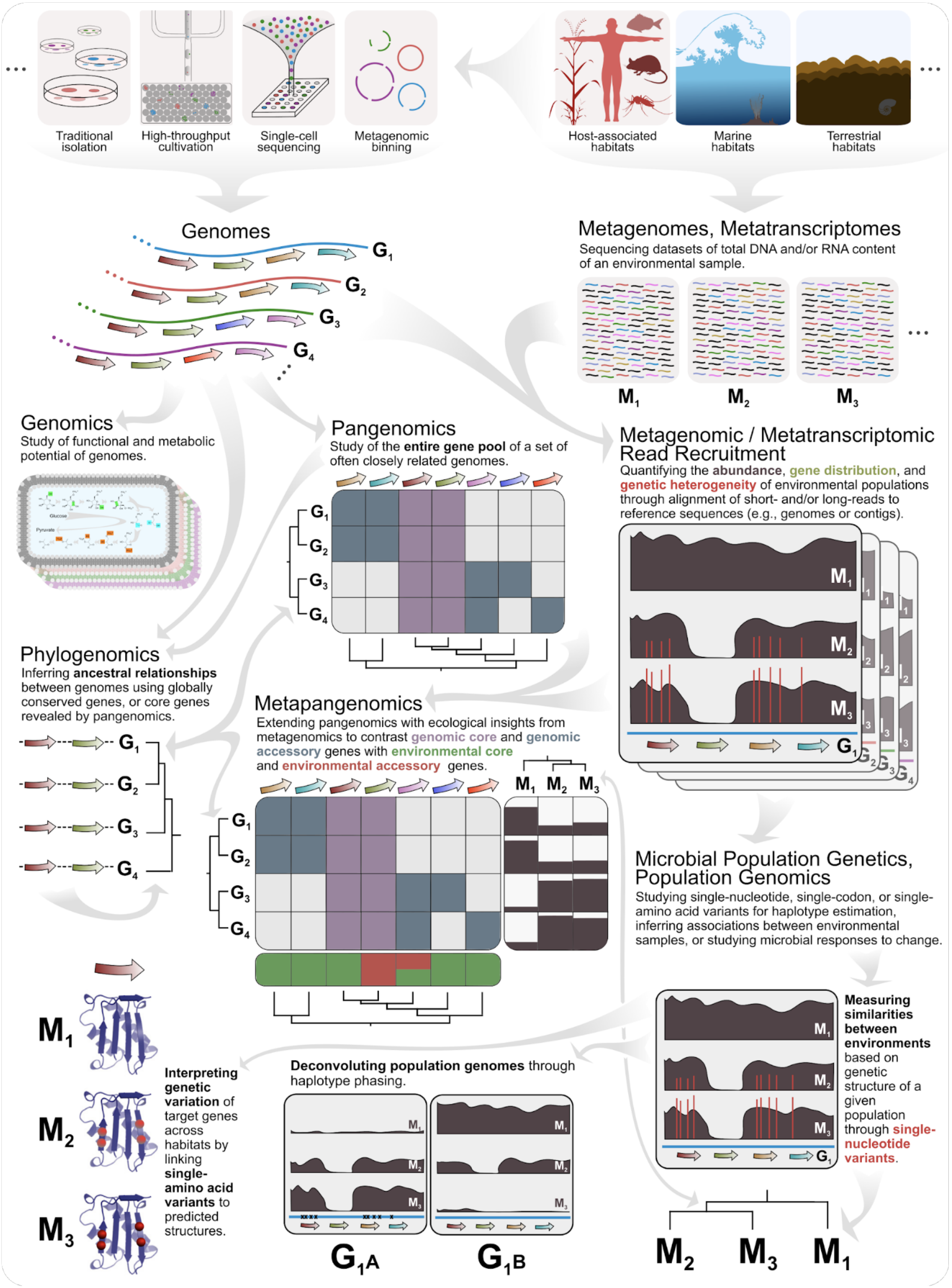

...
Traditional isolation
High-throughput cultivation
Single-cell sequencing
Metagenomic binning
Host-associated habitats
Marine habitats
Terrestrial habitats
...
Genomes
G1
G2
G3
G4
Metagenomes, Metatranscriptomes
Sequencing datasets of total DNA and/or RNA content of an environmental sample.
M1
M2
M3
...
Genomics
Study of functional and metabolic potential of genomes.
Glucose
Pyruvate
Pangenomics
Study of the entire gene pool of a set of often closely related genomes.
G1
G2
G3
G4
Metagenomic / Metatranscriptomic Read Recruitment
Quantifying the abundance, gene distribution, and genetic heterogeneity of environmental populations through alignment of short- and/or long-reads to reference sequences (e.g., genomes or contigs).
M1
M2
M3
G1
G2
G3
G4
Phylogenomics
Inferring ancestral relationships between genomes using globally conserved genes, or core genes revealed by pangenomics.
G1
G2
G3
G4
Metapangenomics
Extending pangenomics with ecological insights from metagenomics to contrast genomic core and genomic accessory genes with environmental core and environmental accessory genes.
M1
M2
M3
G1
G2
G3
G4
Microbial Population Genetics, Population Genomics
Studying single-nucleotide, single-codon, or single-amino acid variants for haplotype estimation, inferring associations between environmental samples, or studying microbial responses to change.
M1
M2
M3
Interpreting genetic variation of target genes across habitats by linking single-amino acid variants to predicted structures.
Deconvoluting population genomes through haplotype phasing.
M1
M2
M3
G1A
G1B
M1
M2
M3
G1
Measuring similarities between environments based on genetic structure of a given population through single-nucleotide variants.
M2
M3
M1

Figure 2.1: A glimpse of the interconnected nature of ’omics analysis strategies anvi’o makes accessible, and their potential applications.

CHAPTER 3

SINGLE-AMINO ACID VARIANTS REVEAL EVOLUTIONARY PROCESSES THAT SHAPE THE BIOGEOGRAPHY OF A GLOBAL SAR11 SUBCLADE

This chapter is derived from the following publication:

> Delmont, Tom O.*, **Evan Kiefl***, Ozsel Kilinc, Ozcan C. Esen, Ismail Uysal, Michael S. Rappé, Steven Giovannoni, and A. Murat Eren. 2019. "Single-Amino Acid Variants Reveal Evolutionary Processes That Shape the Biogeography of a Global SAR11 Subclade." eLife 8 (September). https://doi.org/10.7554/eLife.46497.

*** I share co-first authorship with Tom O. Delmont.**

3.1 Author contributions

TOD, EK, and AME conceived and designed the study. EK, OCE, and AME developed analysis tools to compute single-amino acid variants from metagenomes and visualize them in the context of protein structures. OK and IU developed analysis tools for machine learning. EK, TOD, and AME analyzed data, wrote the paper, prepared figures and tables, and developed the reproducible analysis workflow. MSR and SG contributed to all stages of the data analysis and interpretation. All authors reviewed and revised the drafts of the paper.

3.2 Abstract

Members of the SAR11 order Pelagibacterales dominate the surface oceans. Their extensive diversity challenges emerging operational boundaries defined for microbial 'species' and complicates efforts of population genetics to study their evolution. Here we employed single-amino acid variants (SAAVs) to investigate ecological and evolutionary forces that maintain the genomic heterogeneity within ubiquitous SAR11 populations we accessed through metagenomic read recruitment using a single isolate genome. Integrating amino acid and protein biochemistry with metagenomics revealed that systematic purifying selection against deleterious variants governs non-synonymous variation among very closely related populations of SAR11. SAAVs partitioned metagenomes into two main groups matching large-scale oceanic current temperatures, and six finer proteotypes that connect distant oceanic regions. These findings suggest that environmentally-mediated selection plays a critical role in the journey of cosmopolitan surface ocean microbial populations, and the idea "everything is everywhere but the environment selects" has credence even at the finest resolutions.

3.3 Introduction

The SAR11 order Pelagibacterales [Thrash et al., 2011, Ferla et al., 2013] is one of the most ubiquitous free-living lineages of heterotrophic bacteria in the world's oceans [Giovannoni et al., 1990, Morris et al., 2002, Carlson et al., 2009, Eiler et al., 2009, Treusch et al., 2009]. Successful cultivation efforts and single amplified genomes from the environment have led to studies revealing their critical role in marine carbon cycling [Rappé et al., 2002, Giovannoni et al., 2005, Stingl et al., 2007, Oh et al., 2011a, Tsementzi et al., 2016, White et al., 2019], and environmental sequencing surveys have offered detailed insights into the ecology of this ancient branch of life in aquatic environments across the globe [Zinger et al., 2011, Brown et al., 2012].

The evolution of SAR11 is an active area of research [Giovannoni, 2017] that is critically important to understanding the determinants of its remarkable ability to maintain abundant populations in the global ocean. The evolutionary origins of SAR11 and thus its precise placement in the Tree of Life is debated [Thrash et al., 2011, Rodríguez-Ezpeleta and Embley, 2012, Ferla et al., 2013, Viklund et al., 2013]. Our understanding of the evolutionary processes that define the biogeography of SAR11 cells is not complete: at the level of major SAR11 clades, previous studies have attributed markedly distinct patterns of distribution in the global ocean to both niche-based [Brown et al., 2012, Eren et al., 2013] and neutral processes [Manrique and Jones, 2017]. At the level of individual populations, a key simulation by Hellweger et al. (2014) showed that the intra-population sequence divergence that reflects the geographic patterns of distribution for SAR11 cells could emerge solely as a function of ocean currents, without selection [Hellweger et al., 2014]. Between the extremes of inter-clade and intra-population diversity lies a wealth of variation that potentially can yield insights into the ecological and genetic forces that determine genomic diversity and fitness between closely-related, naturally occurring SAR11 populations.

High-throughput sequencing of metagenomes provides access to genome-wide heterogeneity within environmental populations [Simmons et al., 2008], and current computational strategies can reveal associations between ecological parameters and microdiversity patterns at various levels of resolution [Murat Eren et al., 2015, Scholz et al., 2016, Nayfach et al., 2016, Costea et al., 2017, Truong et al., 2017]. However, SAR11 poses multiple challenges for such investigations, including their remarkable intra-population genomic diversity and the limited success of reconstructing SAR11 genomes from metagenomic data. Comprehensive investigations of the genetic contents of naturally occurring microbial populations (see [Denef, 2019] for a review) often rely on population genomes directly reconstructed from metagenomes [Simmons et al., 2008, Bendall et al., 2016, Anderson et al., 2017, Garcia et al., 2018]. While advances in genome-resolved metagenomics have made microbial clades more accessible without cultivation [Spang et al., 2015, Brown et al., 2015, Anantharaman et al., 2016], reconstructing SAR11 genomes from the surface ocean remains a difficult endeavor, as evident in recent comprehensive surveys of metagenome-assembled genomes (MAGs) from seawater samples from around the globe [Tully et al., 2018, Delmont et al., 2018]. In the absence of population genomes recovered directly from the environment, genomes from isolates can also offer insights into environmental populations through genome-wide recruitment analyses in which short metagenomic reads are aligned to a reference [Denef, 2019].

Using metagenomic read recruitment to investigate the structure of environmental populations is confounded by the challenge of defining the boundaries of microbial populations. Without an established species concept in microbiology, defining units of microbial diversity and their boundaries is a significant challenge (see [Jesse Shapiro, 2018] and [Cohan, 2019] for discussions). Nevertheless, from analyses of isolated microbial strains with formal taxonomic descriptions, a genome-wide average nucleotide identity (gANI) cutoff of 95% emerged as an operational delineation of species [Konstantinidis and Tiedje, 2005, Varghese et al., 2015] and was confirmed in a recent analysis of eight billion pairwise comparisons of whole genomes [Jain et al., 2018]. Both gANI calculations using complete genomes, as well as the average nucleotide identity of metagenomic short reads (ANIr) recruited from environmental metagenomes using reference genomes, show an interesting discontinuity among sequence-discrete populations at sequence identity levels between 80% and 90-95% [Konstantinidis and DeLong, 2008, Caro-Quintero and Konstantinidis, 2012, Jain et al., 2018]. Regardless of their theoretical significance, these cutoffs are essential for multiple practical purposes, such as the identification and subsequent exclusion of metagenomic reads that originate from non-target environmental populations, to avoid inflating variants arising from contaminating non-specific reads in microbial population genetics studies.

Interestingly, the boundaries of environmental SAR11 populations appear to not comply with the 95% ANIr cutoff. For instance, Tsementzi et al (2016) observed substantial sequence diversity within sequence-discrete SAR11 subclades in the environment, and suggested that an ANIr as low as 92% would be required to adequately define the boundaries of the SAR11 populations recovered in their study [Tsementzi et al., 2016]. These findings are consistent with a comprehensive study of isolate genomes and marine metagenomes by Nayfach et al (2016), which suggested that SAR11 is one of the most genetically heterogeneous marine microbial clades [Nayfach et al., 2016]. The substantial sequence diversity within environmental SAR11 populations not only explains the absence of SAR11 population genomes in genome-resolved metagenomics studies, but also challenges conventional approaches to the study of population genetics in microorganisms. For instance, the multiple occurrence of single-nucleotide variants in individual codon positions would render commonly used computational strategies that classify synonymous and non-synonymous variations based on independent nucleotide sites (such as in [Schloissnig et al., 2013, Bendall et al., 2016]) unfeasible. Despite these challenges, SAR11, with its ubiquity in surface seawater samples, extensive diversity in sequence space, and unique evolutionary history, remains one of the exciting puzzles of contemporary microbiology. Here we investigated the evolutionary processes that maintain genetic diversity within a natural SAR11 lineage accessible through a single isolate genome that recruited more than 1% of surface ocean metagenomic reads from a global dataset. Using single-amino acid variants, we were able to (1) delineate multiple proteotypes whose distributions were more closely linked to large-scale oceanic current temperatures than they were to geographic proximity, and (2) resolve positive and negative selection mediated by temperature and its co-variables. Our findings suggest that environmentally mediated selection, rather than neutral processes, dominate the biogeographic partitioning of SAR11 at fine scales of taxonomic resolution. Our study also offers new computational approaches to characterize variation within complex microbial populations, including additional means to integrate amino acid and protein biochemistry into microbial population genetics.

3.4 Results and Discussion

To find the most appropriate SAR11 isolate genome to study the population genetics of naturally occurring SAR11, we used the complete genomes of 21 SAR11 isolates in a competitive recruitment of short reads from 103 metagenomes. Most of these metagenomes were from the TARA Oceans Project [Sunagawa et al., 2015], and correspond to 93 stations across four oceans and two seas. We also included an additional 10 metagenomes from the Ocean Sampling Day Project [Kopf et al., 2015] to cover high-latitude areas of the Northern hemisphere. All metagenomes correspond to small planktonic cells (0.2-3m in size) from the surface (0-15 meters depth; n=71) and deep chlorophyll maximum (17-95 meters depth; n=32) layers of the water column (Table 3.1). The isolates we used belonged to SAR11 subclades Ia.1 (n=6), Ia.3 (n=11), II (n=1), IIIa (n=2) and the related alphaproteobacterium Va (n=1) (Table 3.1), which collectively recruited 1,029,716,339 reads from all metagenomes, or 3.3% of the dataset (Table 3.2).

3.4.1 The metapangenome of SAR11

To investigate associations between ecology and gene content of SAR11 lineages, we first performed a pangenomic analysis in conjunction with read recruitment from the metagenomic data. The pangenome of SAR11 genomes consisted of all 29,719 genes grouped into 6,175 gene clusters (Table 3.3). The clustering of genomes based on shared gene clusters matched that of the previously described phylogenetic clades [Grote et al., 2012] (Figure 3.1A; an interactive version of which is available at `http://anvi-server.org/p/4Q2TNo`). The SAR11 pangenome across metagenomes (*i.e.* the SAR11 metapangenome) revealed distinct distribution patterns for each clade within SAR11 (Figure 3.1A). Clade Ia recruited the most reads compared to other clades (Table 3.2), consistent with previous studies that found this clade to be highly abundant in surface seawater [Field et al., 1997, Brown et al., 2012, Eren et al., 2013, Manrique and Jones, 2017]. Gene clusters divided clade Ia into two main clusters corresponding to the high-latitude subclade Ia.1 and the low-latitude subclade Ia.3 (Figure 3.1A). While all high-latitude genomes displayed a bi-polar geographic distribution in the metagenomic dataset, gene clusters in low-latitude genomes revealed multiple sub-groups that also showed different patterns of geographic distribution (Figure 3.1A). This emphasized the need to further refine subclade 1a.3, in which each genome pair had over 98.6% sequence identity at the 16S rRNA gene level (Table 3.4). Our consideration of geographical co-occurrence patterns, phylogenomic characteristics, and pangenomic properties in this metapangenome revealed six subclades within 1a.3 with cultured representatives (Figure 3.1A). We tentatively name them SAR11 subclade 1a.3.I (HTCC7211, HTCC7214 and HTCC7217; gANI of >93% and 16S rRNA gene identity of >99.4%), 1a.3.II (HIMB5), 1a.3.III (HIMB4 and HIMB1321; gANI of 94.8% and 16S rRNA gene identity of 100%), 1a.3.IV (HTCC8051 and HTCC9022; gANI of 86.9% and 16S rRNA gene identity of 100%), 1a.3.V (HIMB83) and 1a.3.VI (HIMB122 and HIMB140; gANI of 94.6% and 16S rRNA gene identity of 99.7%). Overall, the refinement of SAR11 subclades reveals a striking agreement between phylogeny, pangenome, and the ecology of the members of the SAR11 clade Ia.

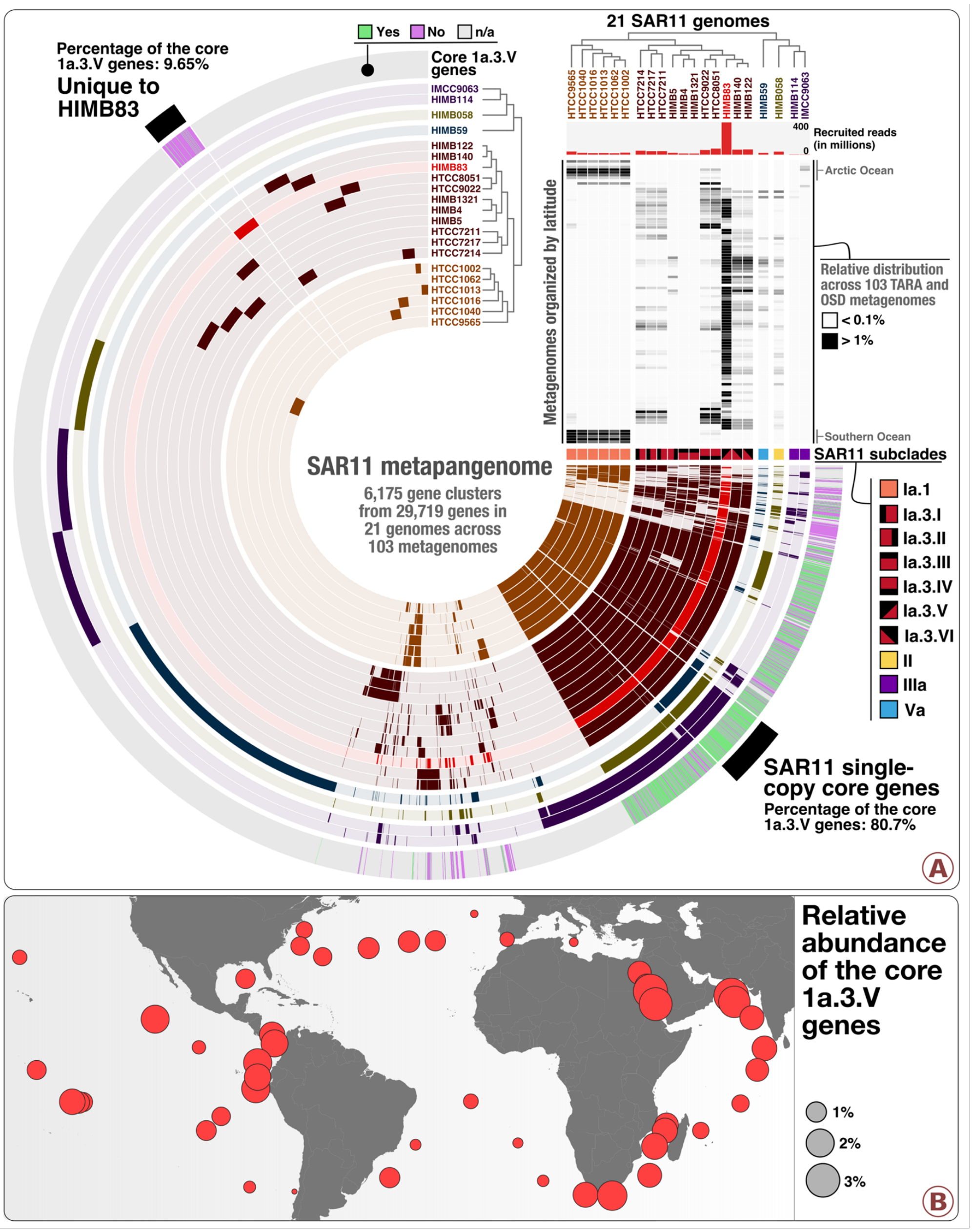


Figure 3.1: The SAR11 metapangenome. Panel A describes the pangenome of 21 SAR11 isolate genomes based on the occurrence of 6,175 gene clusters, in conjunction with their phylogeny (clade level) and relative distribution of recruited reads in 103 metagenomes ordered by latitude from the North Pole to the South Pole (top right heat map). The relative distributions were displayed for a minimum value of 0.1% and a maximum value of 1%. The layer named "Core 1a.3.V genes" displays the occurrence of the 799 core 1a.3.V genes (in green) and those found in HIMB83 but not in the 1a.3.V lineage (in purple). Panel B describes the relative distribution of reads the 799 core 1a.3.V genes recruited across surface metagenomes from TARA Oceans.

3.4.2 A remarkably abundant and widespread SAR11 lineage at low latitudes

While Ia.3 was the most abundant SAR11 subclade in our dataset, the new subclades we defined in this group differed remarkably in their competitive recruitment of short reads from metagenomes (Figure 3.1A, Table 3.2). For example, while the least abundant subclade (1a.3.II; represented by HIMB5) recruited 22.6 million reads, the most abundant one (1a.3.V; represented by HIMB83), recruited 390.9 million reads, or 1.18% of the entire metagenomic dataset (Table 3.2). For perspective, this is roughly two times more reads than the most abundant Prochlorococcus isolate genome recruited from the same dataset [Delmont and Eren, 2018]. Strain HIMB83 contains a 1.4 Mbp genome with 1,470 genes, and was isolated from coastal seawaters off Hawai'i, USA. But it also recruited large numbers of reads from locations that were distant to the source of isolation (Table 3.2). The gANI between HIMB83 and the most similar genome in our dataset, HIMB122 (1a.3.VI) was 82.6%, and the remarkable abundance of HIMB83 has also been recognized by others [Brucks, 2014, Nayfach et al., 2016]. To the best of our knowledge, 1a.3.V is the most abundant and widespread SAR11 subclade in the euphotic zone of low-latitude oceans and seas.

Although it is a member of the subclade 1a.3.V, the genomic context HIMB83 provides does not exhaustively describe the gene content of all members of 1a.3.V. Nevertheless, it gives access to the core 1a.3.V genes through read recruitment. To identify core 1a.3.V genes, we used a conservative two-step filtering approach. First, we defined a subset of the 103 metagenomes within the main ecological niche of 1a.3.V using genomic mean coverage values (Table 3.2). Our selection of 74 metagenomes in which the mean coverage of HIMB83 was >50X encompassed three oceans and two seas between -35.2° and +43.7° latitude, and water temperatures at the time of sampling between 14.1°C and 30.5°C (Figure 3.5, Table 3.1). We then defined a subset of HIMB83 genes as the core 1a.3.V genes if they occurred in all 74 metagenomes and their mean coverage in each metagenome remained within a factor of 5 of the mean coverage of all HIMB83 genes in the same metagenome. This criterion accounted for biological characteristics influencing coverage values in metagenomic surveys of the surface ocean such as cell division rates and variations in coverage as a function of changes in GC-content throughout the genomic context. Figure 3.5 displays the coverage of all HIMB83 genes across all metagenomes, and Table 3.5 reports the coverage statistics. While the 799 genes that met these criteria systematically occurred within the niche boundaries of 1a.3.V, 40% of the remaining 671 HIMB83 genes that were filtered out were present in five or fewer metagenomes and coincided with hypervariable genomic loci (Figure 3.5). Hypervariable genome regions are common features of surface ocean microbes [Coleman et al., 2006, Zaremba-Niedzwiedzka et al., 2013, Kashtan et al., 2014, Delmont and Eren, 2018] that are not readily addressed through metagenomic read recruitment but do influence pangenomic trends. Here, less than 10% of gene clusters unique to HIMB83 were among core 1a.3.V genes (Figure 3.1A), indicating HIMB83's unique genes are mostly accessory to the members of 1a.3.V. In contrast, more than 80% of gene clusters that were core to the 21 SAR11 genomes matched to the core 1a.3.V genes. The overlap between environmental core genes of 1a.3.V revealed by the metagenomic read recruitment and the genomic core of SAR11 revealed by the pangenomic analysis of isolate genomes suggests that these genes represent a large fraction of the 1a.3.V genomic backbone (Figure 3.1A). Core 1a.3.V genes recruited on average 1.25% of reads in the 74 metagenomes (Figure 3.1B, Table 3.5). The broad geographic prevalence of core 1a.3.V genes represents a unique opportunity to study the population genetics of an abundant marine microbial subclade across distant geographies.

3.4.3 SAR11 subclade 1a.3.V maintains a substantial amount of genomic heterogeneity

To investigate the amount of genomic heterogeneity within 1a.3.V, we first studied individual short reads that the HIMB83 genome recruited from metagenomes. The percent identity of reads that matched to the 799 core 1a.3.V genes ranged from 88% to 100% (Figure 3.2), which is considerably more diverse than those observed in similar reference-based metagenomic studies [Konstantinidis and DeLong, 2008, Tsementzi et al., 2016, Meziti et al., 2019]. Notably, we also observed similar trends for the other SAR11 genomes included in this study (Figure 3.6), suggesting that the relatively high sequence diversity observed among core 1a.3.V genes may be a characteristic shared with other SAR11 lineages in the surface ocean.

Overall, our data confirm that ANIr values of >95% used previously to delineate sequence-discrete populations does not apply to SAR11. One immediate implication of this substantial amount of sequence diversity that defies previous empirical observations is our inability to explicitly define what we are accessing in the environment. This challenge is partially because a precise and exhaustive description of what constitutes a 'population' remains elusive [Cohan and Perry, 2007, Shapiro and Polz, 2015, Cohan, 2019], which creates significant practical challenges [Rocha, 2018], such as the accurate determination of the boundaries of naturally occurring microbial populations especially in metagenomic read recruitment results. Nevertheless, the term 'population' is frequently used in literature [Simmons et al., 2008, Kashtan et al., 2014, Bendall et al., 2016], which implies that Charles Darwin's observation in his historical work "On the Origin of Species" continues to summarize our struggle in life sciences to describe fundamental units of life even though microbiology has gone beyond species in this pursuit: "no one definition of species has yet satisfied all naturalists; yet every naturalist knows vaguely what [they mean] when [they speak] of a species" [Darwin, 1859]. Our study is not well-positioned to offer a precise theoretical definition, either. Instead, similar to previous studies, we resort to an operational definition that suggests a population is "an agglomerate of naturally occurring microbial cells, genomes of which are similar enough to align to the same genomic reference with high sequence identity" [Delmont and Eren, 2018] (see also [Denef, 2019] and references therein for a comprehensive discussion of what constitutes a population from a metagenomic perspective). By outsourcing the hypothetical radius of a population in sequence space to the minimum sequence identity of short reads recruited from metagenomes, this approach offers a practical means to study very closely related environmental sequences without invoking theoretical considerations. The broad heterogeneity continuum that possesses no discernible sequence-discrete components we observed within the narrow sequence set defined this way, *i.e.* metagenomic reads that match competitively to conserved HIMB83 genes (Figure 3.2), supports the assumption that this set originates within a population boundary (Figure 3.5). However, due to the incomplete theoretical foundation and limitations associated with the use of short metagenomic reads, in discussions here we more conservatively assume that our reads originate from multiple closely related yet intertwined SAR11 populations within subclade 1a.3.V.

Both high recombination rates between cells displaying low gANI values and frequent transfer of adaptive genes between ecologically distinct clades could explain the high-level of cohesion between SAR11 populations in the surface ocean [Vergin et al., 2007, Cohan, 2019]. The high density of closely related 1a.3.V cells in the surface ocean suggests the strength of these two forces could be high within populations as well. At least two hypotheses reconcile extensive SAR11 sequence diversity and aide in understanding its implications. One hypothesis is that the members of 1a.3.V we access are in the process of evolving into multiple sequence-discrete populations and we are simply observing an emerging fork in the evolutionary journey of SAR11. Alternatively, the observed diversity may represent a cloud of random sequence variants akin to a quasispecies [Domingo Esteban et al., 2012]. To examine these hypotheses, we tested the correlation between basic statistical properties of these curves (*i.e.*, mean, standard deviation, and skewness) and environmental parameters via linear regression (Figure 3.7, Table 3.6). This analysis revealed a significant correlation between *in situ* temperature and distribution shape (mean p-value: 2.0×10^{-3}; standard deviation p-value: 3.4×10^{-8}; skewness p-value: 1.0×10^{-8}), which suggests a strong influence of temperature and its co-variables on the sequence heterogeneity within 1a.3.V (Figure 3.2) and is incompatible with the hypothesis of random sequence variants.

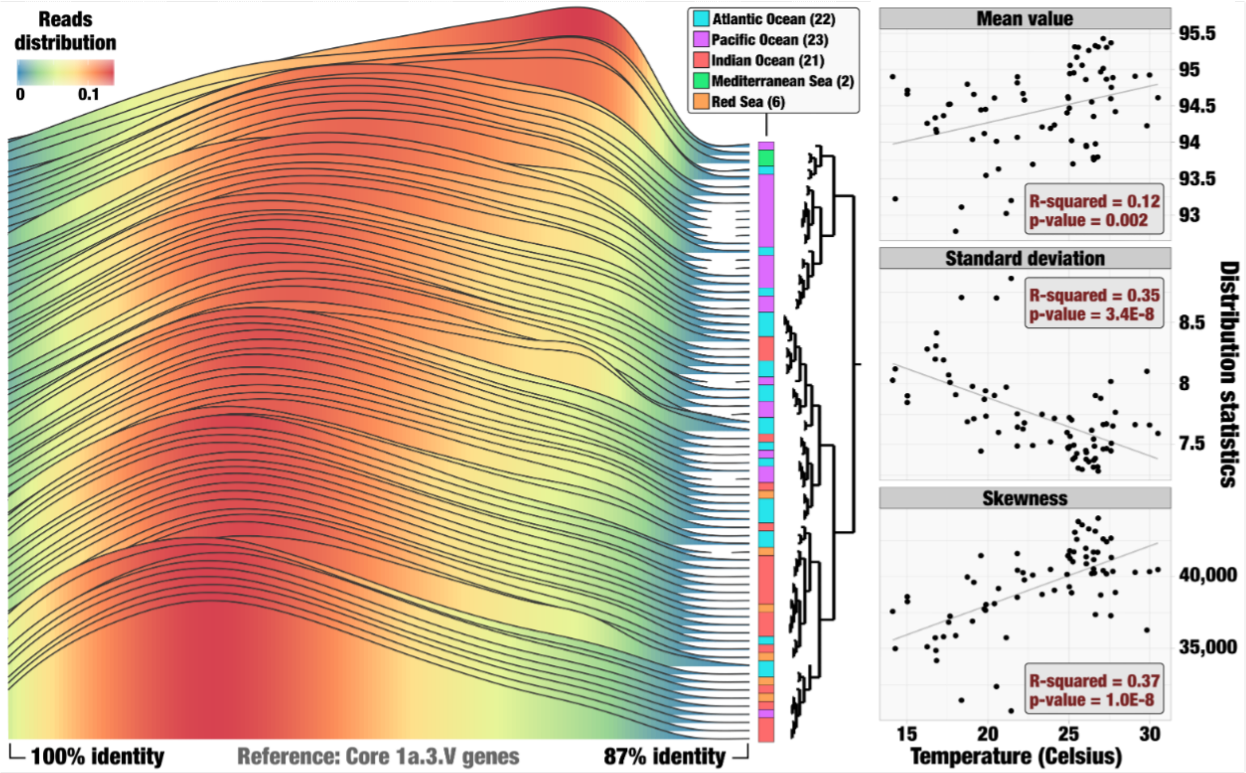


Figure 3.2: Statistics of recruited reads. Left panel shows percent identity distributions in each of the 74 metagenomes. Curves are colored based on height. Metagenomes are ordered according to how the percent identity distributions hierarchically cluster based on Euclidean distance (dendrogram). Right panels display a summary of distribution statistics for each percent identity distribution compared against *in situ* temperature in a linear regression (correlations to all other available parameters are summarized in Figure 3.7). Each point is a metagenome and black lines are lines of best fit. For visual clarity, the data in left panel considers only the median read length and interpolates between data points, whereas the data in right panels considers all read lengths with no interpolation.

3.4.4 SAAVs: Accurate characterization of non-synonymous variation

Percent identity distributions are useful to assess overall alignment statistics of short reads to a reference; however, they do not convey information regarding allele frequencies, their functional significance, or association with biogeography. To bridge this gap, we implemented a framework to characterize amino acid substitutions in metagenomic data and to study genomic variation that impacts amino acid sequences (see Material and Methods). Briefly, our approach employs only metagenomic short reads that cover all three nucleotides in a given codon to determine the frequency of single-amino acid variants (SAAVs) in translated protein sequences. While synonymity is a codon characteristic, in practice it is often determined from a single-nucleotide variant (SNV) with the assumption that the two remaining nucleotides are invariant. However, populations with extensive nucleotide variation can violate this assumption. Indeed, in the case of the core 1a.3.V genes, on average 22.5% of SNVs per metagenome co-occurred with other SNVs in the same codon. Thus, quantifying frequencies of full codon sequences as implemented in the SAAV workflow is a requirement to correctly assess synonymity.

Among the 799 core 1a.3.V genes and 74 metagenomes, we identified 1,074,096 SAAVs in which >10% of amino acids diverged from the consensus (*i.e.*, the most frequent amino acid for a given codon position and metagenome) (Table 3.2). The SAAV density (the percentage of codon positions that harbor a SAAV) of core 1a.3.V genes averaged 5.76% and correlated with SNV density (19.3% on average) across the 74 metagenomes (linear regression, p-value $< 2.2\times10^{16}$; R^2: 0.90; Figure 3.5 and Table S02). SNV and SAAV density metrics did not decrease in metagenomes sampled closest to the source of isolation, suggesting that the location of isolation for strain HIMB83 does not predict the biogeography and population genetics of 1a.3.V. To improve downstream beta-diversity analyses, we discarded codon positions if their coverage in any of the 74 metagenomes was <20X, which resulted in a final collection of 738,324 SAAVs occurred in 37,416 codon positions that harbored a SAAV in at least one metagenome among the total of 252,333 codon positions (14.8%) within the core 1a.3.V genes (Tables 3.7 and 3.8). We considered a protein to be 'invariant' (*i.e.*, absence of variation due to intensive purifying or positive selection) in a given metagenome if it lacked SAAVs. They were rare in our data: in total, we detected 2,548 invariant proteins (only 4.3% of all possibilities across the 74 metagenomes) that encompassed only 113 genes. In addition, all genes, except one 679 nucleotide long ABC transporter (gene id 1469), contained at least one SAAV in at least one metagenome, revealing a wide range of amino acid sequence diversification among core 1a.3.V proteins (Table 3.7).

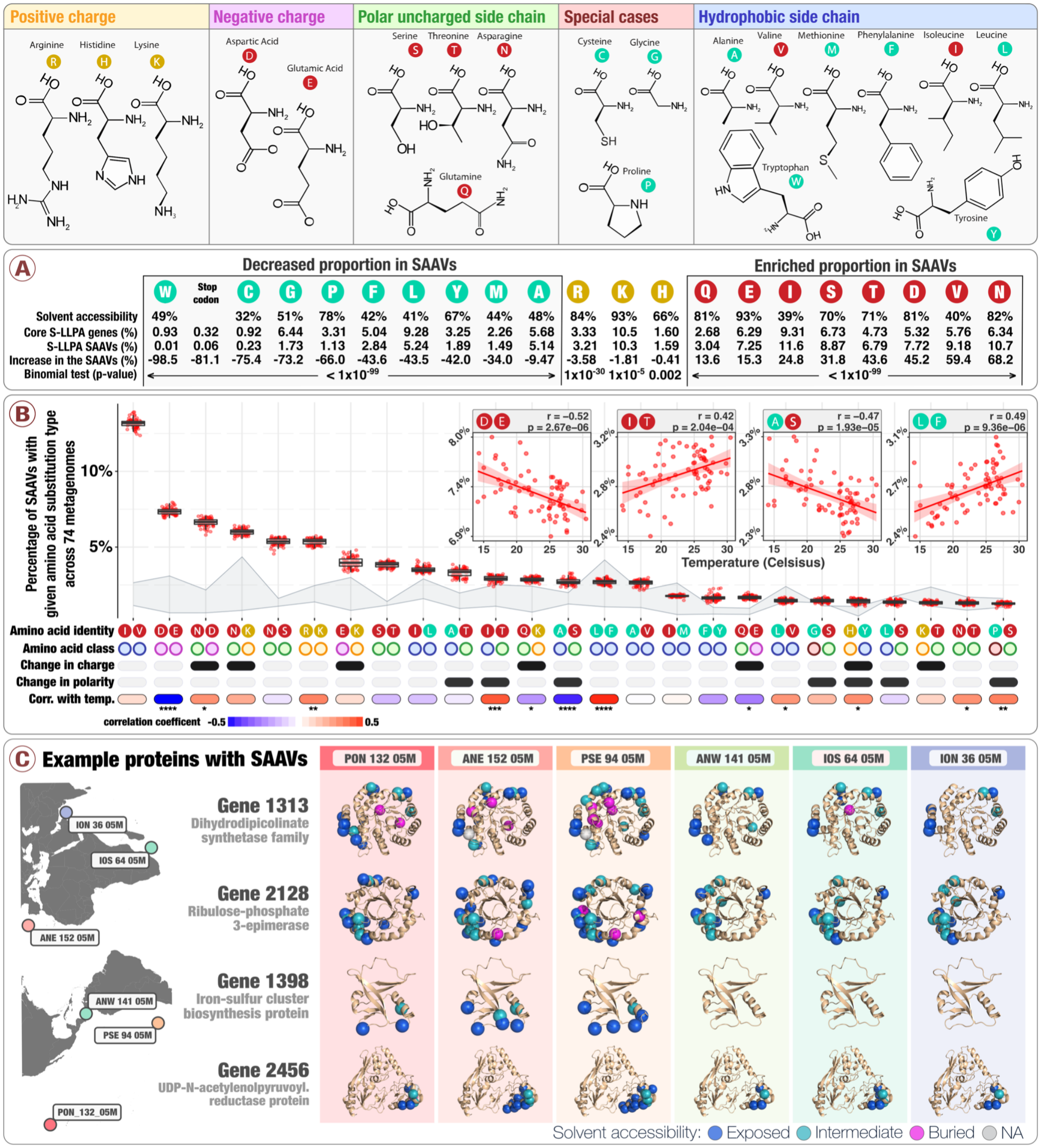
3.4.5 Hydrophobicity influences the strength of purifying selection acting on amino acids

To understand how commonly each amino acid was found in variant sites, we compared the amino acid composition of SAAVs to the amino acid composition of the core 1a.3.V genes (see Material and Methods). In a scenario in which amino acids are as common in SAAVs as they are across all 799 core genes, the frequency that an amino acid occurred in SAAVs (variant sites) would share one-to-one correspondence with its frequency within the core genes (all sites). While these variables were correlated (linear regression, p-value: 9.8×10^{-6}; R^2: 0.65), we observed large deviations from this null expectation, implying strong differential occurrence of amino acids in SAAVs relative to their occurrence in core genes (Figure 3.3A, Figure 3.8, Table 3.9). All negatively charged (Asp, Glu) and uncharged polar (Thr, Asn, Ser, Gln) amino acids were significantly enriched in SAAVs compared to the core 1a.3.V genes (Figure 3.3A). For instance, while asparagine made up only 6.34% of all amino acids in the core genes, on average 10.7% (±0.16%) of SAAVs involved asparagine substitutions across the 74 metagenomes (Table S9). Interestingly, unlike negatively charged amino acids, positively charged amino acids did not exhibit substantial differences (<4% deviation between core 1a.3.V genes and SAAVs). Thus, hydrophilic amino acids were either overrepresented or exhibited little change in SAAVs with respect to their frequency within core genes. In stark contrast, all hydrophobic amino acids, with the very notable exceptions of isoleucine and valine, were underrepresented in SAAVs (Figure 3.3A, Figure 3.8, Table S9).

Hydrophobic interactions within the solvent inaccessible core of proteins are known to be critical for maintaining the stability required for folding and activity, which enforces a strong purifying selection placed on mutations occurring in buried (solvent inaccessible) positions [Bustamante et al., 2000, Chen and Zhou, 2005, Worth et al., 2009]. Since hydrophobic amino acids form the majority of buried positions, they are on average under stronger purifying selection, which is the likely explanation for the underrepresentation of hydrophobic amino acids within SAAVs. On the other hand, mutations in exposed (solvent accessible) positions on the surface of proteins are tolerated more, as they are less likely to disrupt protein architecture. Overall, our compositional analysis revealed that the occurrence of amino acids in SAAVs is roughly correlated with the occurrence of amino acids within the core 1a.3.V genes, and that deviations from this expectation are driven in part by levels of purifying selection that depend upon the suitability of an amino acid's hydrophobicity for a given physicochemical environment (Figure 3.8).

3.4.6 Amino acid exchange rates reveal hallmarks of neutral, purifying, and adaptive evolution

Next, we sought to investigate amino acids that co-occur in variable sites. SAAVs were often dominated by a few amino acids; hence, the frequency vector for a given SAAV contained many zero values. To reduce sparsity, we first simplified our data by associating each SAAV with an amino acid substitution type (AAST), defined as the two most frequent amino acids in a given SAAV. In 738,324 SAAVs, we observed 182 of 210 theoretically possible unique AASTs and a highly skewed AAST frequency distribution (Table S9, Figure 3.3B boxplots). For example, the two most frequent AASTs, 'isoleucine/valine' and 'aspartic/glutamic acid', together comprised 20% of all SAAVs (Figure 3.3B). This is not surprising, since the amino acids in both of these AASTs (1) are common in the genome, (2) share very similar chemical structure (both differing by only a single methylene bridge), and (3) can be substituted through a single nucleotide substitution. On the other hand, the 'glycine/tryptophan' pair



	W	Stop codon	C	G	P	F	L	Y	M	A	R	K	H	Q	E	I	S	T	D	V	N
Solvent accessibility	49%		32%	51%	78%	42%	41%	67%	44%	48%	84%	93%	66%	81%	93%	39%	70%	71%	81%	40%	82%
Core S-LLPA genes (%)	0.93	0.32	0.92	6.44	3.31	5.04	9.28	3.25	2.26	5.68	3.33	10.5	1.60	2.68	6.29	9.31	6.73	4.73	5.32	5.76	6.34
S-LLPA SAAVs (%)	0.01	0.06	0.23	1.73	1.13	2.84	5.24	1.89	1.49	5.14	3.21	10.3	1.59	3.04	7.25	11.6	8.87	6.79	7.72	9.18	10.7
Increase in the SAAVs (%)	-98.5	-81.1	-75.4	-73.2	-66.0	-43.6	-43.5	-42.0	-34.0	-9.47	-3.58	-1.81	-0.41	13.6	15.3	24.8	31.8	43.6	45.2	59.4	68.2
Binomial test (p-value)	← < 1x10^{-99} →										1x10^{-30}	1x10^{-5}	0.002	← < 1x10^{-99} →							

Figure 3.3: Physico-chemical properties of amino acid variants. The top panel describes the structure of 20 amino acids grouped by their main chemical properties. Panel A describes the solvent accessibility of amino acids, their relative distribution in both the core 1a.3.V genes and SAAVs, and their percentage increase in SAAVs as compared to the core 1a.3.V genes. The solvent accessibility of amino acids derives from the analysis of 55 proteins

Figure 3.3 continued: [Bordo and Argos, 1991]. Panel B describes the relative abundance of the top 25 most prevalent amino acid substitution types (AASTs) across 74 metagenomes (boxplots), along with the classes their amino acids belong to and the correlation coefficient between AAST prevalence and *in situ* temperature calculated via linear regression (see Figure 3.9 for p-values). The area shaded in light gray shows bounds for the expected frequency distribution given strictly neutral processes. The upper bound is Model 1 and the lower bound is Model 2 (see Materials and Methods). The 4 insets example the relationship between AAST prevalence and *in situ* temperature for the AASTs 'aspartic/glutamic acid', 'isoleucine/threonine', 'alanine/serine', and 'leucine/phenylalanine' (Figure 3.9 illustrate similar plots for all 25 of the most prevalent AASTs). The 25 AASTs included in the analysis cover 87.1% of all SAAVs. Panel C displays SAAVs on the predicted protein structures of four core 1a.3.V genes across six metagenomes from distant locations.

represents an opposite example: these amino acids (1) are uncommon in the genome, (2) share no chemical or structural similarity to one another, and (3) can only be substituted through a triple nucleotide substitution. Expectedly, ‘glycine/tryptophan’ was exceedingly rare in our data and occurred only once in 738,324 SAAVs (Table S9).

While such a skewed AAST frequency distribution cannot be explained by strictly random mutational process (Figure 3.3B light-gray shaded area), it is compatible with standard theories of neutral or nearly-neutral evolution, since such theories consider the role of purifying selection [Ohta and Gillespie, 1996]. Within subclade 1a.3.V the distribution of AAST frequencies was notably constrained across geographies (Figure 3.3B). For example, the relative standard deviation of ‘aspartic/glutamic acid’ frequencies across the 74 metagenomes was just 3.0%, and the statistical spread of other AASTs was comparable (Figure 3.3B). The overall consistency of AAST frequency distributions across geographies supports the hypothesis that purifying selection controls the permissibility of amino acid exchangeability within 1a.3.V and enables an interpretation of these data through a neutral model: SAAVs composing the AAST frequency distribution represent primarily neutral mutations that have drifted to measurable levels, and the lack of SAAVs in AASTs of dissimilar amino acids that likely represent deleterious mutations reflect the influence of purifying selection. However, a closer inspection reveals a subtle divergence of amino acid exchangeabilities that correlates with water temperature and/or its co-variables (Figure 3.3B insets, Figure 3.9). Note that this divergence is AAST specific; for example, positions with mixed proportions of glutamic and aspartic acid are less commonly found in warm waters (linear regression, uncorrected p-value: 2.7×10^{-6}), yet for isoleucine and valine such a correlation is nonexistent (linear regression, uncorrected p-value: 0.418). These findings suggest that amidst a signal that is predominantly indicative of purifying selection, there appears to be a fingerprint of adaptive/divergent processes caused by temperature and/or its co-variables that subtly shift the mutational profile within 1a.3.V. We were unable to attribute the magnitude or direction of these correlations to differences between amino acids (*i.e.*, changes in hydrophobicity, size, or charge). This was likely due to the insufficiency of characterizing SAAVs with only the chemical properties of the involved amino acids, and disregarding position-specific information, such as the surrounding physicochemical environment that can only be studied with knowledge of the protein's structure.

To address this shortcoming, we next sought to link SAAVs to predicted protein structures of the core 1a.3.V genes, 436 of which had significant matches in Protein Data Bank for template-based structure modeling (see Materials and Methods). Placing SAAVs on predicted protein structures revealed that their occurrence was not randomly distributed but was instead strongly dependent on the local physicochemical environment of the structure (Figure 3.3C, Table 3.10 and `http://data.merenlab.org/sar11-saavs`). Within the subset of the 1a.3.V proteome accessible to us, we found that buried amino acids (0-10% relative solvent accessibility) were approximately 4.4 times less likely to be variant than those that were exposed (41-100% relative solvent accessibility) (ANOVA, p-value: $< 2 \times 10^{-16}$). This observation was strikingly apparent in TIM barrels, where SAAVs mostly occurred in the outer alpha helix and loop regions (*e.g.*, Figure 3.3C gene 2,128). This trend directly confirmed our previous inference (based on the underrepresentation of hydrophobic amino acids) that solvent inaccessible positions are subject to higher levels of purifying selection and thus contain fewer SAAVs. The local physicochemical environment therefore shapes variation, and visual inspection of Figure 3.3C indicates that this is conserved across distant geographies; *i.e.* positions that vary in one metagenome are likely to vary in others, as well. Overall, 91.7% of variant positions in the core 1a.3.V genes varied in 10 or more metagenomes, and 21.7% varied in all 74 metagenomes (Table 3.8).

3.4.7 Temperature correlates with amino acid allele frequency trajectories

In addition to considering patterns of variability that emerged when we pooled data across 37,416 codon positions exhibiting variation within the core 1a.3.V genes, we also investigated the allele frequency trajectories of individual positions (i.e., the relative frequency between the two most prevalent amino acids across the 74 metagenomes) and sought to identify those that correlate with *in situ* temperature and/or its co-variables. Amino acid allele frequencies in 4,592 of the 37,416 positions were correlated with temperature (Table 3.11; Benjamini–Hochberg multiple testing correction on linear regression p-values, false discovery rate 5%). Figure 3.10 illustrates example cases and correlation statistics per AAST. It is statistically implausible that such correlations with temperature could have arisen from neutral evolution, given that distant oceans share similar temperatures (Table 3.1). It is therefore most plausible to conclude that these allele frequency trajectories are the result of environmentally mediated selection. Although we note that, considering the pervasive effect of genetic hitchhiking in microbial evolution [Good et al., 2017], variation in a considerable fraction of positions may be neutral despite their association with temperature.

We then sought to investigate which positions are under selection, and whether the variation at these positions can be explained by differing levels of purifying selection, or diversifying selection that could be evidence of adaptive evolution. Scrutinizing all 4,592 positions to address these critical questions is an intractable problem, so we narrowed our focus to genes possessing disproportionately high ratios of temperature-correlated to temperature-uncorrelated SAAV positions, since we expected this to be a reasonable criterion for identifying likely candidates of adaptive evolution (Table 3.11). Of the 10 genes fitting this criterion (see Materials and Methods), the permease subunit of a glycine betaine ATP-binding cassette (ABC) transporter stood out due to its appreciated relevance to SAR11 biology: glycine betaine transporters of SAR11 are highly translated proteins in the environment and transport osmolyte compounds into cells for energy production [Noell and Giovannoni, 2019]. To investigate the positioning of amino acids in the tertiary structure of the permease relative to the cellular membrane, we first categorized the location of each residue as transmembrane, cytosolic (inside the inner membrane), or periplasmic (outside the inner membrane) (Figure 3.11). Positions that were not correlated with temperature were commonly transmembrane, and infrequently periplasmic. In contrast, most positions that correlated with temperature were periplasmic (Figure 3.11). The probability of observing a similar distribution between temperature-correlated and temperature-uncorrelated positions across transmembrane, periplasmic, and cytosolic regions was only 0.034 (analytic trinomial test, temperature-uncorrelated distribution as prior), which indicates temperature-correlated positions are subjected to unique evolutionary forces. A previous study suggested that periplasmic residues of transmembrane proteins undergo higher rates of adaptive evolution due to their increased exposure to changing environmental conditions [Sojo et al., 2016]. This observation lends additional support to the hypothesis that periplasmic SAAV positions within this gene that correlate with temperature are more likely shaped by adaptive processes.

Allele frequency trajectories also provide an opportunity to study the directionality of exchange rates of AASTs. For example, of the 1,066 positions dominated by 'alanine/serine' SAAVs, 158 positions correlated with temperature (Figure 3.10). If there was no temperature-driven preference for either amino acid in this subset of positions, the frequency of alanine should positively correlate with temperature as often as the frequency of serine does. Yet this expectation is grossly violated: in 103 of 158 positions alanine frequencies positively correlated with temperature (binomial test, Bonferroni-corrected p-value: 0.004). Overall, this result indicates temperature-dependent amino acid substitution preferences that are independent of site (Figure 3.10B).

3.4.8 SAAV partitioning between warm and cold currents

We finally sought to extend the concept of allele frequency tracking at individual SAAV positions to investigate large-scale geographic partitioning of metagenomes. For this, we simplified the 738,324 SAAVs into a presence-absence matrix for codon position-specific AASTs across 74 metagenomes (Table 3.8, also see 'Recovering codon position-specific AASTs from SAAVs' in Material and Methods). Of 57,277 codon position-specific AASTs affiliated with 37,415 unique codon positions, we detected 1.94% in all 74 metagenomes, while 33.3% were found in single metagenomes (Table 3.8). To estimate distances between metagenomes based on these data, we used a Deep Learning approach. Briefly, this approach relies on a graph-based activity regularization technique for competitive learning from hyper-dimensional data, modified to reveal latent groups of variants in a fully unsupervised manner through frequent random sampling of variants [Kilinc and Uysal, 2017]. Hierarchical clustering of samples based on Deep Learning-estimated distances (Table 3.12) resulted in two main groups: the Western (warm) and Eastern (cold) boundary currents (Figure 3.4A). High latitude, relatively cold, and relatively nutrient rich waters are the source of Eastern boundary currents, which warm up and typically decline in nutrients as they transit in an equatorial direction. The opposite is true of Western boundary currents, which move poleward. The first group of 41 metagenomes, which matched cold currents (Benguela, Canary, California and Peru), encompassed most metagenomes from the Eastern Pacific Ocean, as well as the East side of the Atlantic Ocean (except near the southern tip of Africa) and the Mediterranean Sea (Figure 3.4B). The second group of 33 metagenomes, which matched warm currents (Agulhas, Somali, Mozambique, Brazil and Gulf stream), encompassed all metagenomes from the Red Sea and Indian Ocean, as well as metagenomes from the West side of the Atlantic Ocean (Figure 3.4B). Samples collected from the deep chlorophyll maximum layer of the water column mirrored trends observed in the surface samples (Figure 3.12). The association between SAAVs and ocean current type revealed a strong, global signal at the amino acid-level for 1a.3.V and suggested the presence of two main ecological niches for this lineage. Warm and cold currents are dynamic environments that differ in a host of factors in addition to the latitude and temperature of source waters. Factors that could drive adaptive changes in amino acid sequences between warm and cold currents include major differences in phytoplankton communities, altered composition of dissolved organic carbon pools, and the water temperature itself. Interestingly, the niche defined by cold currents exhibited significantly more SAAVs (ANOVA, p-value: 1.66×10^{-12}). This observation could be explained either by (1) extinction/re-emergence events that operate continually on specific codon positions (adaptive evolution), or (2) changes in abundances within a large seed bank of variants due to positive and negative selection as the lineage transits. A recent study using Lagrangian particle tracking and network theory suggested that all regions of the surface ocean are connected to each other with less than a decade of transit [Jönsson and Watson, 2016], which might favor the latter scenario due to lack of time for the extinction and reemergence of variants in abundant marine microbial lineages.

To explore more detailed trends of the relationships between metagenomes, we further divided our dendrogram into six sub-clusters based on the elbow of the intra-cluster sum-of-squares curve of k-means clusters (Figure 3.13). These 1a.3.V 'proteotypes' grouped samples with similar amino acid variations (Figure 3.4A) and could not have been predicted from the clustering of samples based on percent identity distributions of short reads alone (Figure 3.14). Among the environmental measurements for each metagenome latitude and temperature at the time of sampling were the most significant predictors of the proteotypes

(ANOVA, p-values: 8.56×10^{-13} and 3.57×10^{-7}, respectively). These two variables were followed by the concentrations of nitrate, phosphate, oxygen, and to a lesser extent, silicate and latitude (Table 3.13). The number of SAAVs and the number of invariant proteins, however, were more significant predictors of these groups compared to all environmental parameters (ANOVA, p-value: $< 2 \times 10^{-16}$, Table 3.13). Strikingly, most 1a.3.V proteotypes linked samples from distant geographical regions (Figure 3.4B). An exception to this was the proteotype A, which only contained Pacific Ocean metagenomes (Figure 3.4B). For instance, proteotypes E and F occurred both in the Indian Ocean and the West side of the Atlantic Ocean and associated with distinct warm currents: E was characteristic of the Mozambique and Brazil currents while F dominated the Agulhas current (Figure 3.4B). One of the most interesting proteotypes, D, whose reads most closely resembled the HIMB83 genome itself (Figure 3.4A), contained a distinctively low number of SAAVs, and grouped metagenomes sampled from both sides of the Panama Canal with metagenomes from the Red Sea and North of the Indian Ocean (Figure 3.4B). We also clustered the same data set using fixation index, a widely-used metric to measure population structure [Weir, 2012], which we modified in accordance with [Schloissnig et al., 2013] to permit multi-allelic variant positions. Both approaches preserved associations between distant geographies (*i.e.*, Proteotype D, Figure 3.4, Figure 3.15). , however, they were not identical in their organization of metagenomes (*i.e.*, Proteotype E was associated with colder currents according to fixation index rather than warmer ones; Figure 3.4, Figure 3.15), highlighting the non-trivial nature of establishing individual proteotypes from SAAVs.That said, the significance of *in situ* temperature to explain clustering of metagenomes into two main groups and six proteotypes was higher with Deep Learning (Figure 3.4, Figure 3.15), suggesting that Deep Learning was able to better capture the strong association between temperature and the genomic heterogeneity within 1a.3.V through SAAVs.

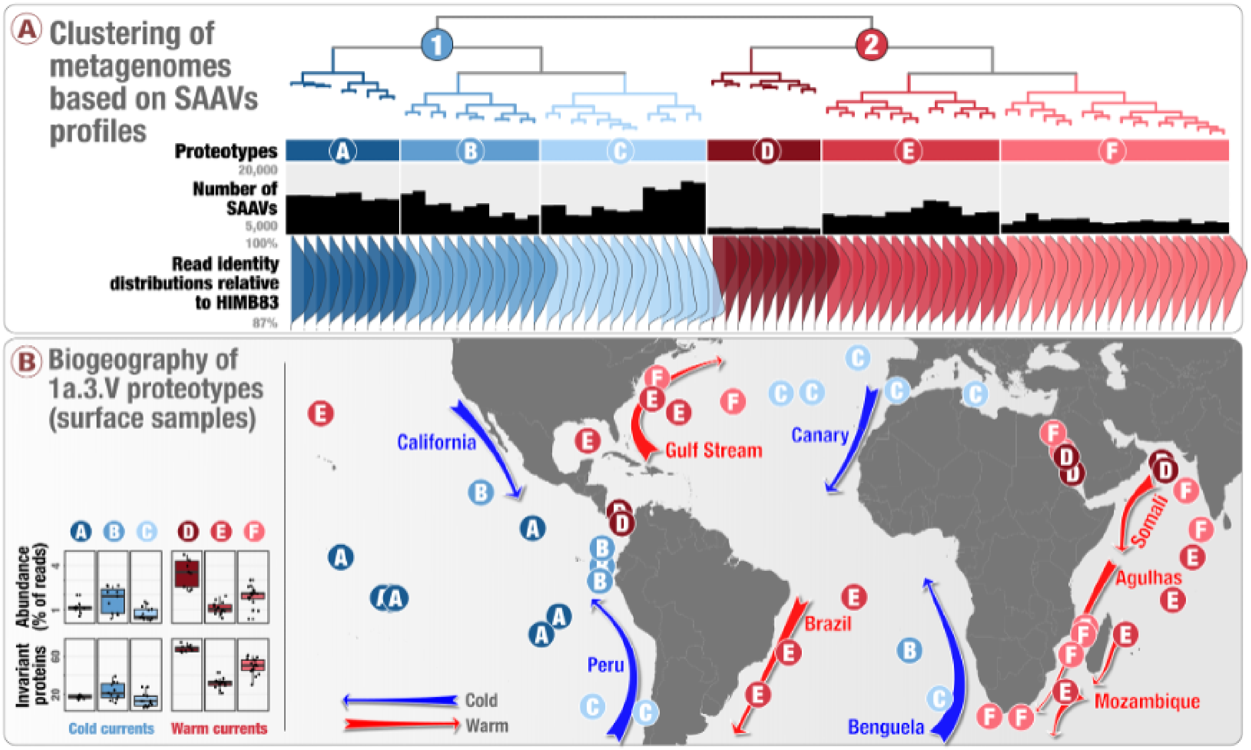


Figure 3.4: Biogeography of SAR11 subclade 1a.3.V based on single amino acid variants. Panel A describes the organization of 74 metagenomes based on 57,277 codon position-specific AASTs affiliated with 37,415 unique codon positions and summarizes the number of detected SAAVs and percent identity of reads HIMB83 recruited for each metagenome. The world map in panel B displays the geographic partitioning of the two main metagenomic groups and six proteotypes. Panel B also describes the relative abundance of 1a.3.V and the number of invariant proteins across the six proteotypes.

The striking connection between geographically distant regions of the oceans through SAAVs suggests a likely role for adaptive processes to maintain the genomic heterogeneity of closely related SAR11 populations within 1a.3.V (Figure 3.16). In fact, both the main ecological niches and more refined proteotypes indicate that SAAVs are not primarily structured by the global dispersal of water masses but instead tend to link distant geographic regions with similar environmental conditions (Figure 3.4B). Overall, these results indicate that environmentally-mediated selection is a strong determinant of SAR11 evolution and biogeography.

One question remains: what is the proportion of distinct evolutionary processes acting upon closely related SAR11 populations within 1a.3.V? Offering a precise answer to this critical question is compounded by multiple theoretical and technical factors. These factors include, but are not limited to, (1) the phenomenon of genetic hitchhiking that prevents accurate determination of amino acid positions that likely confer fitness, (2) the metagenomic short-read recruitment strategy that prevents absolute confidence regarding the origin of each fragment, (3) heavy reliance on temperature as the sole environmental stressor to predict associations between environmental parameters and variation due to limited insights into *in situ* physicochemistry, (4) the lack of a complete understanding of syntrophic relationships between taxa in the environment, and (5) computational bottlenecks to gain rapid and accurate insights into the role of variable amino acid residues even when protein structures are available. With these significant limitations in mind, we could nevertheless speculate that among the 252,333 total codon positions, 37,416 were variable, suggesting purifying selection maintains the conservancy of 85% of the positions within 799 core 1a.3.V genes. Of those 37,416 positions that were within the scope of permissible mutations, 4592 had amino acid frequency trajectories that significantly correlated with temperature, suggesting an upper-bound of 12% for the variable positions that are likely under the influence of temperature-driven adaptive processes, while neutral processes explain at least 88% of the variation. In summary, this global view of the data suggests that among the remarkable amount of variation within some of the most abundant and prevalent microbial populations in the ocean, adaptive evolutionary processes operating on core genes are responsible for variation in about 2% of all codon positions.

3.5 Conclusion

We took advantage of billions of metagenomic reads to investigate single-amino acid variants (SAAVs) within the environmental core genes of the remarkably abundant and closely related SAR11 populations within subclade Ia.3.V, which we defined from a SAR11 metapangenome. The results elicit a highly-resolved quantitative description of purifying selection constraining the scope of permissible mutations to those non-detrimental to protein stability requirements. Of permissible variation, thousands of codon positions harbored allele frequencies that systematically correlated with *in situ* temperatures and, overall, patterns of amino acid diversity reflected the temperature trends of large-scale ocean currents. This was especially apparent regarding the clear SAAV partitioning between Western and Eastern boundary currents. Previous studies have subdivided SAR11 clade Ia into cold-water (Ia.1) and warm-water (Ia.3) subclades with distinct latitudinal distributions [Brown et al., 2012], and reported sinusoidal oscillations between their abundances as a function of seawater temperature at a single temperate ocean site [Murat Eren et al., 2015]. At a much finer evolutionary scale (*i.e.*, closely related populations within Ia.3.V), we observed significantly more protein variants in cold currents and more invariant proteins in warm currents, revealing a global pattern of alternating diversity for SAR11 in surface ocean currents in temperate and tropical latitudes. We were able to track this variation to changes in amino acid sequence preserved by selection.

Trends that emerged from our culture-independent survey of SAR11 were consistent with a recent study that also suggested an important role for environmental and ecological selective processes defining the spatial and temporal distribution of a widespread diatom species [Whittaker and Rynearson, 2017]. Overall, these findings suggest that environmentally-mediated selection plays a critical role in the journey of cosmopolitan microbial populations in the surface ocean, lending credence to the idea for marine systems that "everything is everywhere but the environment selects" [Baas Becking, 1934]. However, identifying environmental variables and their contributions to diversification within a lineage is shrouded by both the dynamism and complexity of surface ocean environments, as well as the rich evolutionary dynamics that arise even in the simplest of conceivable environments [Good et al., 2017]. These formidable challenges stress the importance of designing appropriate experiments to uncover variables that underpin the evolutionary divergence of closely related lineages, and drive transitions between them through space and time.

3.6 Materials and Methods

The URL `http://merenlab.org/data/sar11-saavs` contains a reproducible bioinformatics workflow that extends the descriptions and parameters of programs used here for (1) the metapangenome of SAR11 using cultivar genomes, (2) the profiling of metagenomic reads that the cultivar genomes recruited, (3) the analysis of single nucleotide variants using Deep Learning, and (4) the visualization of single nucleotide variants in the context of protein structures.

3.6.1 SAR11 cultivar genomes

We acquired the genomic content of 21 SAR11 isolates from NCBI and simplified the deflines using anvi'o [Murat Eren et al., 2015]. We then concatenated all contigs into a single FASTA file, and generated an anvi'o contigs database, during which Prodigal [Hyatt et al., 2010] v2.6.3 identified open reading frames in contigs, and we annotated them with InterProScan [Zdobnov and Apweiler, 2001] v1.17. Table 3.1 reports the main genomic features.

3.6.2 Metagenomic datasets

We acquired 103 metagenomes from the European Bioinformatics Institute (EBI) repository under the project IDs ERP001736 (n=93; TARA Oceans project) and ERP009703 (n=10; Ocean Sampling Day project), and removed noisy reads with the illumina-utils library (Eren et al., 2013b) v1.4.1 (available from `https://github.com/meren/illumina-utils` using the program 'iu-filter-quality-minoche' with default parameters, which implements the method previously described by Minoche et al. [Minoche et al., 2011]. Table 3.1 reports accession numbers and additional information (including the number of reads and environmental metadata) for each metagenome.

3.6.3 Pangenomic analysis

We used the anvi'o pangenomic workflow [Delmont and Eren, 2018] to organize translated gene sequences from SAR11 genomes into gene clusters. Briefly, anvi'o uses BLAST [Altschul et al., 1990] to assess the similarity between each pair of amino acid sequences among all genomes, and then resolves this graph into gene clusters using the Markov Cluster algorithm [Enright et al., 2002]. We built the gene clustering metric using a minimum percent identity of 30%, an inflation value of 2, and a maxbit score of 0.5 for high sensitivity. Anvi'o used the occurrence of gene clusters across genomes data, which is also reported in Table 3.3, to compute clustering dendrograms both for SAR11 genomes and gene clusters using Euclidian distance and Ward linkage algorithm.

3.6.4 Estimating distances between isolate genomes based on full-length 16S ribosomal RNA gene sequences

We used the program `anvi-get-sequences-for-hmm-hits` (with parameters `--hmm-source Ribosomal_RNAs` and `--gene-name Bacterial_16S_rRNA`) to recover full-length 16S ribosomal RNA gene sequences from the anvi'o contigs database for the 21 isolate genomes. We then used PyANI [Pritchard et al., 2015] through the program `anvi-compute-ani` to estimate pairwise distances between each sequence.

3.6.5 Competitive recruitment and profiling of metagenomic reads

We mapped reads competitively from each metagenome against a single FASTA file containing all SAR11 genomes using Bowtie2 [Langmead and Salzberg, 2012] v.2.0.5 with default parameters, and converted the resulting SAM files into BAM files using samtools [Li et al., 2009] v1.3.1. Competitive read recruitment ensures that short reads that match to more than one genome are assigned uniquely and randomly to one of the matching genomes. This minimizes computational biases at the mapping level and avoid inflated coverage statistics. To confirm our observations, we also used BWA [Li and Durbin, 2009] to recruit reads (with the option n=0.05). We used anvi'o to generate profile databases from the BAM files and combine these mapping profiles into a merged profile database, which stored coverage and variability statistics as outlined in Eren et al. [Murat Eren et al., 2015]. Table 3.2 reports the mapping results (number of recruited reads, as well as mean coverage and detection statistics) per genome across the 103 metagenomes.

3.6.6 Determining the coverage of HIMB83 genes across metagenomes

The anvi'o merged profile database contains the coverage of individual genes across metagenomes. We normalized the coverage of HIMB83 genes in each metagenome (summarized in Table 3.5) and calculated their coefficient of gene variation. We used the coefficient of gene variation estimates to identify metagenomes in which HIMB83 was well detected, yet the coverage values of its genes were highly unstable, which is an indicator of non-specific read recruitment from other lineages.

3.6.7 Determining the main ecological niche and core genes of 1a.3.V

We considered metagenomes in which HIMB83 was sufficiently abundant (mean genomic coverage >50X) with a stable detection of its genes (coefficient of gene variation <1.25) to represent the main ecological niche of 1a.3.V. To determine the core 1a.3.V genes, we first disregarded metagenomes that displayed an unusually high coefficient of gene coverage variation (Figure 3.5, panel A), which can indicate non-specific read recruitments from other abundant populations. The 74 metagenomes fitting these criteria are summarized in Table

3.1. We defined the subset of HIMB83 genes as the core 1a.3.V genes if in each of the 74 metagenomes, the mean coverage of a gene remained within a factor of 5 of the mean coverage across all genes. The 799 genes fitting this criterion are summarized in Table 3.4. Calculation of percent identity distributions of recruited metagenomic short reads. We used percent identity distributions to broadly characterize how well short reads within a metagenome matched to the reference sequences by which they were recruited. We determined the percent identity for each read as $100 \times (N - n)/N$ where n is the number of mismatches to the reference and N is the read length. For simplicity, visualization of these distributions only included reads lengths of which matched to the median read length, and we defined bins to contain only one unique value. For example, if the median length of reads was 100, the bin domains for visualization purposes were (99,100], (98,99], (97,98], $\ldots$, [0,1]. In contrast, all statistical calculations were carried out using all read lengths.

3.6.8 Generating single-nucleotide variants (SNV) data

We used the program `anvi-gen-variability-profile` to report variability tables describing the nucleotide frequency (*i.e.*, ratio of the four nucleotides) in recruited metagenomic reads per SNV position. To study the extent of variation of the core 1a.3.V genes across all metagenomes, we instructed anvi'o to report positions with more than 1% variation at the nucleotide level (*i.e.*, at least 1% of recruited reads differ from the consensus nucleotide). To compare the densities of SAAVs to SNVs, we instructed anvi'o to report only positions with more than 10% variation at the nucleotide level. Table 3.2 reports the density of SNVs for all SAR11 genomes across all metagenomes. We also used anvi'o to report SNVs for a subset of genes and metagenomes, and by considering only nucleotide positions with a minimum coverage cut-off across metagenomes under consideration. Controlling the minimum coverage of single nucleotide positions across metagenomes improves confidence in variability analyses. Table 3.5 reports the SNV density values for all core 1a.3.V genes.

3.6.9 Definitions of 'SAAV', 'allele frequency' and 'AAST'

A single amino acid variant (SAAV) is a codon position that exhibits variation in a metagenome, and the unique identifier of a SAAV is a single codon position and a metagenome. The position of a SAAV in the reference sequence, and a vector of 21 elements that contain the allele frequencies of each amino acid as well as the stop codon fully characterize a SAAV. The allele frequency of an amino acid is equal to the number of short reads that fully cover the codon that resolves to the amino acid, divided by the total number of reads that fully cover the same position (the sum of all 21 allele frequencies is therefore 1). We also attributed to each SAAV an amino acid substitution type (AAST), which corresponds to the two amino acids with the largest and second largest allele frequencies.

3.6.10 Generating single-amino acid variants (SAAVs) data

The program `anvi-gen-variability-profile` (with an additional `--engine AA` flag) reported variability tables describing the allele frequencies for each SAAV. Anvi'o only considers short reads that cover the entire codon to determine amino acid frequencies at a given codon position in a metagenome. We instructed anvi'o to report only positions with more than 10% variation at the amino acid-level (*i.e.*, at least 10% of recruited reads differ from the consensus amino acid). Table 3.2 reports the density of SAAVs for all SAR11 genome across all metagenomes. We also used anvi'o to report SAAVs for a subset of genes and metagenomes, and by considering only gene codons with a minimum coverage cut-off of 20X across all metagenomes of interest. Controlling the minimum coverage of gene codons across metagenomes improves confidence in variability analyses. Table 3.5 reports the SAAV density values for all core 1a.3.V genes.

3.6.11 Differential occurrence of amino acids in SAAVs and in the core 1a.3.V genes

We determined the amino acid composition in the 799 core 1a.3.V genes as well as in SAAVs maintained in each metagenome using anvi'o programs `anvi-get-aa-counts` and `anvi-get-codon-frequencies` (with the flag `--return-AA-frequencies-instead`). We quantified the amino acid composition of all core 1a.3.V genes of in HIMB83 using the program `anvi-get-aa-counts`. In contrast, we quantified the amino acid composition of SAAVs by calculating the frequency of a given amino acid being one of the two dominant alleles. We then calculated p-values via a binomial test that represents the probability of observing the difference between amino acid frequencies computed over all core 1a.3.V genes versus only 1a.3.V SAAVs, given the null hypothesis that amino acids in 1a.3.V SAAVs are distributed according to the same distribution as the amino acids in the core 1a.3.V genes.

3.6.12 Estimating a neutral AAST frequency distribution

This calculation provides an estimate for the AAST frequency distribution given strictly neutral mutations. Unlike the neutral theory of evolution, it excludes the influential effects of purifying selection (negative selection coefficients). Since all mutations are equally likely to drift to detectable frequencies under a neutral model, the expected number of variant positions that have C_i and C_j as their two dominant alleles, is proportional to the rate that C_i mutates to C_j plus the rate that C_j mutates to C_i. Expressed mathematically,

$$\mathbb{E}(N_{\{C_i,C_j\}}) \propto P(C_i \mid m)P(C_i \to C_j \mid C_i, m) + P(C_j \mid m)P(C_j \to C_i \mid C_j, m)$$

where $\mathbb{E}(N_{\{C_i,C_j\}})$ is the expected number of variant positions that have C_i and C_j as their two dominant alleles, $P(C_i \mid m)$ is the probability that a C_i position mutates given that a mutation has occurred, and $P(C_i \to C_j \mid C_i, m)$ is the probability that such a mutation will mutate to C_j. Assuming all sites are equally likely to mutate, $P(C_i \mid m)$ is equivalent to the fraction of codons in the reference sequence that are C_i, and we denote this quantity as f_{C_i}. To extend the equation to the expected number of variant positions that have amino acids A_1 and A_2 as their two dominant alleles, *i.e.* a quantity proportional to the AAST frequency, one must enumerate over all codons in A_1 and A_2:

$$\mathbb{E}(N_{AAST=\{A_1,A_2\}}) \propto \sum_{C_i \in A_1} \sum_{C_j \in A_2} P(\{C_i, C_j\})$$

In general, $P(C_i \to C_j \mid C_i, m)$ will depend primarily upon the nucleotide edit distance between C_i and C_j, which we denote as d, as well as the transition/transversion rate ratio, which we will denote κ. How the model handles these aspects will critically influence the expected frequency distribution. To encapsulate the broadest possible interpretation of the neutral model, we evaluate expressions for two extreme cases: In the first case (Model 1), we assume that the probability of an edit distance $d > 1$ is 0 (in reality, estimates at least for eukaryotes range from 0.003 [Ellegren et al., 2003] to 0.03 [Schrider et al., 2011]. We also impose a κ value of 2 so that transitions are twice as likely as transversions. Intuitively, these impositions have the effect of skewing the AAST frequency distribution towards AASTs that possess highly similar codons. In the second case (Model 2), we assume all codon transitions are equally likely regardless of edit distance or the number of transitions/transversions ($\kappa = 1$). Intuitively, this has the effect of homogenizing the AAST frequency distribution towards a more uniform-like distribution.

In Model 1, $P(C_i \to C_j \mid C_i, m) = \frac{1}{3}\delta_{d,1}P(m)$, where $\delta_{d,1}$ is a Kronecker delta function describing the probability the mutation has an edit distance d, $\frac{1}{3}$ is the probability that the correct nucleotide position is mutated, and $P(m)$ is the probability that the mutation occurs based on whether or not it is a transition. Formally,

$$\begin{cases} \kappa/\kappa+2 & m = \text{transition} \\ 1/\kappa+2 & m = \text{transversion} \end{cases}$$

In Model 2, $P(C_i \to C_j \mid C_i, m) = 1/63$, since all 63 possible mutations are permissible and equally probable. The expressions for $\mathbb{E}(N_{AAST=\{A_1,A_2\}})$ for Model 1 and Model 2 thus simplify to:

$$(M_1)\mathbb{E}(N_{AAST=\{A_1,A_2\}}) \propto \langle f_{C_i}, f_{C_j} \rangle \delta_{d,1} P(m)$$
$$^{(M_2)}\mathbb{E}(N_{AAST=\{A_1,A_2\}}) \propto \langle f_{C_i}, f_{C_j} \rangle$$

where M_1 and M_2 refer to Model 1 and Model 2, respectively. To compare directly with observation, we extracted f_{C_i} for the 64 codons from the HIMB83 reference sequence using `anvi-get-codon-frequencies` and the distributions under both models were calculated from the above equations.

3.6.13 3D structure of proteins using template-based protein structure modeling

We used a template-based structure modeling tool, RaptorX Structure Prediction [Källberg et al., 2012], to predict structures of 1a.3.V amino acid sequences based on available data from the Protein Data Bank (PDB) [Bernstein et al., 1977]. We used the program blastp in NCBI's BLAST distribution to identify core 1a.3.V genes that matched to an entry with at least 30% similarity over the length of the given core gene. We then programmatically mapped SAAVs from metagenomes onto the predicted tertiary structures, and used PyMOL [DELANO and W. L, 2002] to visualize these data. We colored SAAVs based on RaptorX-predicted structural properties, including solvent accessibility and secondary structure.

3.6.14 Identifying genes with disproportionately high number of temperature-correlated positions

First, we calculated the number of temperature-correlated and temperature-uncorrelated positions for each of the 1a.3.V core genes. Then, we performed a one-sided binomial test that these numbers are biased towards higher proportion of temperature-correlated positions compared to a model distribution defined from the total number of temperature-correlated positions in 1a.3.V. Since there were 4,592 such positions out of 37,416, the model probability of success was defined as $p_0 = 4592/37416 = 0.123$. In other words, the expected proportion of variant positions in a gene that are temperature-correlated is 0.123 under the model. We corrected the resulting p-values for each gene for multiple testing using Benjamini & Hochberg's method [Benjamini and Hochberg, 1995].

3.6.15 Predicting transmembrane, periplasmic, and cytosolic regions in the glycine betaine permease

To categorize amino acid positions as transmembrane, periplasmic, and cytosolic, we used Phobius [Käll et al., 2004, 2007], a membrane topology and prediction software through the webserver at `http://phobius.sbc.su.se`. The output is a probability of the 4 classes for each residue, and to simplify the data we categorized each residue into the class found to be most probable. We removed residues with signaling peptide association from downstream analyses.

3.6.16 Recovering codon position-specific AASTs from SAAVs

We simplified the hyper-dimensional SAAV data into a simpler presence-absence matrix for downstream analyses. For this, we defined codon position-specific AASTs (cAASTs) and summarized their occurrence across metagenomes. In such a table the value of '1' indicates that a given metagenome had a SAAV at a given codon position that resolved to a given AAST. In contrast, the value '0' indicates that the metagenome did not have a SAAV that resolved to this AAST. In the latter case a given metagenome may have another AAST in this particular codon position (in which case this information would appear in another row in the same table that is affiliated with the same AAST with the same codon position). Hence, each AAST listed in the first column of the table will be unique to a single codon position, yet a given codon position may have different AASTs in different metagenomes, resulting in multiple AASTs in the resulting table that belong to the same codon position. Combining AAST with the codon position would then result in a unique cAAST.

3.6.17 Application of Deep Learning to codon-position-specific AASTs data

To estimate an unbiased distance between our metagenomes based on SAAVs, we used a novel deep neural network modification called the auto-clustering output layer (ACOL). Briefly, ACOL relies on a recently introduced graph-based activity regularization (GAR) technique for competitive learning from hyper-dimensional data to demarcate fine clusters within user-defined 'parent' classes [Kilinc and Uysal, 2017]. In this application of ACOL, however, we modified the algorithm so it can reveal latent groups in our SAAVs in a fully unsupervised manner through frequent random sampling of SAAVs to create pseudo-parent class labels instead of user-defined classes [Kilinc and Uysal, 2018]. See the URL `http://merenlab.org/data/sar11-saavs` for the details of the pseudo parent-class generation algorithm, and the reproducible distance estimation workflow in Python.

3.6.18 Other statistical tests and visualization

We used the `aov` function in R to perform one way ANOVA tests, used the ggplot2 [Ginestet, 2011a] package for R to visualize the relative distribution of 1a.3.V genes and geographic distribution of proteotypes, and finalized all figures using an open-source vector graphics editor, Inkscape (available from http://inkscape.org/).

3.6.19 Code and data availability

The vast majority of analyses relied on the open-source software platform anvi'o v2.4.0 (available from http://merenlab.org/software/anvio). The URL http://merenlab.org/data/sar11-saavs serves the remaining custom code used in our analyses. We made available (1) SAR11 isolate genomes (http://doi.org/10.6084/m9.figshare.5248945), (2) the anvi'o contigs database and merged profile for SAR11 genomes across metagenomes (http://doi.org/10.5281/zenodo.835218) and the static HTML summary for the mapping results (http://doi.org/10.6084/m9.figshare.5248453), (3) the SAR11 metapangenome (http://doi.org/10.6084/m9.figshare.5248459), single-nucleotide and single-amino acid variant reports for 1a.3.V across 74 TARA Oceans metagenomes (http://doi.org/10.6084/m9.figshare.5248447), and (4) SAAVs overlaid on predicted tertiary structures of 58 core 1a.3.V genes (http://doi.org/10.6084/m9.figshare.5248432). The URL http://anvi-server.org/p/4Q2TNo serves an interactive version of the SAR11 metapangenome, and the URL http://data.merenlab.org/sar11-saavs serves an interactive web page to investigate the link between SAAVs and predicted protein structures.

3.7 Acknowledgements

We thank the TARA Oceans consortium for generating metagenomic datasets of great legacy, as well as all researchers involved in the characterization of SAR11 isolates. We thank Kostas Konstantinidis, Edward Delong, Lois Maignien, Mike Lee, and the members of the Meren Lab for helpful discussions. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) to E.K, as well as the Frank R. Lillie Research Innovation Award and the start-up funds from the University of Chicago to A.M.E.

3.8 Supplementary Figures

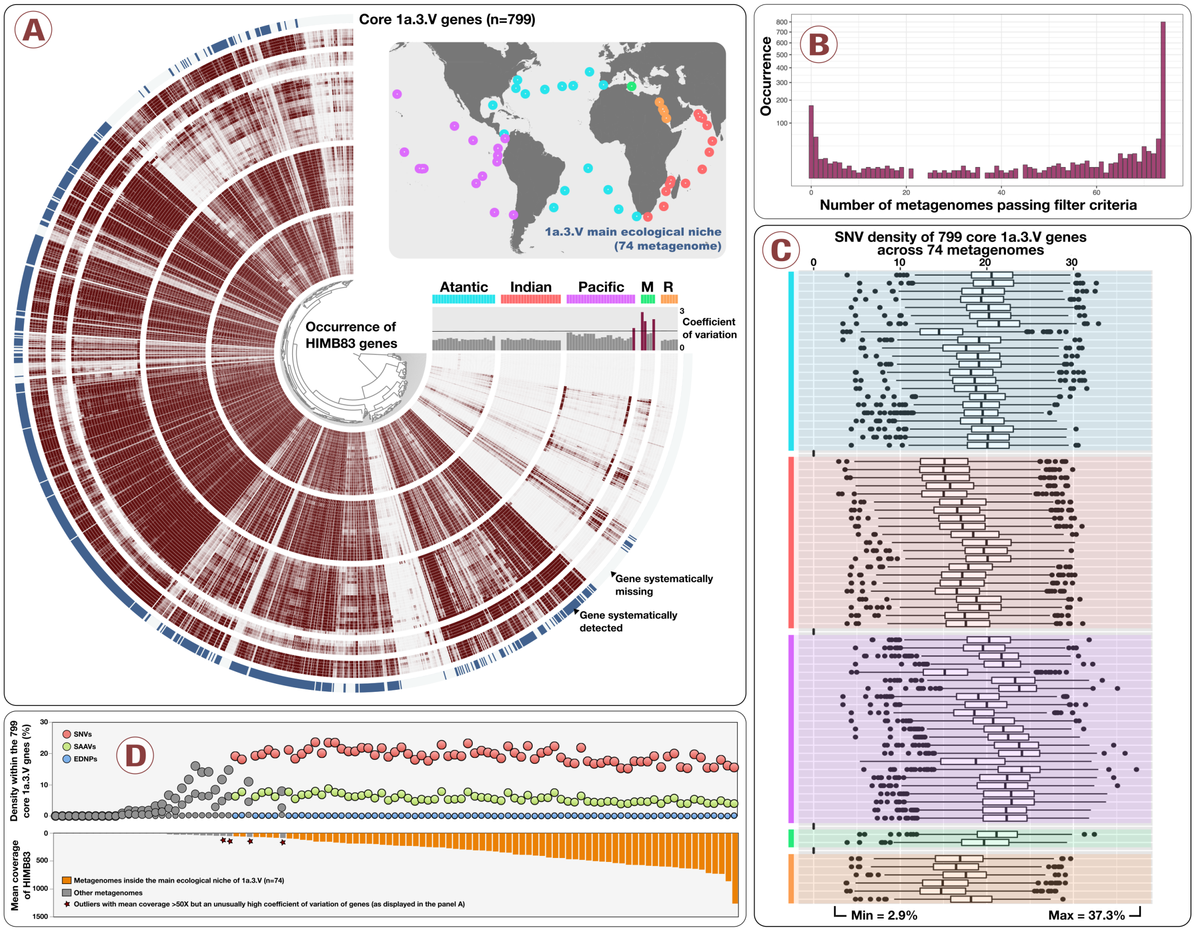


Figure 3.5: Distribution and diversity of the core 1a.3.V genes. Panel A displays the relative distribution of HIMB83 genes across 78 metagenomes, along with their coefficient of variation and the selection of 799 core 1a.3.V genes (blue outer circle). A world map provides the location of 74 metagenomes corresponding to the main ecological niche of 1a.3.V metagenomes (four metagenomes were disregarded due to high coefficients of variation). Panel B shows the number of metagenomes in which genes were consistently present. Genes were considered to be present in a metagenome consistently only if their coverage was within a factor of 5 of the average coverage of all genes for that metagenome. Those that passed the filter criteria in all 74 metagenomes (far right) were defined as the core 1a.3.V genes. Panel C displays the SNV density of core 1a.3.V genes across these 74 metagenomes. SNV density varied between 2.9% and 37.3% across genes and metagenomes. Panel D summarizes the heterogeneity extent of the core 1a.3.V genes within the population main ecological niche. Specifically, the panel displays the density of single nucleotide variants (>1% from consensus), environmentally disconnected nucleotide position (*i.e.*, positions stable in the environment but differing from the reference, <1% from consensus) and single amino acid variants (>10% from consensus) within the core genes of 1a.3.V across 103 metagenomes as a function of the mean coverage of HIMB83.

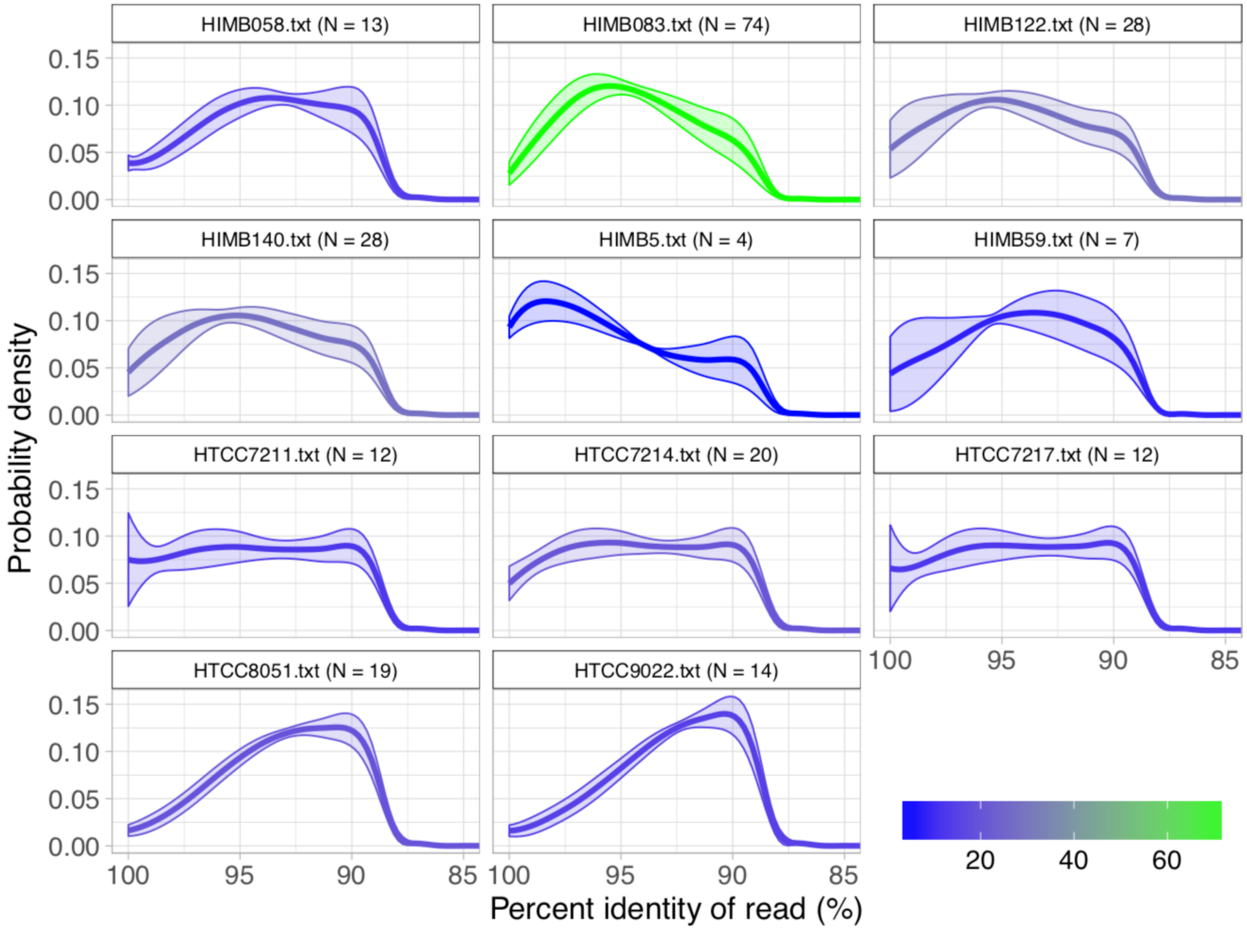


Figure 3.6: Percent identity distributions resulting from the competitive mapping experiment of the metagenomic short reads onto the 21 SAR11 reference genomes. For each reference genome, metagenomes were only included if they recruited at least 50X coverage. 10 references failed to recruit 50X coverage in any metagenome and were excluded from the plot. Curves were colored according to N, the number of metagenomes passing the 50X threshold, and each curve represents the mean distribution of these metagenomes, where the shaded area reflects the ± standard deviation. For visual clarity, the data only includes reads equal to the median read length.

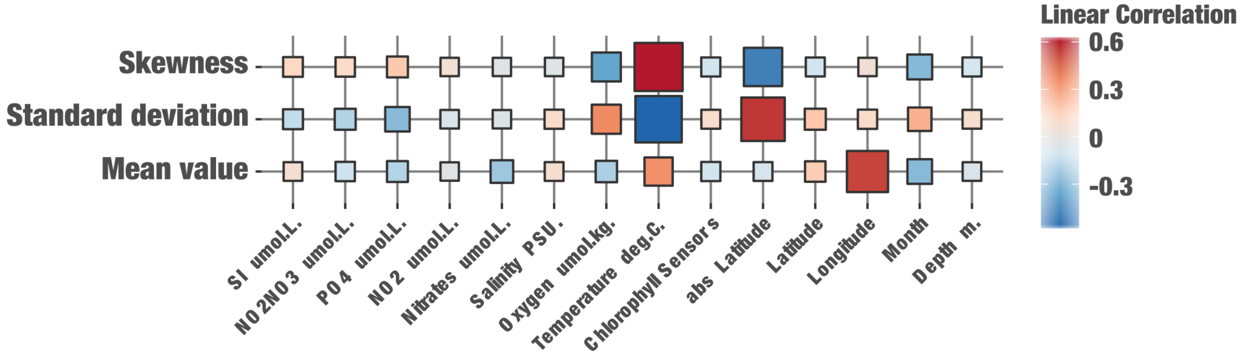


Figure 3.7: A matrix illustrating the degree of correlation (via linear regression) between oceanic metadata and the statistics (mean, standard deviation, skewness) of the percent read identity distributions of reads recruited by HIMB83 for the 74 metagenomes in which HIMB83 was covered at least 50X. For example, cells in the temperature column of this matrix quantify the linear correlation coefficients of the scatterplots shown in Figure 3.2B. Cell colors correspond to their linear correlation coefficient and sizes are proportional to R-squared values. We did not correct for multiple testing due to the potential for strong inter-dependence of the parameters. Table 3.4 reports full numerical statistics including p-values.

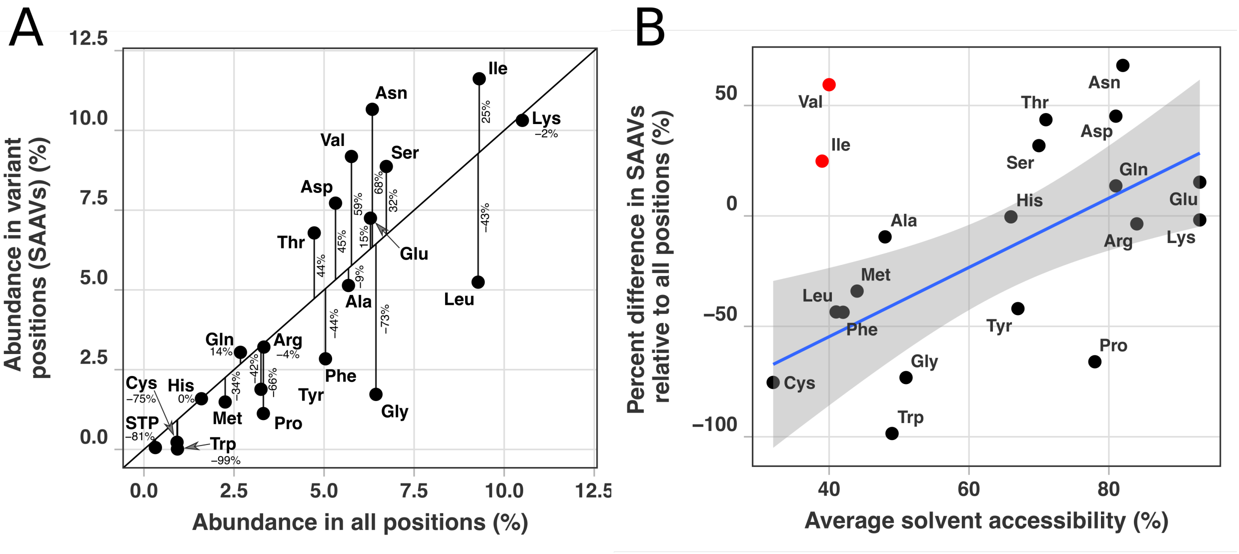


Figure 3.8: Panel A shows a direct comparison between the amino acid composition in all positions compared to the amino acid composition within SAAVs. The amino acid composition of all positions (y-axis) was quantified by calculating the frequency that amino acids appeared within the core 1a.3.V genes of the reference sequence, HIMB83, whereas amino acid composition in SAAVs (x-axis) was quantified by calculating the frequency that an amino acid was one of the two dominant alleles across the 738,324 SAAVs. The black diagonal line signifies a one-to-one correspondence between these two variables, and black vertical lines illustrate the deviation of amino acids from this null expectation. Vertical lines are labelled with percent differences between frequencies of amino acids in SAAVs relative to all positions. The linear correlation coefficient of panel A is 0.81, with an R-squared of 0.65, and a probability that no correlation exists of 9.8×10^{-6}. Panel B shows a comparison between the average solvent accessibility of amino acids (x-axis) to the percent difference between frequencies of amino acids in SAAVs relative to all positions (y-axis). Average solvent accessibilities for amino acids were taken from Table 2 of [Bordo and Argos, 1991]. The linear correlation coefficient of panel B excluding isoleucine and valine (in red) is 0.64, with an R-squared of 0.41, and a probability that no correlation exists of 0.004. Including isoleucine and valine, these values are 0.40, 0.16, and 0.08, respectively. The blue line shows the line of best fit excluding isoleucine and valine, with shaded in regions representing 95% confidence intervals.

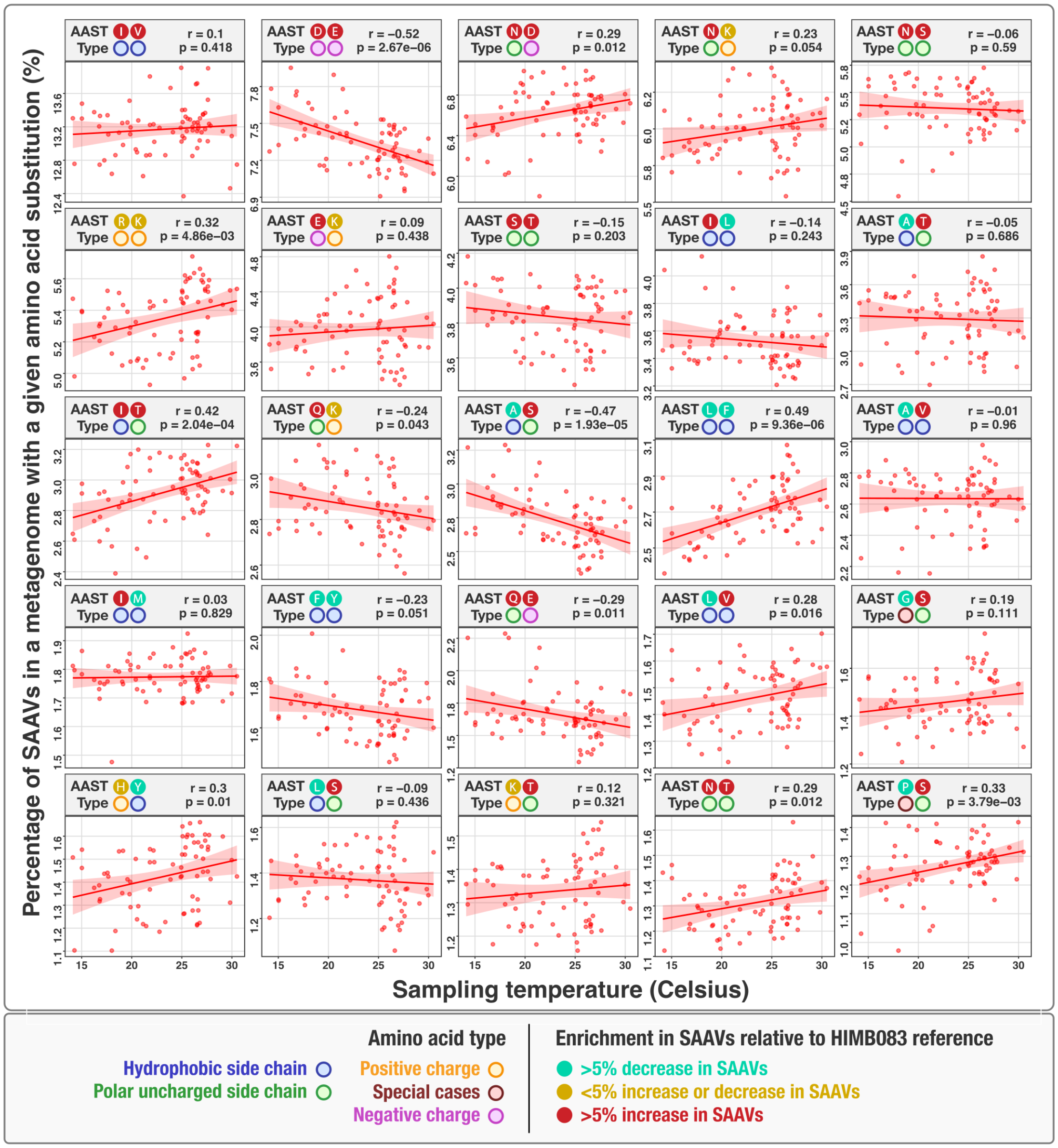


Figure 3.9: The top 25 most abundant amino acid substitution types (AASTs) and their relationship with *in situ* temperature. Each dot represents a metagenome, the x-axis is *in situ* temperature, and the y-axis is the percentage of SAAVs that were a given AAST. The red line is the line of best fit and the shaded-in region illustrates the 95% confidence interval. The linear correlation coefficient is given as r, and the probability of no relationship with temperature is given as p (uncorrected for multiple testing). Corrected p-values can be obtained by dividing p by 25, the number of linear regressions performed (Bonferonni multiple testing).

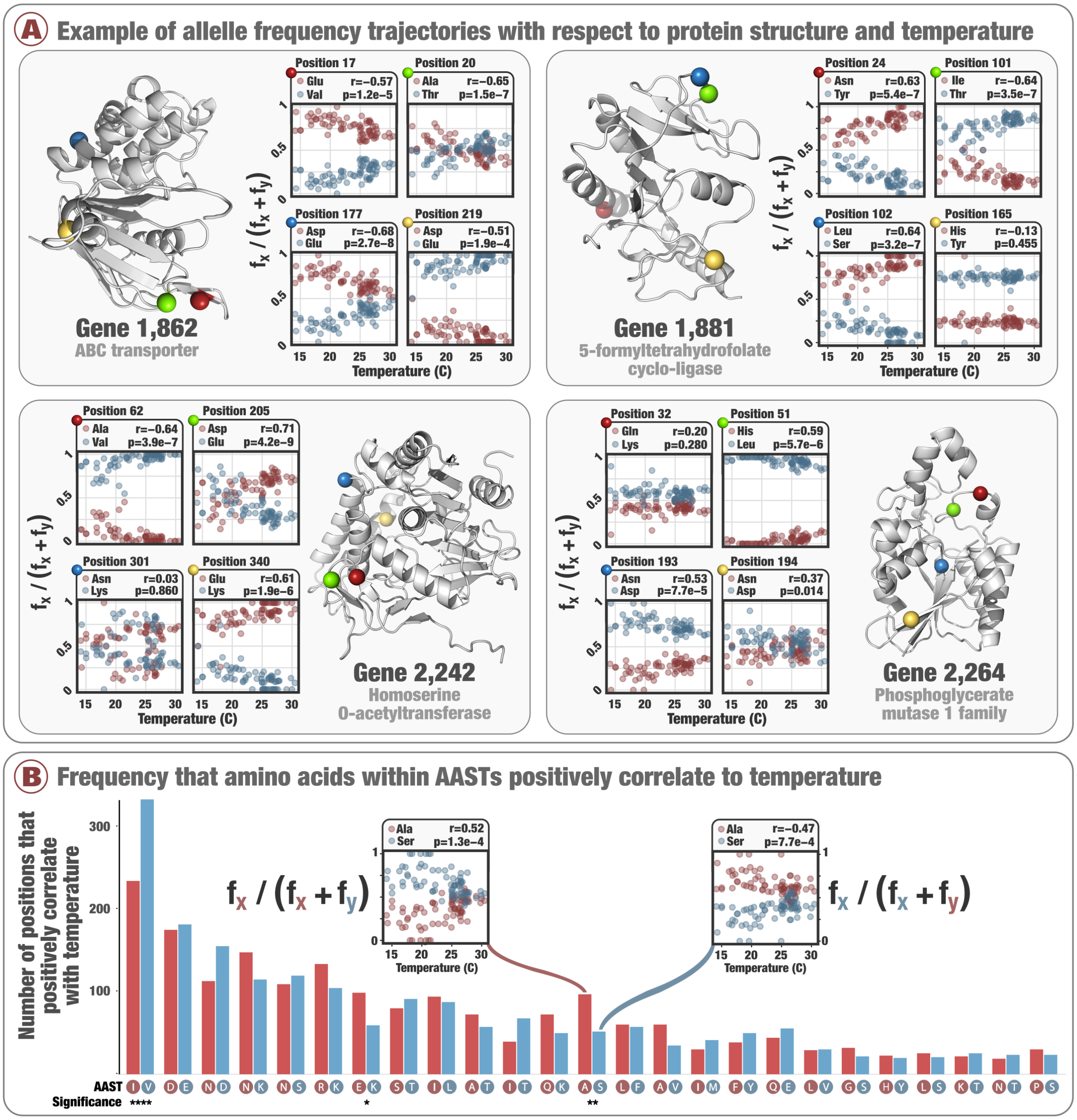


Figure 3.10: Allele frequency trajectories and *in situ* temperature. Panel A illustrates allele frequency trajectories across 74 metagenomes with respect to temperature for 16 manually chosen SAAV positions in the context of protein structures predicted from the core 1a.3.V genes. Only the two most abundant amino acids were considered for each SAAV. The linear correlation coefficient is denoted as r and the probability that no correlation exists is denoted as p (Benjamini–Hochberg corrected p-values <0.05 (i.e. false discovery rate of 5%)). r is defined such that a positive value refers to the first amino acid (in dark red) positively correlating with temperature. Panel B shows the 4,592 positions within the core 1a.3.V

Figure 3.10 continued: genes that had temperature-correlated allele frequency trajectories (Benjamini–Hochberg corrected p-value <0.05), and the number of times these positions corresponded to the top 25 most prevalent AASTs. On the x-axis are the 25 most prevalent AASTs, where red and blue bars indicate the number of times the first and second amino acid, respectively, had allele frequency trajectories that positively correlated with temperature. The insets illustrate two example allele frequency trajectories for the AAST 'alanine/serine'.

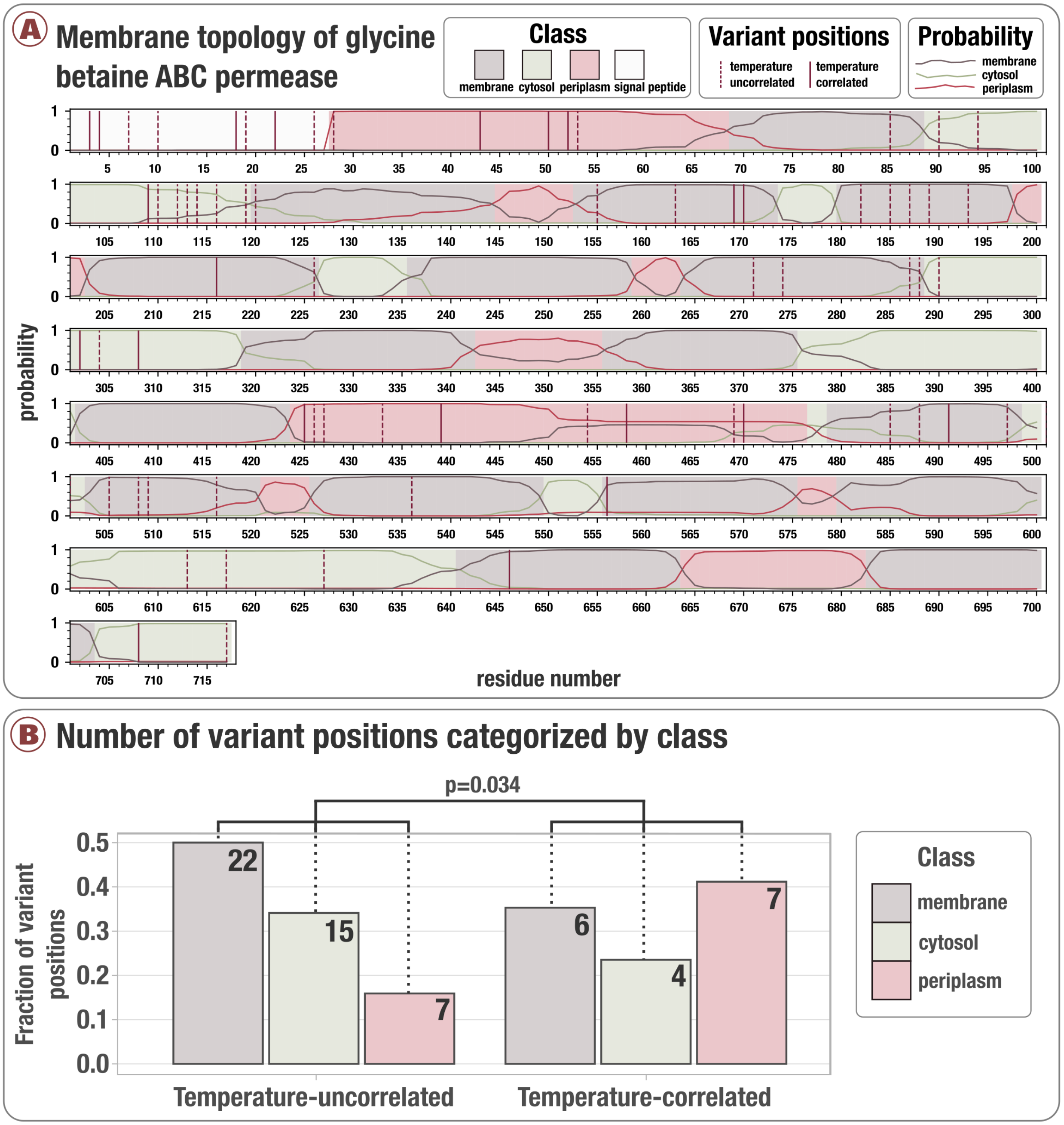


Figure 3.11: Analysis of how temperature-correlated variant positions distribute within Gene 1727, a glycine betaine ATP-binding cassette permease subunit identified for its rare proportion of temperature-correlated variant positions. Panel A illustrates the membrane topology predicted by Phobius (See Materials and Methods), which associates to each position a probability it is periplasmic, cytosolic, transmembrane, or within the signaling peptide (colored lines). We categorized each position according to the maximum probability (shaded regions). Temperature-correlated and uncorrelated variant positions are denoted by solid

Figure 3.11 continued: red and dashed red vertical lines, respectively. Panel B illustrates the frequency that variant positions are found in the membrane, cytosol, and periplasm, based on whether or not they correlate with temperature. The y-axis is the fraction of variant positions observed (excluding positions observed in the signaling peptide), and numbers within each bar denote the number of variant positions observed within each class. The probability that positions are distributed independent of temperature correlation was 0.034. Formally, this is the probability that temperature-correlated positions were distributed according to a trinomial distribution with a probability vector equal to the empirical distribution observed in temperature-uncorrelated positions.

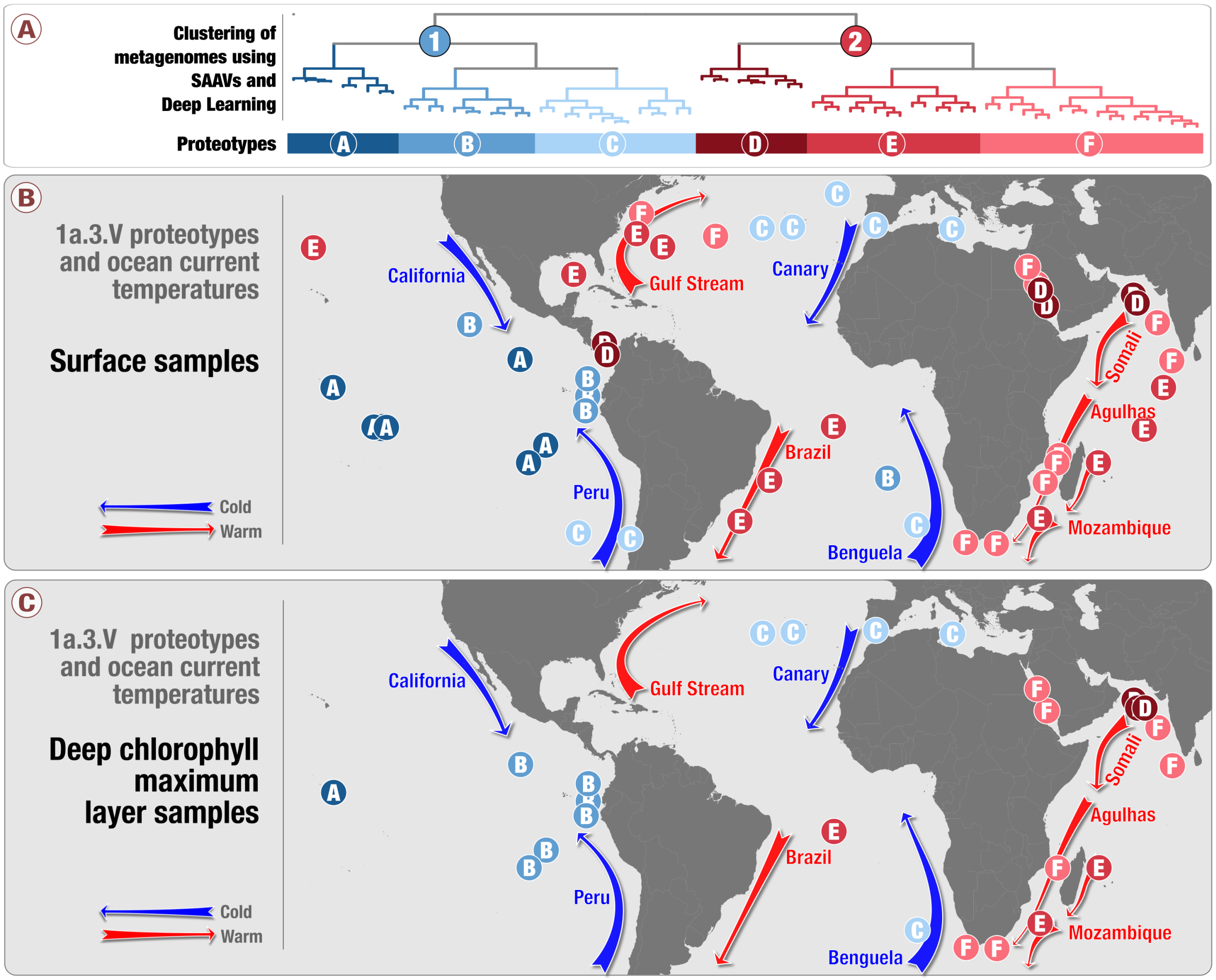

A
Clustering of metagenomes using SAAVs and Deep Learning
1
2
Proteotypes
A
B
C
D
E
F
B
1a.3.V proteotypes and ocean current temperatures
Surface samples
Cold
Warm
California
Gulf Stream
Canary
Somali
Agulhas
Brazil
Peru
Benguela
Mozambique
C
1a.3.V proteotypes and ocean current temperatures
Deep chlorophyll maximum layer samples
Cold
Warm
California
Gulf Stream
Canary
Somali
Agulhas
Brazil
Peru
Benguela
Mozambique

Figure 3.12: A comparison of the geographic partitioning of the 1a.3.V groups and proteotypes between metagenomes sampled from the surface versus those sampled from the deep maximal chlorophyll layer. Panels A and B contain the same information as shown in Figure 3.4, and panel C shows the locations of deep maximal chlorophyll layer metagenomes and the proteotypes to which they belong.

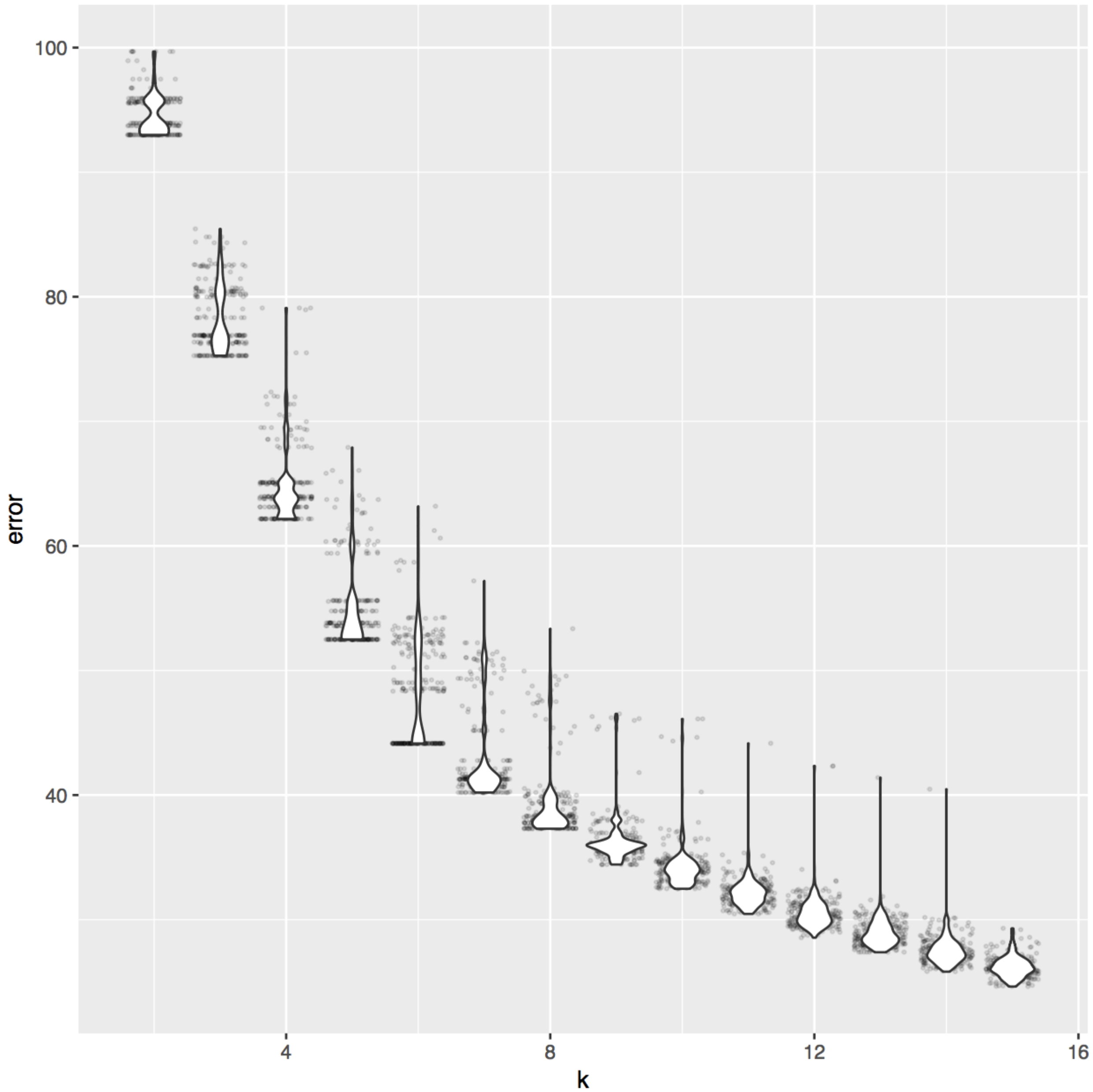


Figure 3.13: K-means clustering results (250 iterations) of the Deep Learning distance metric of 74 metagenomes based on the coordinates and identity of 738,324 SAAVs. The elbow of this curve is $k = 6$, which informed our decision to define six proteotypes.

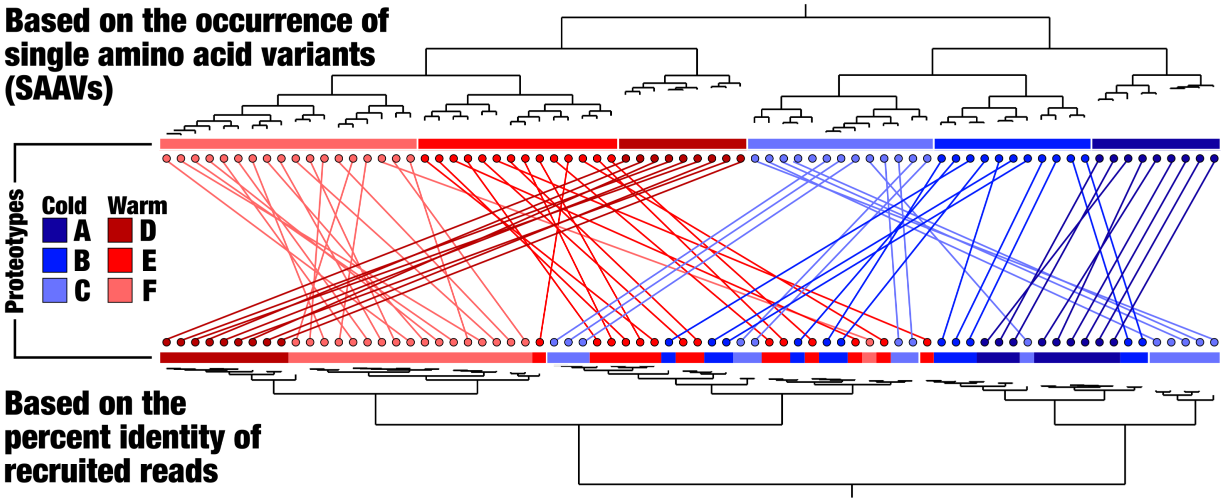
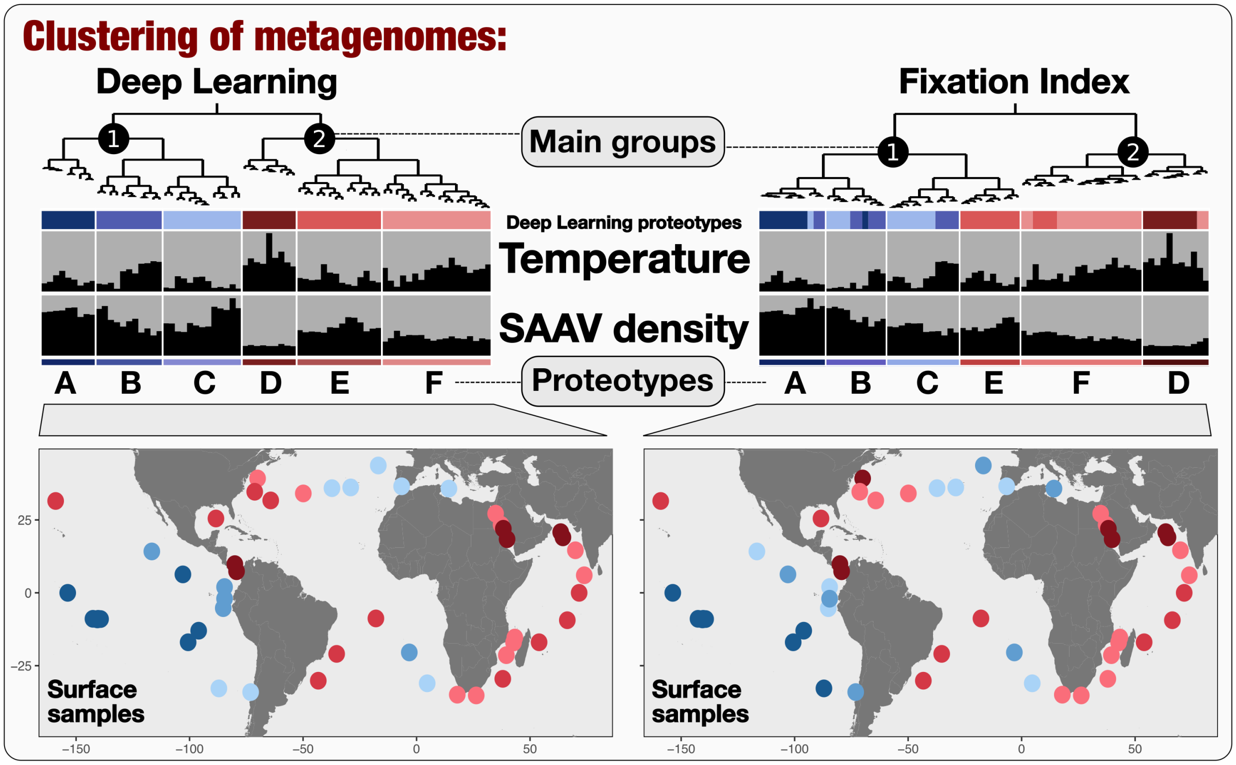


Figure 3.14: A comparison of dendrograms that organize metagenomes based on the genomic variability observed in the core 1a.3.V genes. Above, the metagenomes are organized by applying a novel graph-based activity regularization technique for competitive learning from hyper-dimensional data to the 738,324 SAAVs (see Materials and Methods). Below, the metagenomes are organized by a less comprehensive approach in which the percent identity distributions are hierarchically clustered via Euclidean distance. The first method therefore is sensitive to the identity and positions of amino acid variability, whereas the second method is based solely on a summary statistic that quantifies the degree of sequence divergence at the DNA level and is agnostic to position and identity. Each metagenome is connected to itself through a straight line and is colored according to the proteotype to which it belongs.



One way ANOVA test:

	Deep Learning		Fixation Index	
	Main groups	Proteotypes	Main groups	Proteotypes
Temperature	0.0031**	3.57e-07***	0.0975	0.00196**
SAAV density	7.75e-13***	<2.e-16***	2.71e-16***	<2.e-16***

Figure 3.15: Biogeography of SAR11 subclade 1a.3.V based on single amino acid variants using Deep Learning (left) and Fixation Index (right) (see Material and Methods). The world maps display the geographic partitioning of six proteotypes based on the two methodologies. Finally, ANOVA tests determine whether there are statistically significant differences between the means of either *in situ* temperature or SAAV density across (1) the two main groups and (2) the six proteotypes as inferred from the two methodologies.

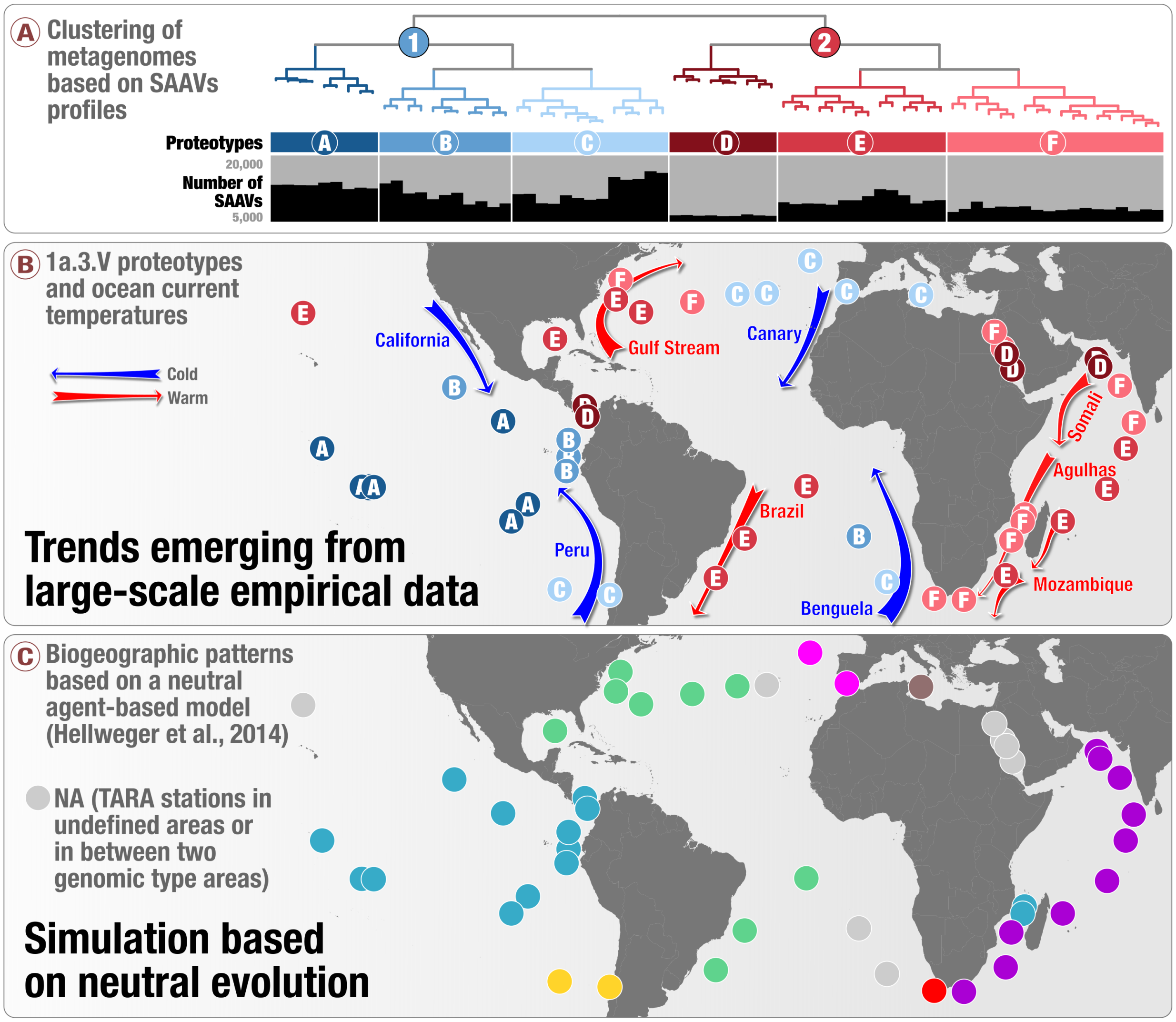

A Clustering of metagenomes based on SAAVs profiles
1
2
Proteotypes
A
B
C
D
E
F
20,000
Number of SAAVs
5,000
B 1a.3.V proteotypes and ocean current temperatures
Cold
Warm
California
Gulf Stream
Canary
Somali
Agulhas
Brazil
Peru
Benguela
Mozambique
Trends emerging from large-scale empirical data
C Biogeographic patterns based on a neutral agent-based model (Hellweger et al., 2014)
NA (TARA stations in undefined areas or in between two genomic type areas)
Simulation based on neutral evolution

Figure 3.16: Geographic partitioning of SAR11 by matching surface metagenomes analyzed in our study to simulated results determined using a neutral-agent based model [Hellweger et al., 2014]. The figure emphasizes biogeographic differences between this simulation focused on neutral evolution and our large-scale empirical analysis.

3.9 Supplementary Tables

All supplementary tables are available at `https://figshare.com/articles/SAR11_SAAVs_Tables_and_Figures/5254537`.

Table 3.1: Summary of 21 SAR11 genomes and 103 metagenomes from TARA Oceans and Ocean Sampling Day.

Table 3.2: Summary of metagenomic reads recruitment. The table describes the number of recruited reads, mean coverage, relative distribution and level of detection for 21 SAR11 genomes across 103 metagenomes, along with the total number of single nucleotide variants and single amino acid variants identified in each metagenome. The table also summarizes the relative distribution of 21 SAR11 genomes, along with 31 Prochlorococcus genomes and 957 marine population genomes across 103 metagenomes.

Table 3.3: Summary of the SAR11 metapangenomic analysis. The table describes the functionality of genes identified in the 21 SAR11 genomes and links each gene to a gene cluster in the SAR11 pangenome. In addition, the table describes gene clusters containing proteins translated from the 799 core 1a.3.V genes, which were independently identified using the coverage values of HIMB83 genes across 74 metagenomes.

Table 3.4: Distance metric of 21 SAR11 genomes based on their 16S rRNA gene sequence similarities.

Table 3.5: Summary of the HIMB83 genes. The table describes the length, functionality and nucleotide sequence of 1,470 genes identified from HIMB83, along with their normalized distribution across 103 metagenomes. The table also summarizes the coefficient of variation of genes in each metagenome determined from these distribution values.

Table 3.6: Summary of the degree of correlation of oceanic metadata to the percent identity histograms of reads recruited to HIMB83.

Table 3.7: Summary of variability of the core 1a.3.V genes. The table describes the density of single nucleotide variants and amino acid variants of each core 1a.3.V genes across 74 metagenomes, as well as the occurrence of invariant proteins.

Table 3.8: Summary of SAAVs in the core 1a.3.V genes. The table describes the coordinates and identity (defined by the two most frequent amino acids) for codon positions in 738,324 core 1a.3.V genes (1) that were covered more than 20X across 74 metagenomes, (2) and in which a divergence >10% from consensus was observed in the frequency of amino acids. The table also links each SAAV to the metagenome it was identified from, and summaries a gene-level analysis of SAAVs.

Table 3.9: Summary of individual amino acids and AASTs involved in 738,324 SAAVs identified within the core 1a.3.V genes. The table describes the proportion of amino acids involved in SAAVs across the 74 metagenomes compared to the core 1a.3.V genes. The table also describes the proportion of acid substitution types (defined by the two most abundant amino acids involved in each SAAV) across 74 metagenomes, in the context of amino acid frequencies in the core 1a.3.V genes.

Table 3.10: SAAV characteristics, including solvent accessibility, for each SAAV belonging to a core 1a.3.V gene with a successfully predicted protein structure.

Table 3.11: The correlation of allele frequencies to temperature for each position containing at least one SAAV. The summary of proportion of temperature-correlated to temperature-uncorrelated positions per gene.

Table 3.12: Deep Learning distance metric of 74 metagenomes based on the coordinates and identity of 738,324 SAAVs.

Table 3.13: Statistical significance of metadata collected in this study and by the TARA Oceans consortium to explain the grouping of six proteotypes with Deep Learning.

CHAPTER 4

STRUCTURE-INFORMED MICROBIAL POPULATION GENETICS ELUCIDATE SELECTIVE PRESSURES THAT SHAPE PROTEIN EVOLUTION

This chapter is derived from the following publication:

Kiefl, Evan, Ozcan C. Esen, Samuel E. Miller, Kourtney L. Kroll, Amy D. Willis, Michael S. Rappé, Tao Pan, and A. Murat Eren. 2022. "Structure-Informed Microbial Population Genetics Elucidate Selective Pressures That Shape Protein Evolution." bioRxiv. https://doi.org/10.1101/2022.03.02.482602.

4.1 Author contributions

EK and AME conceptualized the study and interpreted findings. EK curated data, developed software tools, and performed primary analyses. OCE and AME contributed software. EK and AME wrote the paper. SEM, KK, and ADW helped with data analyses and interpretation. MSP and TP helped with project management and funding acquisition. AME supervised the project. All authors commented on the drafts of the study.

4.2 Abstract

Comprehensive sampling of natural genetic diversity with metagenomics enables highly resolved insights into the interplay between ecology and evolution. However, intra-population genomic variation represents the outcome of both stochastic and selective forces, making it difficult to identify whether variants are maintained by adaptive, neutral, or purifying processes. This is partly due to the reliance on gene sequences to interpret variants, which

disregards the physical properties of three-dimensional gene products that define the functional landscape on which selection acts. Here we describe an approach to analyze genetic variation in the context of predicted protein structures, and apply it to study a marine microbial population within the SAR11 subclade 1a.3.V, which dominates low-latitude surface oceans. Our analyses reveal a tight association between the patterns of nonsynonymous polymorphism, selective pressures, and structural properties of proteins such as per-site relative solvent accessibility and distance to ligands, which explain up to 59% of genetic variance in some genes. In glutamine synthetase, a central gene in nitrogen metabolism, we observe decreased occurrence of nonsynonymous variants from ligand binding sites as a function of nitrate concentrations in the environment, revealing genetic targets of distinct evolutionary pressures maintained by nutrient availability. Our data also reveals that rare codons are purified from ligand binding sites when genes are under high selection, demonstrating the utility of structure-aware analyses to study the variants that likely impact translational processes. Overall, our work yields insights into the governing principles of evolution that shape the genetic diversity landscape within a globally abundant population, and makes available a software framework for structure-aware investigations of microbial population genetics.

4.3 Significance

Increasing availability of metagenomes offers new opportunities to study evolution, but the equal treatment of all variants limits insights into drivers of sequence diversity. By capitalizing on recent advances in protein structure prediction capabilities, our study examines subtle evolutionary dynamics of a microbial population that dominates surface oceans through the lens of structural biology. We demonstrate the utility of structure-informed metrics to understand the distribution of nonsynonymous polymorphism, establish insights into the impact of changing nutrient availability on protein evolution, and show that even synonymous variants are scrutinized strictly to maximize translational efficiency when selection is high. Overall, our work illustrates new opportunities for discovery at the intersection between metagenomics and structural bioinformatics, and offers an interactive and scalable software platform to visualize and analyze genetic variants in the context of predicted protein structures and ligand-binding sites.

4.4 Introduction

Genetic diversity within populations emerges from and is shaped by a combination of stochastic and selective pressures, which often lead to phenotypic differences between closely related individuals, sometimes within a few generations [Burke et al., 2010, Lenski et al., 1991]. Surveys of microbial communities within natural habitats through phylogenetic marker genes [Olsen et al., 1986, Acinas et al., 2004, Sogin et al., 2006] and metagenomics [Simmons et al., 2008, Allen et al., 2007] have revealed a tremendous amount of genetic variation within environmental populations [Curtis and Sloan, 2005, Curtis et al., 2006], and an ever-increasing number of available genomes and metagenomes have provided insight into the selective pressures that shape such variation. However, the overwhelming complexity and dynamicity of these selective pressures, which occur even in the simplest environments [Good et al., 2017], has hindered our ability to determine which variants are under the influence of which pressures [Ochman, 2003, Mes, 2008].

Inferring selective pressures through the isolation of microbial strains and comparative genomics has been widely successful. More recently, metagenome-assembled genomes [Chen et al., 2020] and single-amplified genomes [Woyke et al., 2017] have dramatically increased the number [Almeida et al., 2021, Pachiadaki et al., 2019, Paoli et al., 2021] and diversity [Hug et al., 2016] of microbial clades represented in genomic databases, offering an even larger sampling of environmental microbes to study the emergence and maintenance of genetic variation [Garud and Pollard, 2020]. Nevertheless, genomes represent static snapshots of individual members of often complex environmental populations, and thus, working with genomic sequences alone substantially undersamples genetic variability in natural habitats and its associations with environmental and ecological forces [Van Rossum et al., 2020]. This shortcoming is partially addressed by shotgun metagenomics [Quince et al., 2017] and metagenomic read recruitment, where environmental sequences that are aligned to a reference can be studied to identify genetic variants at the resolution of single nucleotides [Whitaker and Banfield, 2006, Denef, 2019]. In particular, using genomes to recruit reads from metagenomes enables a comprehensive sampling of all genetic variants within environmental populations [Simmons et al., 2008]. Due to the immensity of sequencing data generated by metagenomic studies, even subtle genetic variation in natural populations is now resolvable, making it possible to explicitly correlate patterns of genomic variation with temporal or spatial environmental variables to elucidate the interplay between ecology and evolution [Schloissnig et al., 2013, Bendall et al., 2016, Anderson et al., 2017, Delmont et al., 2019, Garud et al., 2019, Zhao et al., 2019, Shenhav and Zeevi, 2020, Olm et al., 2021, Conwill et al., 2022]. Although quantification and analysis of sequence variants derived from metagenomic data has improved dramatically, inferring the functional impact of individual nucleotides remains a fundamental challenge in part due to the sole reliance on DNA sequences, which do not represent physical properties of proteins they encode, and thus disguise the functional impact of individual mutations.

Given the intermediary role that structure plays within the ‘sequence-structure-function paradigm’ [Anfinsen, 1973], including protein structures as a dimension of analysis is commonplace in studies of protein evolution [Siltberg-Liberles et al., 2011, Harms and Thornton, 2013, Sikosek and Chan, 2014], and it is appreciated that the accuracy of evolutionary models improves with combined analyses of protein structures and the evolution of underlying sequences [Wilke, 2012]. In contrast, the state-of-the-art approaches that quantify genetic variants in environmental microbial populations typically treat genes as strings of nucleotides [Schloissnig et al., 2013, Murat Eren et al., 2015, Nayfach et al., 2016, Costea et al., 2017, Olm et al., 2021]. While this strategy enables rapid surveys of population dynamics through single-nucleotide variants, it disregards the physical properties of three-dimensional gene products that selection acts upon, and thus misses a critical intermediate to understand the relationship between selection and fitness [Golding and Dean, 1998, Chen and Arnold, 1993]. The importance of mapping sequence variants on predicted protein structures to identify genetic determinants of phenotypic variation has been noted more than two decades ago [Sunyaev et al., 2001], yet the limited availability of protein structures has historically rendered protein structure-informed microbial population genetics impractical. Given dramatic advances in both solving and predicting protein structures in recent years [Kuhlman and Bradley, 2019], most notably deep learning approaches such as AlphaFold [Jumper et al., 2021] that offer highly accurate protein structure predictions, this constraint is likely a problem of the past. Altogether, open questions in microbial ecology and evolution, advances in computation, and increased availability of data are culminating in a research landscape that is ripe for new software solutions that integrate protein structures with 'omics data in order to observe and interpret evolutionary processes that shape sequence variation in natural populations.

Here we develop an interactive and scalable software solution for the analysis and interactive visualization of metagenomic sequence variants in the context of predicted protein structures and ligand binding sites as a new module in anvi'o, an open-source, community-led multi-omics platform (`https://anvio.org`). By importing AlphaFold-predicted protein structures into anvi'o structure, we (1) demonstrate the shortcomings of purely sequence-based approaches to interpret patterns of polymorphism observed within complex microbial populations, (2) propose two structural features to interpret genetic variation, relative solvent accessibility (RSA) and distance-to-ligand (DTL), (3) illustrate that nonsynonymous polymorphism is more likely to encroach upon active sites when selection is low, but is purged from active sites when selection is high, and (4) provide evidence that common codons are more translationally robust than their rare synonymous counterparts, which appear within structurally/functionally noncritical sites when selection is low.

4.5 Results and Discussion

To investigate selective pressures that drive protein evolution within microorganisms inhabiting complex naturally occurring environments, we chose a single microbial taxon and a set of metagenomes that match to its niche boundaries: SAR11 (Candidatus Pelagibacter ubique), a microbial clade of free-living heterotrophic alphaproteobacteria that dominates surface ocean waters [Morris et al., 2002], and Tara Oceans Project metagenomes [Sunagawa et al., 2015], a massive collection of deeply sequenced marine samples from oceans and seas across the globe. SAR11 is divided into multiple subclades with distinct ecology [Giovannoni, 2017]. Thus, we further narrowed our focus to HIMB83, a single SAR11 strain genome that is 1.4 Mbp in length. HIMB83 is a member of the environmental SAR11 lineage 1a.3.V, one of the most abundant [Nayfach et al., 2016] and most diverse [Delmont et al., 2019] microbial lineages in marine systems, which recruits as much as 1.5% of all metagenomic short reads in surface ocean metagenomes [Delmont et al., 2019].

To quantify the genetic variability of 1a.3.V, we used HIMB83 as a reference genome of the subclade, and competitively recruited short reads (see Methods) from 93 low-latitude surface ocean metagenomes (Table 4.1), resulting in 390 million reads that were 94.5% identical to HIMB83 on average (Figure 4.5). As an individual member of a diverse subclade, HIMB83 possesses a genomic context that is insufficient for resolving the extent of genetic diversity within 1a.3.V. Regardless, HIMB83 possesses the 'core' gene set of 1a.3.V, and so reads recruited by these genes represent the diversity of the 1a.3.V core genome. Of the 1,470 genes in HIMB83, we restricted our analysis to 799 genes that we determined to form the 1a.3.V core genes, and 74 metagenomes in which the average coverage of HIMB83 exceeded 50X (see Methods). The reads recruited to the 1a.3.V core represent a dense sampling of the diversity within this environmental lineage that far exceeds the evolutionary resolution and volume of sequence data achievable through comparisons of cultured SAR11 genomes alone (Figure 4.5). As a result, these data provide a unique opportunity to zoom in and track how genomic variation in one of the most abundant microbial populations on Earth shifts in response to ecological parameters throughout the global ocean (Figure 4.6).

4.5.1 Polymorphism rates reveal intense purification of nonsynonymous mutants

To quantify genomic variation in 1a.3.V, in each sample we identified codon positions of HIMB83 where aligned metagenomic reads did not match the reference codon. We considered each such position to be a single codon variant (SCV). Analogous to single nucleotide variants (SNVs), which quantify the frequency that each nucleotide allele (A, C, G, T) is observed in the reads aligning to a nucleotide position, SCVs quantify the frequency that each codon allele (AAA, . . . , TTT) is observed in the reads aligning to a codon position (see Methods for a more complete description). Since SCVs are defined to be 'in-frame', they provide inherent convenience when relating nucleotide variation in the genomic coordinates to amino acid variation in the corresponding protein coordinates, as well as for determining whether or not nucleotide variation leads to synonymous or nonsynonymous change. Within the 1a.3.V core genes, we found a total of 9,537,022 SCVs, or 128,879 per metagenome on average. These SCVs distributed throughout the genome such that 78% of codons (32% of nucleotides) exhibited minor allele frequencies >10% in at least one metagenome. Despite this extraordinary level of diversity, our read recruitment strategy is stringent and yields reads that on average differ from HIMB83 in only 6 nucleotides out of 100 (Table 4.2), precluding the possibility that this diversity is generated from excessive nonspecific mapping. While puzzling, this level of diversity is not surprising as it agrees with numerous studies that have pointed out the astonishing complexity of the SAR11 subclade 1a.3.V [Nayfach et al., 2016, Delmont et al., 2019, Haro-Moreno et al., 2020] that could not be further divided into sequence-discrete populations [Delmont et al., 2019].

We found this diversity to be overwhelmingly synonymous. By splitting each SCV into its synonymous (s) and nonsynonymous (ns) proportions, we calculated per-site rates of s-polymorphism and ns-polymorphism as $\mathrm{pS}^{(\mathrm{site})}$ and $\mathrm{pN}^{(\mathrm{site})}$, not to be confused with the related concepts dS and dN. While dS and dN quantify rates of synonymous and nonsynonymous substitution between diverged species, $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ can (1) resolve shorter evolutionary timescales than the characteristic fixation rate, (2) be calculated from metagenomic read recruitment data without complete haplotypes, and (3) define rates on a per-sample basis, thus enabling inter-sample comparisons. Overall, we found that the average $\mathrm{pS}^{(\mathrm{site})}$ outweighed $\mathrm{pN}^{(\mathrm{site})}$ by 19:1 (Table 4.3), revealing an overwhelming fraction of the 1a.3.V diversity to be synonymous and illustrating how nonsynonymous mutants are purified at a much higher rate than synonymous mutants in the population at large.

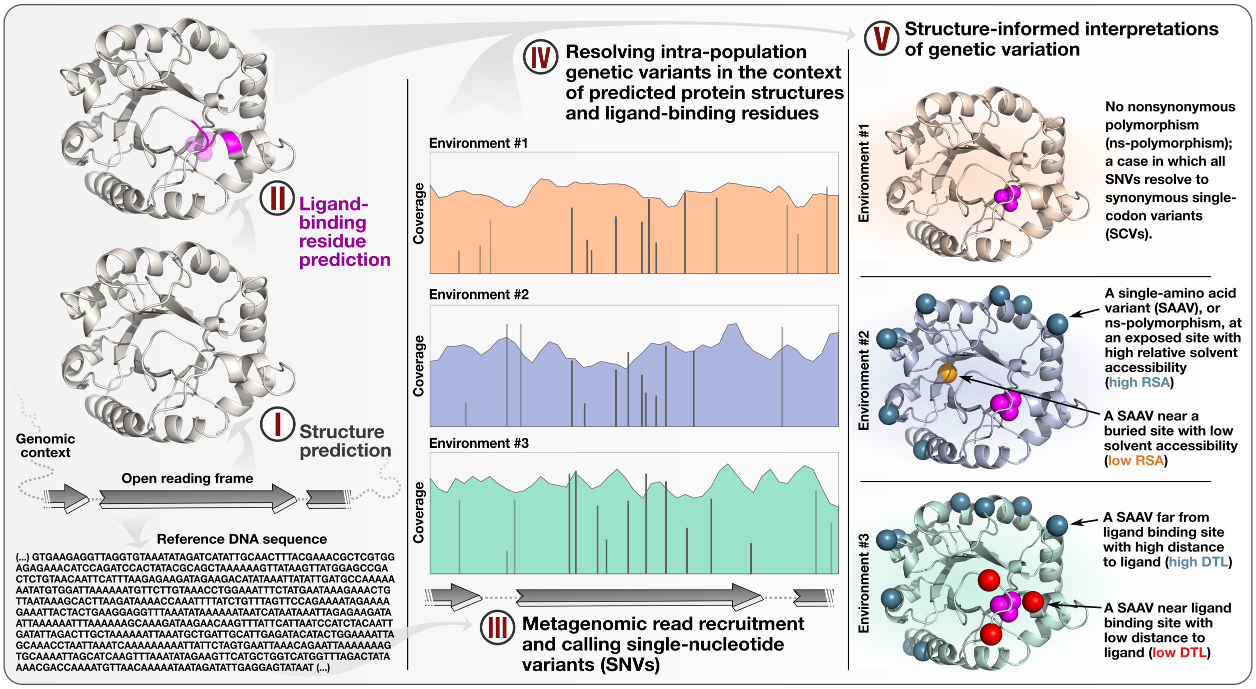

IV Resolving intra-population genetic variants in the context of predicted protein structures and ligand-binding residues
V Structure-informed interpretations of genetic variation
II Ligand-binding residue prediction
I Structure prediction
Genomic context
Open reading frame
Reference DNA sequence
(...) GTGAAGAGGTTAGGTGTAAATATAGATCATATTGCAACTTTACGAAACGCTCGTGG
AGAGAAACATCCAGATCCACTATACGCAGCTAAAAAAGTTATAAGTTATGGAGCCGA
CTCTGTAACAATTCATTTAAGAGAAGATAGAAGACATATAAATTATATTGATGCCAAAAA
AATATGTGGATTAAAAAATGTTCTTGTAAACCTGGAAATTTCTATGAATAAAGAAACTG
TTAATAAAGCACTTAAGATAAAACCAAATTTTATCTGTTTAGTTCCAGAAAATAGAAAA
GAAATTACTACTGAAGGAGGTTTAAATATAAAAAATAATCATAATAAATTAGAGAAGATA
ATTAAAAAATTTAAAAAAGCAAAGATAAGAACAAGTTTATTCATTAATCCATCTACAATT
GATATTAGACTTGCTAAAAAATTAAATGCTGATTGCATTGAGATACATACTGGAAAATTA
GCAAACCTAATTAAATCAAAAAAAAATTATTCTAGTGAATTAAACAGAATTAAAAAAAG
TGCAAAATTAGCATCAAGTTTAAATATAGAAGTTCATGCTGGTCATGGTTTAGACTATA
AAACGACCAAAATGTTAACAAAAATAATAGATATTGAGGAGTATAAT (...)
Environment #1
Environment #2
Environment #3
Coverage
III Metagenomic read recruitment and calling single-nucleotide variants (SNVs)
Environment #1
No nonsynonymous polymorphism (ns-polymorphism); a case in which all SNVs resolve to synonymous single-codon variants (SCVs).
Environment #2
A single-amino acid variant (SAAV), or ns-polymorphism, at an exposed site with high relative solvent accessibility (high RSA)
A SAAV near a buried site with low solvent accessibility (low RSA)
Environment #3
A SAAV far from ligand binding site with high distance to ligand (high DTL)
A SAAV near ligand binding site with low distance to ligand (low DTL)

Figure 4.1: Anvi'o workflow for structure-informed population genetics.

4.5.2 Nonsynonymous polymorphism avoids buried sites

$\text{pN}^{(\text{site})}$ values varied significantly from site-to-site and from sample-to-sample, but overall, more variation existed between sites in a given sample than between samples of a given site (Figure 4.7). The extent that a given site can tolerate ns-polymorphism is largely determined by the local physicochemical environment of the encoded residue, which is defined by the 3D structure of the protein. Thus, we broadened our focus by developing a computational framework, anvi'o structure (Supplementary Information), that enabled the integration of environmental sequence variability with predicted protein structures (Figure 4.1).

We used two independent methods to predict protein structures for the 799 core genes of 1a.3.V: (1) a template-based homology modeling approach with MODELLER [Webb and Sali, 2016], which predicted 346 structures, and (2) a transformer-like deep learning approach with AlphaFold [Jumper et al., 2021], which predicted 754. Our evaluation of the 339 genes for which both methods predicted structures (Supplementary Information) revealed a comparable accuracy between AlphaFold and MODELLER (Figure 4.8, Table 4.4). Thus, we opted to use AlphaFold structures for all downstream analyses due to its higher structural coverage. Indeed, AlphaFold-predicted protein structures covered over 90% of the core genes, highlighting the emerging opportunities afforded by recent advances in *de novo* structure prediction.

Aligning single-codon variants to predicted structures enabled us to directly compare the distributions of s-polymorphism and ns-polymorphism rates relative to biophysical characteristics of the encoded proteins. We first investigated the association between polymorphism rates and relative solvent accessibility (RSA), a biophysical measure of how exposed (RSA = 1) or buried (RSA = 0) a site is. Since nonsynonymous mutations at buried sites are more likely to disrupt folding and stability, RSA serves as a powerful proxy to discuss the strength of structural constraints acting at a site [Echave et al., 2016]. By calculating RSA for each site in the predicted structures, and then weighting every site by the $\text{pN}^{(\text{site})}$ and $\text{pS}^{(\text{site})}$ across all samples, we established proteome-wide distributions for $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ relative to RSA (Figure 4.2a). These data showed that $\mathrm{pS}^{(\mathrm{site})}$ closely resembled the null distribution, which illustrates the lack of influence of RSA on s-polymorphism, while $\mathrm{pN}^{(\mathrm{site})}$ deviated significantly and instead exhibited strong preference for sites with higher RSA. This finding aligns well with the expectation that buried sites are likely to purify nonsynonymous change due to disruption of protein stability while being relatively more tolerant to synonymous change, and validates our methodology.

4.5.3 Nonsynonymous polymorphism avoids active sites

While structural constraints ensure a given protein folds properly and remains stable, they do not guarantee its function. Comprehensive analyses of diverse protein families show that residues that bind or interact with ligands are depleted of mutations [Kobren and Singh, 2019] due to strong selective pressures that maintain active site conservancy. This constraint is not limited to the immediate vicinity of ligand-binding residues and has been observed to radiate outwards from the active site with a strength inversely correlated with distance from active site [Dean et al., 2002, Jack et al., 2016]. More generally, it has been observed that conserved sites induce ‘conservation gradients’ that surround them, leading to increased conservation amongst neighboring sites [Sharir-Ivry and Xia, 2021]. Based on these ideas, we conceptualized the metric ‘distance-to-ligand’ (DTL) as the distance of a given site to the closest active site, and hypothesized that DTL may be a suitable proxy for investigating functional constraints in a manner complementary to RSA, a proxy for investigating structural constraints. To test this, we investigated distributions of $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ as a function of DTL for each predicted structure by first predicting sites implicated in ligand binding using InteracDome [Kobren and Singh, 2019], and then calculating a DTL for each site, given the closest predicted ligand-binding site (Table 4.5).

The average per-site ns-polymorphism rate throughout the 1a.3.V core genome was

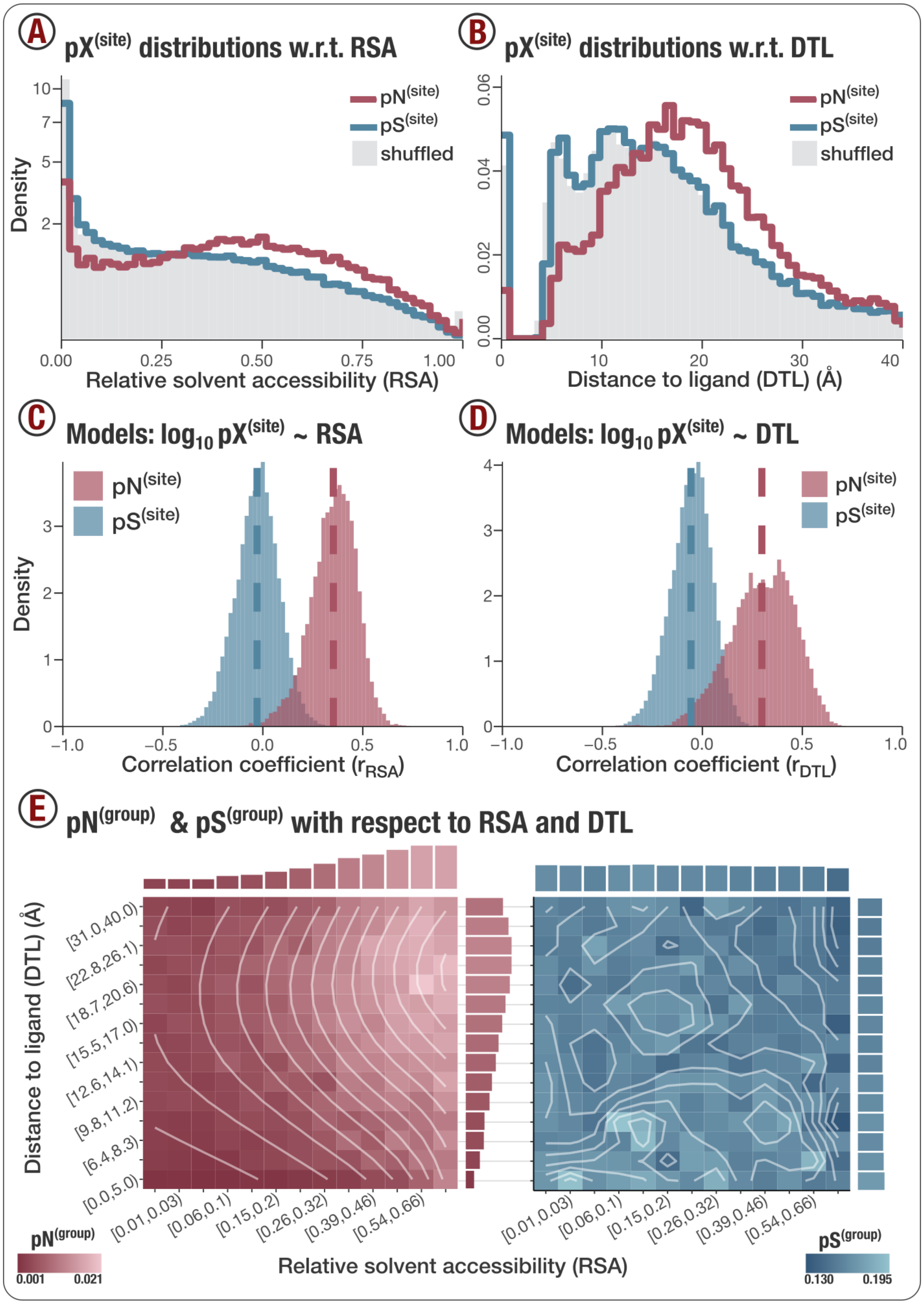


Figure 4.2: (A) Structural constraints shift the pN$^{(site)}$ distribution towards high relative solvent accessibility (RSA). The pN$^{(site)}$ distribution (red line) and pS$^{(site)}$ distribution (blue line) were created by weighting the RSA values of 239,528 sites (coming from the 754

Figure 4.2 continued: genes with predicted structures) by the $pN^{(site)}$ and $pS^{(site)}$ values observed in each of the 74 samples, totaling 17,725,072 $pN^{(site)}$ and $pS^{(site)}$ values. The average distribution of 10 independent, randomly shuffled datasets of $pN^{(site)}$ is depicted by the grey-regions for $pN^{(site)}$, and represents the null distribution expected if no association between $pN^{(site)}$ and RSA existed. Since the null distribution for $pS^{(site)}$ so closely resembles the null distribution for $pN^{(site)}$, it has been excluded for visual clarity, but can be seen in Figure 4.9. (B) Functional constraint shifts the $pN^{(site)}$ distribution towards high distance-to-ligand (DTL) values. The $pN^{(site)}$ distribution (red line) and $pS^{(site)}$ distribution (blue line) were created by weighting the DTL values of 155,478 sites (coming from 415 genes that had predicted structures and at least one predicted ligand) by the $pN^{(site)}$ and $pS^{(site)}$ values observed in each of the 74 samples, totaling 11,505,372 $pN^{(site)}$ and $pS^{(site)}$ values. The $pN^{(site)}$ null distribution was calculated according to the procedure described in panel A), where again, the $pS^{(site)}$ null distribution closely resembled the $pN^{(site)}$ null distribution, and can be seen in Figure 4.9. (C) Linear models reveal positive correlations between $pN^{(site)}$ and RSA. The two distributions show Pearson correlation coefficients produced by linear models of the form $\log_{10}(pN^{(site)}) \sim$ RSA (red-filled region) and $\log_{10}(pS^{(site)}) \sim$ RSA (blue-filled region). A model has been fit to each gene-sample pair that passed filtering criteria (see Supplementary Information), resulting in 16,285 nonsynonymous models and 24,553 synonymous models. Distribution means are visualized as dashed lines. (D) Per-group polymorphism rates explain the major selective pressure trends with respect to RSA and DTL. The left and right panels show heatmaps of $pN^{(group)}$ and $pS^{(group)}$. Each cell represents a group defined by RSA and DTL ranges shown on the x- and y- axes, respectively. The color of each cell represents the respective value for the group, where dark refers to low values and light refers to high values. White lines show the contour lines of smoothed data.

0.0088, however, we observed a nearly 4-fold reduction in this rate to just 0.0024 at predicted ligand binding sites (DTL = 0), indicating stronger purifying selection at ligand-binding sites (Figure 4.2b). Sites neighboring ligand-binding regions also harbored disproportionately low rates of ns-polymorphism, as indicated by the significant deviation towards larger DTL values. This illustrates that purifying selection that preserves proper ligand-binding functionality is not limited to residues at ligand-binding sites, but extends to proximal sites as well. When we defined DTL in sequence space rather than Euclidean space, this effect was no longer observable beyond sequence distances of ~5-10 amino acids (Figure 4.10). Comparatively, $pS^{(site)}$ deviated minimally from the null distribution. Overall, integrating predicted protein structures and ligand-binding sites into the analysis of the genetic diversity of an environmental population has enabled us to demonstrate that (1) structural constraints bias $\mathrm{pN}^{(\mathrm{site})}$ distributions towards solvent exposed sites (*i.e.* high RSA) (Figure 4.2a), and (2) functional constraints bias $\mathrm{pN}^{(\mathrm{site})}$ distributions towards sites that are distant from ligand-binding sites (*i.e.* high DTL) (Figure 4.2b).

4.5.4 Proteomic trends in purifying selection are explained by RSA and DTL

Given the clear shift in ns-polymorphism rates towards high RSA and DTL sites across genes, we next investigated the extent that RSA and DTL can predict per-site polymorphism rates. By fitting a series of linear models to log-transformed polymorphism data (Table S6), we conclude that RSA and DTL can explain 11.83% and 6.89% of $\mathrm{pN}^{(\mathrm{site})}$ variation, respectively. Based on these models we estimate that for any given gene in any given sample, (1) a 1% increase in RSA corresponds to a 0.98% increase in $\mathrm{pN}^{(\mathrm{site})}$, and (2) a 1% increase in DTL (normalized by the maximum DTL in the gene) corresponds to a 0.90% increase in $\mathrm{pN}^{(\mathrm{site})}$. In a combined model, RSA and DTL jointly explained 14.12% of $\mathrm{pN}^{(\mathrm{site})}$ variation, and after adjusting for gene-to-gene and sample-to-sample variance, 17.07% of the remaining variation could be explained by RSA and DTL. In comparison, only 0.35% of $\mathrm{pS}^{(\mathrm{site})}$ variation was explained by RSA and DTL. Using a complementary approach, we constructed models for each gene-sample pair (Supplementary Information), the correlations of which we used to visualize the extent that $\mathrm{pN}^{(\mathrm{site})}$ can be modeled by RSA and DTL relative to $\mathrm{pS}^{(\mathrm{site})}$ (Figures 4.2c, 4.2d). Analyzing gene-sample pairs revealed that the extent of ns-polymorphism rate that can be explained by RSA and DTL is not uniform across all genes (Table S7) and can reach up to 52.6% and 51.4%, respectively (Figures 4.11, 4.12). Finally, we averaged polymorphism rates within groups of sites that shared similar RSA and DTL values, which demonstrated the tight association between the rate of within population ns-polymorphism rate and protein structure (Table S8, Figure 4.2e). Linear regressions of these data show that 83.6% of per-group ns-polymorphism rates and 20.7% of per-group s-polymorphism rates are explained by RSA and DTL (Supplementary Information).

The true predictive power of RSA and DTL for polymorphism rates is most likely higher than we report, since our approaches suffer from methodological shortcomings. For instance, we calculate RSA from the steric configurations of residues in predicted structures. Thus, errors in structure prediction propagate to errors in RSA. Errors in structure also propagate to errors in DTL, since DTL is calculated using Euclidean distances between residues, which is exacerbated by the uncertainty associated with ligand-binding site predictions. Furthermore, RSA and DTL calculations assume that the protein is monomeric, even though oligomeric proteins are common, and they represent the majority of proteins in some organisms [Goodsell and and Olson, 2003]. In these cases, exposed sites in the monomeric structure could be buried once assembled into the quaternary structure, and this is similarly true for estimates of DTL. Even if we assume structural predictions are 100% accurate, it is notable that binding site predictions exclude (1) ligands that are proteins, (2) ligand-protein complexes that have not co-crystallized with each other, (3) ligands of proteins with no shared homology in the InteracDome database, and (4) unknown ligand-protein complexes. Each of these shortcomings leads to missed binding sites, which leads to erroneously high DTL values in the proximity of unidentified binding sites (Figure 4.13). Furthermore, our predictions assume that if a homologous protein in the InteracDome database binds to a ligand with a particular residue, then so too does the corresponding residue in the HIMB83 protein. This leads to uncertain predictions, since homology does not necessitate binding site conservancy. Additionally, studies have shown that conservation gradients are stronger for catalytic versus non-catalytic binding sites [Sharir-Ivry and Xia, 2019], yet we do not distinguish between these ligand classes. Finally, since we do not control for conformational changes induced by allostery, there are likely instances of sites under strong functional constraint that we have labeled as high DTL. Yet despite all these methodological shortcomings, our analyses show that RSA and DTL prevail as significant predictors of per-site and per-group variation.

Clear partitioning of environmental genetic variation by RSA and DTL (Figure 4.2) highlights the utility of these metrics for studies of evolution following the increasing availability of protein structures. Analyses of total genetic variation lacking the ability to delineate distinct processes of evolution limit opportunities to identify determinants of fitness in rich and complex data afforded by environmental metagenomes. Indeed, the application of RSA and DTL to SAR11 demonstrate that not all variants are created equal; a notion considered common knowledge by all life scientists, and yet such a treatment is lacking in studies of genomic heterogeneity that rely upon metagenomic read recruitment. RSA and DTL provide quantitative means to bring a level of scrutiny to distinguish variants based on their distributions in proteins. For instance, a collection of high-RSA and high-DTL sites will be more likely to be enriched in neutral variants. In contrast, residues under strong purifying selection will more likely be enriched in low-RSA and/or low-DTL sites of proteins. The ability to tease apart distinct evolutionary processes with absolute accuracy will indeed remain difficult due to a multitude of factors. But by providing structure-informed means to partition the total intra-population variation into distinct pools, RSA and DTL offer a quantitative framework that enables new opportunities to study distinct evolutionary processes.

4.5.5 Measuring purifying selection between genes and environments with $pN/pS^{(gene)}$

So far, our structure-informed investigation has focused on trends of sequence variation within the gene pool of an environmental population. Next, we shifted our attention to individual proteins. $\mathrm{pN/pS}^{(\mathrm{gene})}$ is a metric that quantifies the overall direction and magnitude of selection acting on a single gene [Schloissnig et al., 2013, Shenhav and Zeevi, 2020], where $\mathrm{pN/pS}^{(\mathrm{gene})} < 1$ indicates the presence of purifying selection, the intensity of which increases as the ratio decreases. Since $\mathrm{pN/pS}^{(\mathrm{gene})}$ is defined for a given gene in a given sample, $\mathrm{pN/pS}^{(\mathrm{gene})}$ values for a single gene can be compiled from multiple samples, enabling the tracking of selective pressures across environments [Shenhav and Zeevi, 2020]. Taking advantage of the large number of metagenomes in which 1a.3.V was present, we calculated $\mathrm{pN/pS}^{\mathrm{(gene)}}$ for all 799 protein-coding core genes across 74 samples (see Methods), resulting in 59,126 gene/sample pairs (Table 4.9). We validated our calculations by comparing sample-averaged $\mathrm{pN/pS}^{\mathrm{(gene)}}$ to $\mathrm{dN/dS}^{\mathrm{(gene)}}$ calculated from homologous gene pairs between HIMB83 and HIMB122, another SAR11 isolate genome that is closely related to HIMB83 (gANI: 82.6%), which we found to yield commensurable results (Figure 4.14, Table 4.12, Supplementary Information).

We found significantly more $\mathrm{pN/pS}^{\mathrm{(gene)}}$ variation between genes of a given sample ('gene-to-gene' variation) than between samples of a given gene ('sample-to-sample' variation) (ANOVA, Figure 4.15). All but one gene (gene #2031, unknown function) maintained $\mathrm{pN/pS}^{\mathrm{(gene)}} \ll 1$ in every sample, whereby 95% of values were less than 0.15 (Figure 4.16, Table 4.9), indicating an intense purifying selection for the vast majority of 1a.3.V genes across environments. This was foreshadowed by our earlier analysis in which $\mathrm{pS}^{\mathrm{(site)}}$ outweighed $\mathrm{pN}^{\mathrm{(site)}}$ by 19:1 within the aggregated data across genes and samples. However, the magnitude of purifying selection was not uniform across all genes. In fact, gene-to-gene variance, as opposed to sample-to-sample variance, explained 93% of $\mathrm{pN/pS}^{\mathrm{(gene)}}$ variation (ANOVA, Figure 4.15). By analyzing the companion metatranscriptomic data [Salazar et al., 2019] that were available for 50 of the 74 metagenomes, we were able to explain 29% of gene-to-gene variance with gene transcript abundance (Table 4.13, Supplementary Information), a known predictor of evolutionary rate [Pál et al., 2001]. Overall, these data demonstrate the utility of $\mathrm{pN/pS}^{\mathrm{(gene)}}$ as a metric to understand the overall extent of selection acting on genes.

The amount of $\mathrm{pN/pS}^{\mathrm{(gene)}}$ variation attributable to sample-to-sample variance was only 0.7% (Figure 4.15). While it represents a small proportion of the total variance, the sample-to-sample variance in $\mathrm{pN/pS}^{\mathrm{(gene)}}$ encapsulates the extent that polymorphism varies in response to the range of environmental parameters observed across samples. These data therefore provide the opportunity to relate how differences in genetic diversity of individual genes manifests from differences in environmental parameters (Table 4.10), which we focused on next.

4.5.6 Nitrogen availability governs rates of non-ideal polymorphism at critical sites of glutamine synthetase

To gain a more highly resolved picture of how selection shapes protein evolution, we searched for a biologically relevant gene within 1a.3.V that exhibited evolutionary patterns that could be understood by leveraging structural information. Glutamine synthetase (GS) is a critical enzyme for the recycling of cellular nitrogen [Bernard and Habash, 2009], a limiting nutrient for microbial productivity in surface oceans [Bristow et al., 2017]. GS yields glutamine and ADP from glutamate, ammonia, and ATP, an essential step in the biosynthesis of nitrogenous compounds.

Given the central role that GS plays in nitrogen metabolism, we expected GS to be under high selection. Indeed, the sample-averaged $\mathrm{pN/pS}^{(\mathrm{GS})}$ was 0.02, ranking GS amongst the top 11% most purified genes (Figure 4.3b, Table 4.9). Although highly purified, we observed significant sample-to-sample variation in $\mathrm{pN/pS}^{(\mathrm{GS})}$ (min = 0.010, max = 0.036) suggesting that the strength of purifying selection on GS varies from sample to sample (Figure 4.3b inset), perhaps due to unique environmental conditions (*e.g.*, nutrient compositions) that differentially impact the need for glutamine synthesis. Since previous work has shown that SAR11 upregulates its transcriptional and translational production of GS in response to nitrogen limitation [Smith Daniel P. et al.], we hypothesized that purifying selection should be highest in nitrogen-limited environments, and lowest in nitrogen-replete environments. We utilized measured concentrations of nitrate as an indication of the level of nitrogen limitation in each sample, and found a positive correlation between measured nitrate concentrations and pN/pS$^{(\mathrm{GS})}$ values across samples (Pearson correlation p-value = 0.009, $R^2 = 0.11$) (Figure 4.3c), which ranked amongst the top 12% of positive correlations between pN/pS$^{(\mathrm{gene})}$ and nitrate concentration (Figure 4.3c inset, Table 4.10). In summary, we find that although GS is under high selection, subtle differences in selection strength are observed between samples and are most likely driven by nitrogen availability.

Next, we focused on the GS protein structure to further investigate the associations between GS polymorphism and processes of selection. Since the native quaternary structure of GS is a dodecameric complex (12 monomers), our monomeric estimates of RSA and DTL are unrepresentative of the active state of GS. We addressed this by aligning 12 copies of the predicted structure to a solved dodecameric complex of GS in Salmonella typhimurium (PDB ID 1FPY), which HIMB83 GS shares 61% amino acid similarity with (Figure 4.3a). From this stitched quaternary structure we recalculated RSA and DTL, and as expected, this yielded lower average RSA and DTL estimates due to the presence of adjacent monomers (0.17 versus 0.24 for RSA and 17.8Å versus 21.2Å for DTL). With these quaternary estimates of RSA and DTL, we found that ns-polymorphism was 30x less common than s-polymorphism, and it strongly avoided sites with low RSA and the three glutamate active sites to which any given monomer was proximal (Figure 4.3d). In comparison, s-polymorphism distributed relatively homogeneously throughout the protein, whereby 17% of s-polymorphism occurred within 10Å of active sites (compared to 3% for ns-polymorphism) and 19% occurred in sites with 0 RSA (compared to 9% for ns-polymorphism). Averaged across samples, the mean RSA was 0.15 for s-polymorphism and 0.33 for ns-polymorphism (Figure 4.3e left panel). Similarly, the mean DTL was 17.2Å for s-polymorphism and 22.9Å for ns-polymorphism (Figure 4.3f left panel). These observations highlight in a single gene what we previously observed across the 1a.3.V core: selection purifies the majority of ns-polymorphism and does so with increased strength at structurally/functionally critical sites.

We next investigated whether variance in selection strength (Figure 4.3b inset) affects the spatial distribution patterns of polymorphism. For each sample, we calculated how polymorphism rates in GS distributed with respect to RSA and DTL and associated these distributions with pN/pS$^{(\mathrm{GS})}$. While the mean RSA of s-polymorphism remained relatively invariant (standard deviation 0.005) (Figure 4.3e right panel), the mean RSA of ns-polymorphism varied dramatically from 0.27 to 0.37 and was profoundly influenced by sample pN/pS$^{(\mathrm{GS})}$; samples exhibiting low selection of GS harbored lower mean RSA and samples exhibiting high selection of GS harbored higher mean RSA (Figure 4.3e right panel). In fact, 82.9% of mean RSA ns-polymorphism variance could be explained by pN/pS$^{(\mathrm{GS})}$ alone (Pearson correlation, p-value $< 1 \times 10^{-16}$, $R^2 = 0.829$). ns-polymorphism distributions with respect to DTL were equally governed by selection strength, where 80.4% of variance could be explained by pN/pS$^{(\mathrm{GS})}$ (Pearson correlation, p-value $< 1 \times 10^{-16}$, $R^2 = 0.804$, Figure 4.3f).

When selection is low, we observe high nitrate concentrations (Figure 4.3c inset) and ns-polymorphism distributions towards lower RSA/DTL (Figures 4.3e, 4.3f). When selection is high, we observe low environmental nitrate concentrations (Figure 4.3c inset) and ns-polymorphism distributions towards higher RSA/DTL (Figures 4.3e, 4.3f). Given that proper functionality of GS is most critical in nitrogen-limited environments and that mutations with low RSA/DTL are more likely to be deleterious, the most likely explanation for the body of evidence presented is that GS accumulates non-ideal polymorphism in samples exhibiting low selection of GS that cannot be effectively purified at the given selection strength. As selection increases, so too does the purifying efficiency, which we indirectly measure as increases in mean RSA and DTL of ns-polymorphism. Our approach illustrates this ‘use it or lose it’ evolutionary principle over a spectrum of selection strengths which have been sampled from natural *in situ* environmental conditions.

Under this hypothesis, there should exist low DTL amino acid alleles that create a negative, yet tolerable impact on fitness when selection is low, yet incur an increasingly detrimental fitness cost as selection increases. One would expect such alleles to be at low frequency

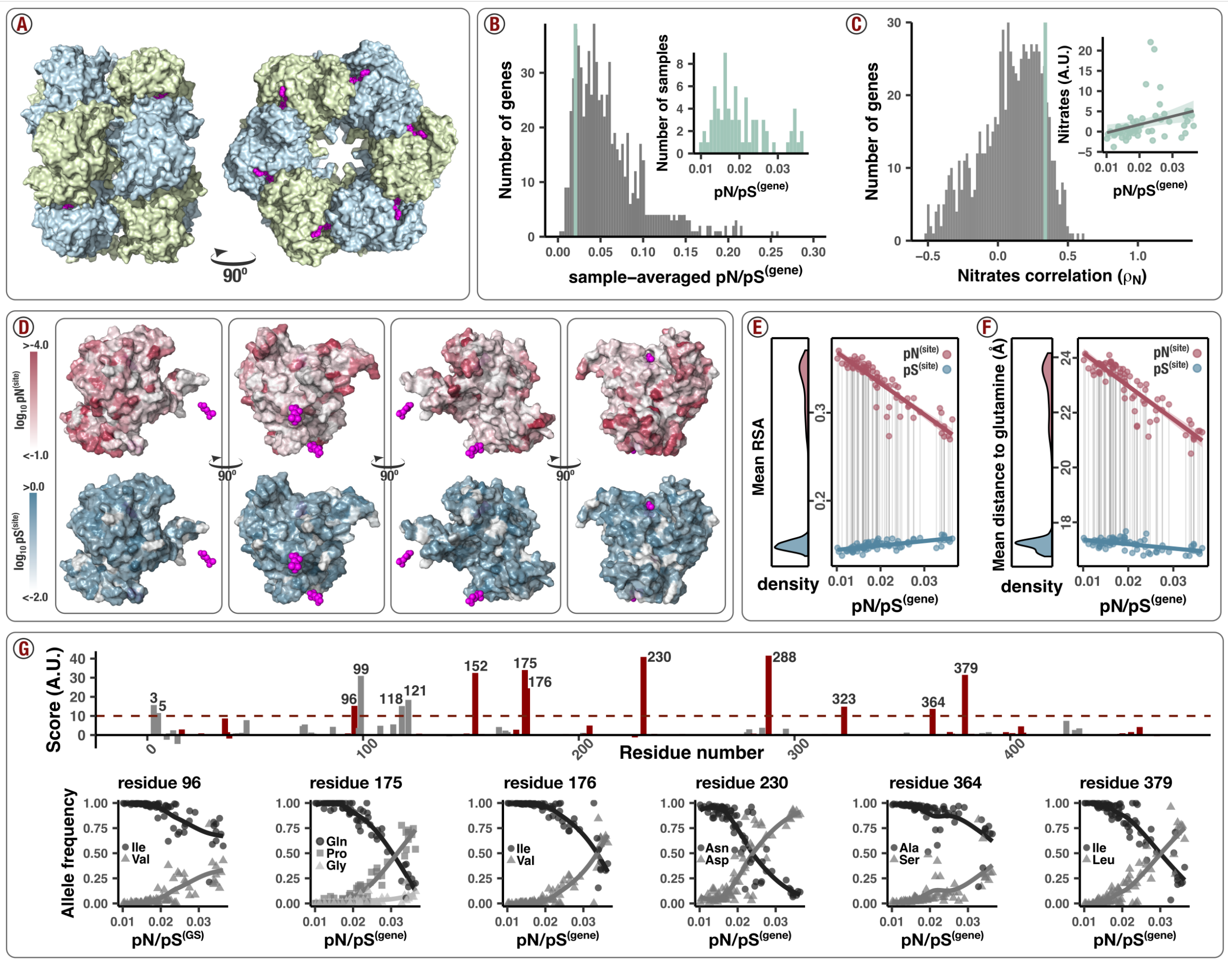


Figure 4.3: Polymorphism distribution patterns in glutamine synthetase (GS). (A) GS forms a dodecameric complex. The structure (PDB ID 1FPY) comes from Salmonella typhimurium (61% sequence similarity to HIMB83) and is shown from two different views. Pink molecules are ADP and phosphinothricin (steric inhibitor of glutamate), and are situated within the active site of GS. (B) GS is one of the most highly conserved genes in 1a.3.V. The main plot shows the distribution of sample-averaged pN/pS$^{(gene)}$ for all 799 genes in the 1a.3.V core (truncated at 0.30). The vertical green line depicts the sample-averaged pN/pS$^{(gene)}$ for GS (0.020). The inset plot shows the distribution of pN/pS$^{(gene)}$ value for GS as seen across the 74 samples, which vary from 0.010 to 0.036.

Figure 4.3 continued: (C) Selection strength on GS correlates with environmental concentration of nitrates. The main plot shows a histogram of Pearson correlation coefficients (one per gene) between $\mathrm{pN/pS}^{(\mathrm{gene})}$ and measured concentration of nitrates in each sample. The vertical green line depicts the correlation coefficient for GS (0.34). The inset shows a scatter plot of $\mathrm{pN/pS}^{(\mathrm{gene})}$ vs nitrate concentrations from which the GS correlation coefficient was calculated. (D) ns-polymorphism polymorphism rates are reduced in the vicinity of the active sites. Each image is a view of the predicted structure of monomeric GS. Phosphinothricin substrates were situated by aligning the predicted GS structure to the complex in panel A. Red surfaces are colored according to the sample-averaged $\log_{10}\mathrm{pN}^{(\mathrm{site})}$ value of each residue, and blue surfaces are colored according to the sample-averaged $\log_{10}\mathrm{pS}^{(\mathrm{site})}$ value of each residue. In each case, darker colors refer to higher rates. Left-to-right, each view is a 90° clockwise rotation of the previous view about the vertical axis. Each image was rendered programmatically using a PyMOL script that was generated from the anvi'o structure interactive interface. (E) As selection decreases, ns-polymorphism creeps into low-RSA sites. The left panel shows the distribution of samples' average RSA of nonsynonymous (red) and synonymous (blue) polymorphisms. The right panel shows how these average RSA values (y-axis) correlate with the samples' $\mathrm{pN/pS}^{(\mathrm{gene})}$ values (x-axis). Each data point is calculated by weighting the RSA of each residue by the $\mathrm{pN}^{(\mathrm{site})}$ (red) or $\mathrm{pS}^{(\mathrm{site})}$ (blue) values observed in that sample. The red and blue lines show the nonsynonymous and synonymous linear fits, respectively, and the corresponding shaded regions show the 95% confidence intervals for the fit. (F) As selection decreases, ns-polymorphism creeps closer to the binding site. The scheme is identical to panel E, where RSA is replaced with the distance-to-glutamate substrate (DTL). (G) Some sites exhibit amino acid minor allele frequencies that co-vary with $\mathrm{pN/pS}^{(\mathrm{GS})}$. The top panel shows the extent that sites co-vary with $\mathrm{pN/pS}^{(\mathrm{GS})}$. The x-axis shows the residue number and the y-axis the slope estimate of a linear regression between the sum of minor allele frequencies and $\mathrm{pN/pS}^{(\mathrm{GS})}$. Sites with DTL values less than the average are indicated in red and are gray otherwise. All sites above an arbitrary cutoff (dashed horizontal line) are annotated with their residue number. Scatter plots below show the allele frequency trajectories for a select number of these sites.

in low $\mathrm{pN/pS}^{(\mathrm{GS})}$ samples, and to reach increasingly higher frequencies in higher $\mathrm{pN/pS}^{(\mathrm{GS})}$ samples. We identified putative sites fitting this description by scoring sites based on the extent that their amino acid minor allele frequencies co-varied with $\mathrm{pN/pS}^{(\mathrm{GS})}$, including only sites with DTL less than the mean DTL of ns-polymorphisms (22.9Å). Using an arbitrary cutoff, we identified 9 top-scoring polymorphisms that co-varied with $\mathrm{pN/pS}^{(\mathrm{GS})}$ (Figure 4.3g): I96V, L152I, Q175P/G, I176V, N230D, S288A/D, I323V, A364S, I379L. Though each of these sites exhibited DTL lower than the average ns-polymorphism, the closest site (residue number 323) was still 9Å away from the glutamate substrate. This suggests there are no 'smoking gun' polymorphisms occurring in the binding site that abrasively disrupt functionality. After all, in absolute terms GS is highly purified regardless of sample – the largest pN/pS$^{(\mathrm{GS})}$ is 0.036, which is just over half the genome-wide average pN/pS$^{(\mathrm{gene})}$ of 0.063. Our data therefore represents a subtle, yet resolvable signal of minute decreases in selection strength manifesting as minute shifts in the distribution of ns-polymorphism towards the active site.

While identifying signatures of positive selection is typically the primary pursuit in evolutionary analysis, our data instead illustrates a highly resolved interplay between purifying selection strength and polymorphism distribution. The geography and unique environmental parameters associated with each sample yielded a spectrum of selection strengths which enabled us to quantify how polymorphism distributions of a gene under high selection shift in response to small perturbations in selection strength. In the case of GS, we were able to attribute these shifts to the availability of nitrogen, thereby linking together environment, selection, and polymorphism.

Throughout the 1a.3.V core genes, we observed that samples exhibiting low overall selection of 1a.3.V were strongly associated with increased accumulation of ns-polymorphism at low RSA/DTL sites (Figures 4.4a, 4.4b, Supplementary Information), suggesting this signal is not specific to GS, but rather a general feature of the 1a.3.V core genes. Though highly significant (one sided Pearson p-values 9×10^{-12} for RSA and 2×10^{-4} for DTL), the magnitude that ns-polymorphism distributions shift with respect to DTL and RSA were subtle: across samples, the mean DTL of ns-polymorphism varied by less than 1Å, and the mean RSA varied between 0.230 and 0.236. Resolving such a minute signal with such robust statistical power is owed to the immense quantities of sequence data afforded by metagenomics.

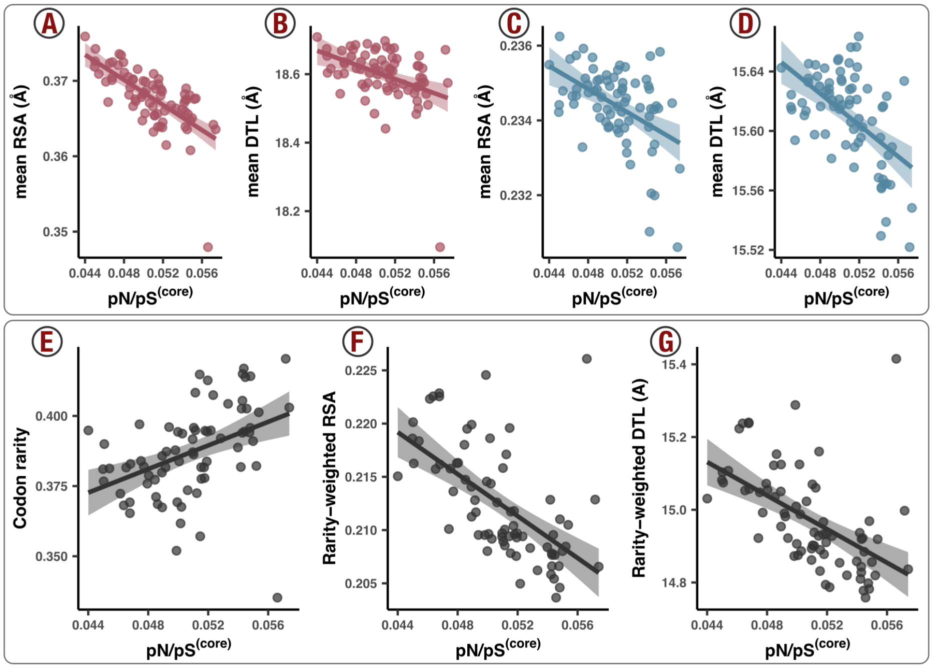


Figure 4.4: Polymorphism distribution patterns with respect to genome-wide selection strength. Each data point is a sample (metagenome). Lines represent lines of best fit and corresponding translucent areas represent 95% confidence intervals. The x-axis is $\text{pN/pS}^{(\text{core})}$, which is calculated across the whole core genome and is an inverse proxy of genome-wide purifying selection strength (see Methods). (A) The ns-polymorphism distribution mean with respect to RSA is negatively associated with $\text{pN/pS}^{(\text{core})}$ (one-sided Pearson p-value $= 9\times10^{-12}$). (B) The ns-polymorphism distribution mean with respect to DTL is negatively associated with $\text{pN/pS}^{(\text{core})}$ (one-sided Pearson p-value $= 2\times10^{-4}$). (C) The s-polymorphism distribution mean with respect to RSA is negatively associated with $\text{pN/pS}^{(\text{core})}$ (one-sided Pearson p-value $= 1\times10^{-5}$). (D) The s-polymorphism distribution mean with respect to RSA is negatively associated with $\text{pN/pS}^{(\text{core})}$ (one-sided Pearson p-value $= 3\times10^{-7}$). (E) Rare synonymous codons are more abundant in samples with high $\text{pN/pS}^{(\text{core})}$ (one-sided Pearson p-value $= 4\times10^{-5}$). (F) Rare synonymous codons avoid low RSA sites when $\text{pN/pS}^{(\text{core})}$ is low (one-sided Pearson p-value $= 1\times10^{-10}$). (G) Rare synonymous codons avoid low DTL sites when $\text{pN/pS}^{(\text{core})}$ is low (one-sided Pearson p-value $= 7\times10^{-9}$).

4.5.7 Synonymous but not silent: selection against rare codons at critical sites

Thus far we have observed that purifying efficiency observably decreases in response to lowered selection strength, as evidenced by ns-polymorphism occurring nearer to binding sites and in more buried sites. Given the influence of synonymous substitutions in translational processes [Plotkin and Kudla, 2011], as a final analysis we focused on within-population trends of s-polymorphism.

Compared to ns-polymorphism, s-polymorphism distributes more uniformly throughout protein structures (Figures 4.2a, 4.2b). Yet our data also revealed an association between selection strength and the distribution of s-polymorphism. In samples under higher selection, s-polymorphism systematically tended to occur (1) in more solvent-exposed sites (Figure 4.4c, one-sided Pearson p-value $= 1 \times 10^{-5}$) and (2) farther from binding sites (Figure 4.4d, one-sided Pearson p-value $= 3 \times 10^{-7}$). These trends indeed mimic the nonsynonymous trends in glutamine synthetase (Figures 4.3d, 4.3e, 4.3f) as well as the core genes in general (Figures 4.4a, 4.4b), and cannot be reasonably explained by neutral processes. The surprising association suggests a relationship between selection and synonymous change that is at least partly determined by structural features of proteins.

With a GC-content lower than 30%, SAR11 genomes maintain a non-uniform yet conserved codon composition (Figure 4.17). Previous work has shown that rare codons can significantly reduce translation rates [Sørensen et al., 1989], cause delays in the production of the polypeptide chain at the ribosome [Komar, 2009], which can lead to protein misfolding [Drummond and Wilke, 2008, Agashe et al., 2013], and impair fitness [Walsh et al., 2020]. Thus, we hypothesized that rare codons in 1a.3.V may incur fitness costs relative to their more common, synonymous counterparts. To test this hypothesis, we investigated the relationship between selection strength and the occurrence of rare codons, which required us to define a 'codon rarity' metric based on the frequency that codons are found in the HIMB83 genome relative to their synonymous counterparts (Table 4.11). We then attributed an overall rarity score to each sample by weighting the rarity of all synonymous codon alleles by the frequencies with which they were observed (see Methods). Our analysis of these data revealed a positive correlation between codon rarity in a sample and its pN/pS$^{(\mathrm{core})}$ (Figure 4.4d, one-sided Pearson p-value $= 1 \times 10^{-5}$), illustrating that rare codons are more likely to be found in samples where genome-wide selection is low. We found this to be the case for s-polymorphism within all 18 amino acids that possess two or more codons (Figure 4.18), illustrating that this evolutionary process acts ubiquitously throughout the genetic code of 1a.3.V. Rare codons did not distribute throughout protein structures uniformly, either. In samples with low genome-wide selection, where rarity was highest, rare codons occurred farther away from binding sites (one-sided Pearson p-value $= 1 \times 10^{-10}$) and occurred more frequently in more solvent-exposed sites (one-sided Pearson p-value $= 7 \times 10^{-9}$), as compared to low selection samples (Figures 4.4e, 4.4f).

Overall, these data show that when genome-wide selection strength is low, rare codons both (1) incorporate into the genome with increased propensity, and (2) manifest in sites that are statistically more likely to be structurally/functionally important. As previous research suggests, the most likely explanation for these observations is that rare codons are less fit due to decreased translational accuracy compared to their more common, synonymous counterparts. Yet the environmental and structural dimensions of our data reveal the dynamic nature of the evolutionary processes that maintain synonymous polymorphism as a function of changing conditions in naturally occurring habitats and elucidates the intensity of such processes as a function of their physical locations in the structure. Indeed, 1a.3.V maintains the lowest proportion of rare codons in samples where genome-wide selection is highest, and rare codons in these samples are statistically more likely to be incorporated in noncritical sites of proteins, most likely due to the increased efficiency with which purifying selection operates in an environment- and site-dependent manner. These rare codon data provide a lens into the potential fitness costs associated with suboptimal translational accuracy in complex populations, and by including structural data, we demonstrate where optimal translational accuracy matters most.

4.6 Conclusions

With recent breakthroughs in predicting protein structures and ligand binding sites, microbial ecology need not be limited to just sequences. By offering an interactive, scalable, and open-source software solution that integrates environmental genetic variants with structural bioinformatics, our study takes advantage of recent advances to connect environmental 'omics and structural biology. Indeed, by leveraging structure and ligand-binding predictions we were able to describe striking patterns of nucleotide polymorphism in an environmental microbial population that we could ascribe to evolutionary constraints that preserve protein structure (folding & stability) and protein function (ligand-binding activity). By tracking a SAR11 population across metagenomes we were able to demonstrate the presence of dynamic processes that purge both synonymous and nonsynonymous polymorphism from the vicinity of ligand binding sites of proteins as a function of selection strength. Overall, our study proposes a structure-informed computational framework for microbial population genetics and offers a glimpse into the emerging interdisciplinary opportunities made available at the intersection of ecology, evolution, and structural biology.

4.7 Methods

4.7.1 Overview

The URL `https://merenlab.org/data/anvio-structure` provides a complete reproducible workflow for all analysis steps detailed below, including (1) downloading the publicly available metagenomes and genomes, (2) recruiting reads from metagenomes, (3) calculating single amino-acid and single codon variants, (4) predicting protein structures and ligand binding sites, and (5) visualizing metagenomic sequence variants and binding sites onto protein structures.

4.7.2 Metagenomic and metatranscriptomic read recruitment and processing

To study the population structure of the environmental SAR11 population 1a.3.V defined previously [Delmont et al., 2019], we used anvi'o v7.1 [Eren et al., 2021], and its metagenomics workflow [Shaiber et al., 2020] which uses snakemake v5.10 [Köster and Rahmann, 2012] to automate gene calling, gene function annotation, metagenomic and metatranscriptomic read recruitment steps. The compendium of anvi'o programs the metagenomics workflow called upon employed Prodigal v2.6.3 [Hyatt et al., 2010] for gene calling, NCBI's Clusters of Orthologous Groups (COGs) database [Tatusov et al., 2003] and Pfams [El-Gebali et al., 2019] for gene function annotation, HMMER v3.3 [Eddy, 2011] for profile HMM searches, DIAMOND v2.0.6 [Buchfink et al., 2015] for sequence searches, Bowtie2 v2.4 [Langmead and Salzberg, 2012] for read recruitment, and samtools v1.9 [Li et al., 2009] to generate BAM files. The metagenomic workflow resulted in a 'contigs database' and a 'merged profile database' (two anvi'o artifacts detailed at https://anvio.org/help/), which gives access to gene and genome coverages (with metagenomic or metatranscriptomic short reads), as well as the sequence variability data to study population genetics as detailed below. We adopted a competitive read recruitment strategy by using all SAR11 genomes, rather than only HIMB83, as reference to recruit reads from Tara Oceans Project metagenomes and metatranscriptomes to maximize the exclusion of reads that matched better to other known SAR11 genomes, thereby narrowing our scope of probed diversity and minimizing the impacts of non-specific read recruitment. In all subsequent analyses we focused on the core genes of the 1a.3.V subclade by only considering (a) reads that mapped to HIMB83 (b) the

74 metagenomes in which HIMB83 was found above 50X, and (c) the 799 HIMB83 genes that were previously found to maintain consistent coverage patterns [Delmont et al., 2019].

4.7.3 Quantifying SCVs and SAAVs in metagenomes

To characterize the variants in metagenomic read recruitment results we used and extended the microbial population genetics framework implemented in anvi'o. The program `anvi-profile` with the flag `--profile-SCVs` characterizes single codon variants (SCVs), from which single amino acid variants (SAAVs) can also be calculated. Anvi'o determines allele frequency vectors for SCVs by tallying the frequencies of codons observed in the 3-nt segments of reads that fully map to a given codon position. The frequencies of amino acids encoded by each 3-nt segment yield SAAVs observed in a given position, which represent allele frequency vectors of positions after collapsing synonymous redundancy. For a given codon position, anvi'o excludes any reads that do not map to all 3 nucleotides, which can happen either if the read terminates within the codon position, or there exists a deletion in the read relative to the reference genome. Reads that contain insertions within the codon relative to the reference genome are also excluded during this step. We exported variant profiles as tabular data using the program `anvi-gen-variability-profile`, where each row is a SCV (or SAAV) and the columns specify (1) identifying information such as the corresponding gene, codon position, and sample id, (2) the number of mapped reads corresponding to each of the 64 codons (or 20 amino acids), and (3) numerous miscellaneous statistics, all of which can be explored at `https://merenlab.org/analyzing-genetic-varaibility`.

4.7.4 Calculations of polymorphism rates of individual codon sites, $pN^{(site)}$ and $pS^{(site)}$

We calculated the polymorphism rates of individual codon sites from allele frequencies defined from each SCV based on a recent study by Shenhav and Zeevi [Shenhav and Zeevi, 2020], where a given codon allele contributes (to either $\mathrm{pN}^{(\mathrm{site})}$ or $\mathrm{pS}^{(\mathrm{site})}$) an amount that is equal to its observed relative abundance (frequency). To which rate the allele contributes is determined by its synonymity relative to the popular consensus, *i.e.* the allele most common across all samples. After summing the contributions for each of the 63 codons (excluding the popular consensus), we normalized the resulting values of $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ by the number of nonsynonymous and synonymous sites of the popular consensus, respectively. For example, if the popular consensus is 'ACC' (Thr), there are 9 possible single point mutations, 3 synonymous and 6 nonsynonymous, therefore $\mathrm{pS}^{(\mathrm{site})}$ will be divided by $3/3 = 1$ and $\mathrm{pN}^{(\mathrm{site})}$ will be divided by $6/3 = 2$. This procedure can be mathematically expressed as

$$\mathrm{pN}^{(\mathrm{site})} = \frac{1}{n_n} \sum_{c \in C \backslash r} f_c \, N(c, r), \quad \mathrm{pS}^{(\mathrm{site})} = \frac{1}{n_s} \sum_{c \in C \backslash r} f_c \, S(c, r)$$

where $C \backslash r$ is the set of all codons excluding the popular consensus r; n_n and n_s are the number of nonsynonymous and synonymous sites of r, respectively; f_c is the frequency of the cth allele; $N(c, r)$ is the indicator function where,

$$N(c, r) = 1 \text{ if not synonymous}(c, r) \text{ else } 0$$

and $S(c, r)$ is the indicator function where,

$$S(c, r) = 1 \text{ if synonymous}(c, r) \text{ else } 0.$$

We implemented this strategy into the program `anvi-gen-variability-profile` as a new flag `--include-site-pnps`, which when declared, adds $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ values as additional columns to the tabular output after calculating them for 3 different choices of the reference codon r: (1) the popular consensus (as used in this paper), (2) the consensus (the allele with the highest frequency), and (3) the codon found in the reference sequence (the sequence used for read recruitment). For efficient computation, this calculation uses the Python package numba [Lam et al., 2015] for just-in-time compilation. For a dataset with 12,583,626 SCVs, the current implementation computes $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ terms in less than a minute on a laptop computer.

4.7.5 Calculations of polymorphism rates within a group of sites, $pN^{(group)}$, $pS^{(group)}$, and $pN/pS^{(group)}$

We defined groups such that all sites in a group share similar RSA and DTL values. Formally, we defined $\mathrm{pN}^{(\mathrm{group})}$ and $\mathrm{pS}^{(\mathrm{group})}$ as

$$\mathrm{pN}^{(\mathrm{group})} = \frac{\sum_{g=1}^{G} \sum_{c \in C \setminus r} f_c^{(g)} \, N(c, r^{(g)})}{\sum_{g=1}^{G} n_n^{(g)}}, \quad \mathrm{pS}^{(\mathrm{group})} = \frac{\sum_{g=1}^{G} \sum_{c \in C \setminus r} f_c^{(g)} \, S(c, r^{(g)})}{\sum_{g=1}^{G} n_s^{(g)}}.$$

G is the number of sites in the group; $r^{(g)}$ is the popular consensus of the gth site; $f_c^{(g)}$ is the frequency of the cth allele at the gth site; $n_n^{(g)}$ and $n_s^{(g)}$ are the number of nonsynonymous and synonymous sites of r, respectively. All other definitions are the same as for $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$. $\mathrm{pN}^{(\mathrm{group})}$ and $\mathrm{pS}^{(\mathrm{group})}$ can be expressed in terms of weighted sums of $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$, respectively:

$$\mathrm{pN}^{(\mathrm{group})} = \frac{\sum_{g=1}^{G} n_n^{(g)} \, \mathrm{pN}^{(\mathrm{g,\ site})}}{\sum_{g=1}^{G} n_n^{(g)}}, \quad \mathrm{pS}^{(\mathrm{group})} = \frac{\sum_{g=1}^{G} n_s^{(g)} \, \mathrm{pS}^{(\mathrm{g,\ site})}}{\sum_{g=1}^{G} n_s^{(g)}}.$$

Finally, $\mathrm{pN/pS}^{(\mathrm{group})}$ is defined as

$$\mathrm{pN/pS}^{(\mathrm{group})} = \mathrm{pN}^{(\mathrm{group})} \,/\, \mathrm{pS}^{(\mathrm{group})}.$$

4.7.6 Calculations of polymorphism rates for individual and core genes, $pN^{(gene)}$, $pS^{(gene)}$, $pN/pS^{(gene)}$, and $pN/pS^{(core)}$

We calculated rates of polymorphism for genes and the 1a.3.V core genome identically to the calculations of $\mathrm{pN}^{(\mathrm{group})}$, $\mathrm{pS}^{(\mathrm{group})}$, and $\mathrm{pN/pS}^{(\mathrm{group})}$. For example, $\mathrm{pN}^{(\mathrm{gene})}$ refers to the ns-polymorphism rate of all sites in a given gene, and $\mathrm{pS}^{(\mathrm{core})}$ refers to the s-polymorphism rate of all sites in the 1a.3.V core genome.

4.7.7 Predicting and processing protein structures

We attempted to predict protein structures for each gene in the HIMB83 genome that belonged to the 1a.3.V core using both AlphaFold [Jumper et al., 2021] and MODELLER [Webb and Sali, 2016]. To process, store, and access the resulting protein structures we developed a novel program, `anvi-gen-structure-database`, which gives access to all atomic coordinates as well as per-residue statistics such as relative solvent accessibility, secondary structure, and phi & psi angles calculated using DSSP (Touw et al., 2015; Kabsch and Sander, 1983). For AlphaFold predictions we used a version of the codebase that closely resembles v2.0.1 (the URL `https://github.com/johnaparker/alphafold/tree/3829f4e0ba01aa1b4f01916c83e9ca5de771d98a` gives access to its exact state) and ran predictions using 6 GPUs, which took a week on a high-performance computing system. AlphaFold predicted structures for 795 of 799 proteins, and after removing structures with gene-averaged pLDDT scores <80, we were left with 754 structures we deemed 'trustworthy' for downstream analyses. To predict protein structures with MODELLER, we developed a pipeline that, for each gene, (1) searches the Research Collaboratory for Structural Bioinformatics Protein Data Bank [Berman et al., 2000] (RSCB PDB) for homologs using DIAMOND [Buchfink et al., 2015], then downloads tertiary structures for matching entries, and (2) uses these homologs as templates to predict the gene's structure with MODELLER [Webb and Sali, 2016]. We discarded any proteins if the best template had a percent similarity of <30%. Unlike more sophisticated homology approaches that make use of multi-domain templates [Källberg et al., 2012], we used single-domain templates which are convenient and are accurate up to several angstroms, yet can lead to physically inaccurate models when the templates' domains match to some, but not all of the sequences' domains. To avoid this, we discarded any templates if the alignment coverage of the protein sequence to the template was <80%. Applying these filters resulted in 408 structures from the 1a.3.V core, which was further refined by requiring that the root mean squared distance (RMSD) between the predicted structure and the most similar template did not exceed 7.5 Å, and that the GA341 model score exceeded 0.95. After applying these constraints, we were left with 348 structures in the 1a.3.V that we assumed to be 'trustworthy' structures as predicted by MODELLER. These structures were on average 44.8% identical to their templates, which is within the sequence similarity regime where template-based homology modeling generally produces the correct overall fold [Rost, 1999].

4.7.8 Predicting ligand-binding sites

For the 1a.3.V core genes we estimated per-residue binding frequencies for a diverse collection of ligands by using InteracDome, a database that annotates the sites (match states) of Pfam profile hidden Markov models (HMMs) with ligand binding frequencies predicted from experimentally-determined structural data [Kobren and Singh, 2019]. To associate match state binding frequencies of the profile HMMs to the sites of HIMB83 genes, we applied a protocol similar to that described in Kobren & Singh.

First, we downloaded the Representable-NR Interactions (RNRI) from the InteracDome web server (`https://interacdome.princeton.edu/`) that "correspond to domain-ligand interactions that had nonredundant instances across three or more distinct PDB structures" (Table S5). Next, we downloaded the profile HMMs for Pfam v31.0 and kept only those 2,375 profiles that belonged to the RNRI dataset. Then, we searched each HIMB83 gene against this set using HMMER's hmmsearch. After the removal of HMM hits that were below the gathering threshold (GA) noise cutoffs defined in Pfam models, 940 of the 1,470 HIMB83 coding genes had at least one domain hit, with a total of 1,770 domain hits from 832 unique profile HMMs. Of these, we removed 177 for being too partial (length of the hit divided by the profile HMM length was less than 0.5), and 1 hit because the query sequence did not match all the consensus residues for match states in which the information content exceeded 4 (Table S5). We then associated binding frequencies for a collection of ligand types to the HIMB83 genes by parsing alignments of the profile HMMs to the HIMB83 gene amino acid sequences, which are provided in the standard output of hmmsearch. If a given HIMB83 residue aligned to multiple match states, each which had the same ligand type, we attributed the average binding frequency to the HIMB83 residue. We then filtered out binding frequency scores less than 0.5, yielding 40,219 predicted ligand-residue interactions across 11,480 unique sites (Table S5). We considered each of these sites to be 'ligand-binding sites'.

Our study includes two novel programs to automate this procedure and make it accessible to the community. The first, `anvi-setup-interacdome`, downloads the RNRI and Pfam datasets, and only needs to be run once. The second, `anvi-run-interacdome`, is a multi-threaded program that takes an anvi'o contigs database as input, and runs the remainder of the workflow described for each gene in the database. Predicted binding frequencies are stored internally in the database, which enables a seamless integration with other anvi'o programs to accomplish various tasks, such as the interactive visualization of the binding sites of predicted structures for any given gene with `anvi-display-structure` (see Supplementary Information), or exporting the underlying data as TAB-delimited files with `anvi-export-misc-data`. In the present study, `anvi-run-interacdome` processed the HIMB83 genome in 53 seconds on a laptop computer using a single thread.

4.7.9 Calculating relative solvent accessibility (RSA)

We calculated RSA for each residue of each predicted structure, where RSA was defined as the accessible surface area (ASA) probed by a 1.4Å radius sphere, divided by the maximum ASA, *i.e.* the ASA of a Gly-X-Gly tripeptide. RSA values were calculated in the program `anvi-gen-structure-database` using Biopython's DSSP module [Cock et al., 2009].

4.7.10 Calculating distance-to-ligand (DTL)

DTL was calculated for all sites that belonged to genes with (a) a predicted structure and (b) at least one predicted ligand-binding residue. Ideally, one would calculate DTL as the Euclidean distance of a residue to the predicted ligand, however our predictions did not yield the 3D coordinates of ligands. Instead, we approximated DTL as the Euclidean distance of a residue to the closest ligand-binding residue (see Methods), which lies within a few angstroms of the predicted ligand. Specifically, we defined this distance according to the sites' side chain center of masses. A consequence of approximating DTL with respect to the closest ligand-binding sites is that by definition, any ligand-binding residue has a DTL of 0.

As discussed in *Proteomic trends in purifying selection are explained by RSA and DTL*, missed binding sites lead to erroneously high DTL values. We assessed the magnitude of this error source by comparing our distribution of predicted DTL values in the 1a.3.V core to that found in BioLiP, an extensive database of semi-manually curated ligand-protein complexes [Yang et al., 2013]. We found the 1a.3.V DTL distribution had a much higher proportion of values >40 Å, suggesting these likely result from incomplete characterization of binding sites (Figure 4.13). To mitigate the influence of this inevitable error source, we conservatively excluded DTL values >40 Å (8.0% of sites) in all analyses after Figure 4.2b.

4.7.11 Calculating polymorphism null distributions for RSA and DTL

The null distributions for polymorphism rates with respect to RSA and DTL were calculated by randomly shuffling the RSA and DTL values calculated for each site, yielding distributions one would expect if there was no association between polymorphism rate and RSA. To avoid biases, each null distribution is the average of 10 shuffled datasets.

4.7.12 Proportion of polymorphism rate variance explained by RSA and DTL

To calculate the extent that RSA and DTL can explain polymorphism rates, we constructed 3 synonymous models (s-models) and 3 nonsynonymous models (ns-models) (Table 4.6). s-models fit linear regressions of $\log_{10}(\mathrm{pS}^{(\mathrm{site})})$ to RSA (s #1), DTL (s #2), and both RSA & DTL (s #3). Similarly, ns-models fit linear regressions of $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ to RSA (ns #1), DTL (ns #2), and both RSA & DTL (ns #3). Additionally, each model included the gene and sample of the corresponding polymorphism as independent variables, in order to account for gene-to-gene and sample-to-sample differences. Polymorphism rates were log-transformed because it helped linearize the data, yielding better models. The data used to fit each model included all codon positions across all samples in each gene that had a predicted protein structure and at least 1 predicted ligand-binding residue. After excluding monomorphic sites ($\mathrm{pN}^{(\mathrm{site})} = 0$ for ns-models, $\mathrm{pS}^{(\mathrm{site})} = 0$ for s-models), this yielded 5,838,445 data points for s-models and 3,850,182 for ns-models. While every protein has RSA values that span the domain [0,1], protein size creates dramatic gene-to-gene differences in observed DTL values. We accounted for this by standardizing DTL values on a per-gene basis, which improved variance explained by DTL. The variance explained by RSA, DTL, sample, and gene was determined by performing an ANOVA on each model and partitioning the sum of squares (Table S6).

4.7.13 Calculating transcript abundance (TA)

Since proper transcription level metrics such as molecules per cell are incalculable from metatranscriptomic data, we estimated the transcript abundance (TA) to be

$$\text{TA} = \frac{C^{(MT)}}{D^{(MT)}} \Big/ \frac{C^{(MG)}}{D^{(MG)}},$$

where $C^{(MT)}$ is the coverage of the gene in the metatranscriptome, $D^{(MT)}$ is the sequencing depth (total number of reads) of the metatranscriptome, $C^{(MG)}$ is the coverage of the gene in the metagenome, and $D^{(MG)}$ is the sequencing depth (total number of reads) of the metagenome. This means, for example, that a gene with a metatranscriptomic relative abundance 10% of its metagenomic relative abundance would have a TA of 0.10.

4.7.14 Definition of codon rarity

Our definition of codon rarity quantifies how rare a codon is compared to the codons it is synonymous with. Let f_c be the frequency that codon c is observed in the HIMB83 genome sequence, *i.e.* the proportion of HIMB83 codons corresponding to the codon c. Then, the rarity of codon i is equal to

$$R_i = 1 - \frac{f_i}{\sum_j f_j \, S(i,j)},$$

where $\sum$ is a sum over all 64 codons and $S(i,j)$ is the indicator function describing whether codons i and j are synonymous with one another:

$$S(i,j) = 1 \text{ if synonymous}(i,j) \text{ else } 0.$$

We utilized this definition to calculate the codon rarity of synonymous polymorphic sites ($\text{pN}^{(\text{site})} < 0.0005$) by weighting each codon's rarity by the frequency that the codon was observed in the short reads mapping to that position. For example, a polymorphic site with a coverage of 200, where 50 reads resolve to GCC ($R_{GCC} = 0.94$) and 150 resolve to GCT ($R_{GCT} = 0.58$) would get a rarity score of $50/200 \times 0.94 + 150/200 \times 0.58 = 0.67$. Extending this to multiple sites, we take the codon rarity of an entire sample to be the average rarity across all codon sites, excluding those with $\text{pN}^{(\text{site})} > 0.0005$.

4.7.15 Statistical data analysis and visualization

We used R v3.5.1 [R Development Core Team, 2011] for the analysis of numerical data reported from anvi'o. For data visualization we used ggplot2 [Ginestet, 2011b] library in R and anvi'o, and finalized images for publication using Inkscape v1.1 (https://inkscape.org/).

4.8 Acknowledgements

We thank Stephen Giovannoni (https://microbiology.oregonstate.edu/dr-stephen-giovannoni), Shilpa Nadimpalli Kobren (http://shilpakobren.com/), Michael K. Yu (@michaelkuyu), Tobin Sosnick (http://sosnick.uchicago.edu/), Craig DeValk (http://ranganathanlab.org/people_v2/), as well as the members of our laboratory (https://merenlab.org/people/) for helpful discussions. EK acknowledges support from the Natural Sciences and Engineering Research Council of Canada. This work was supported by Alfred P. Sloan Foundation Fellowship in Ocean Sciences to AME, and by a Simons Foundation grant (#687269) to AME.

4.9 Supplementary Figures

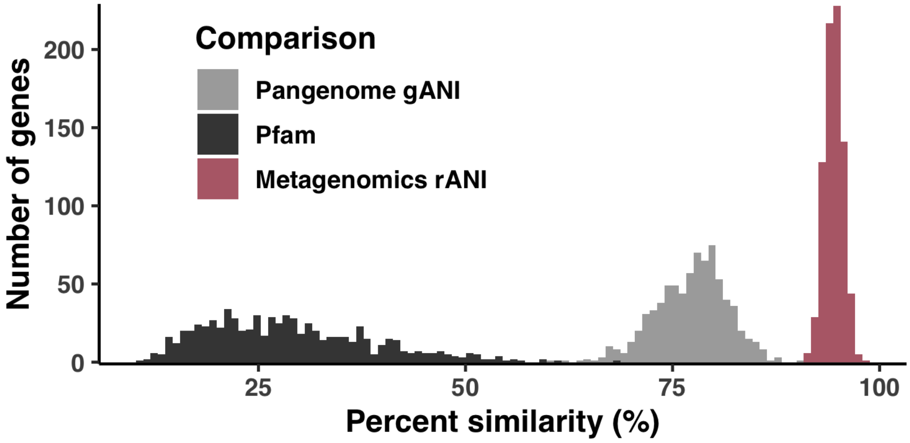


Figure 4.5: Regimes of sequence similarity probed by metagenomics, SAR11 cultured genomes, and protein families. Empirical distributions of gene-level percent similarity for HIMB83 compared with recruited metagenomic reads (pink), homologous SAR11 genomes (blue), and homologous Pfams (orange). For calculation details, see Supplementary Information.

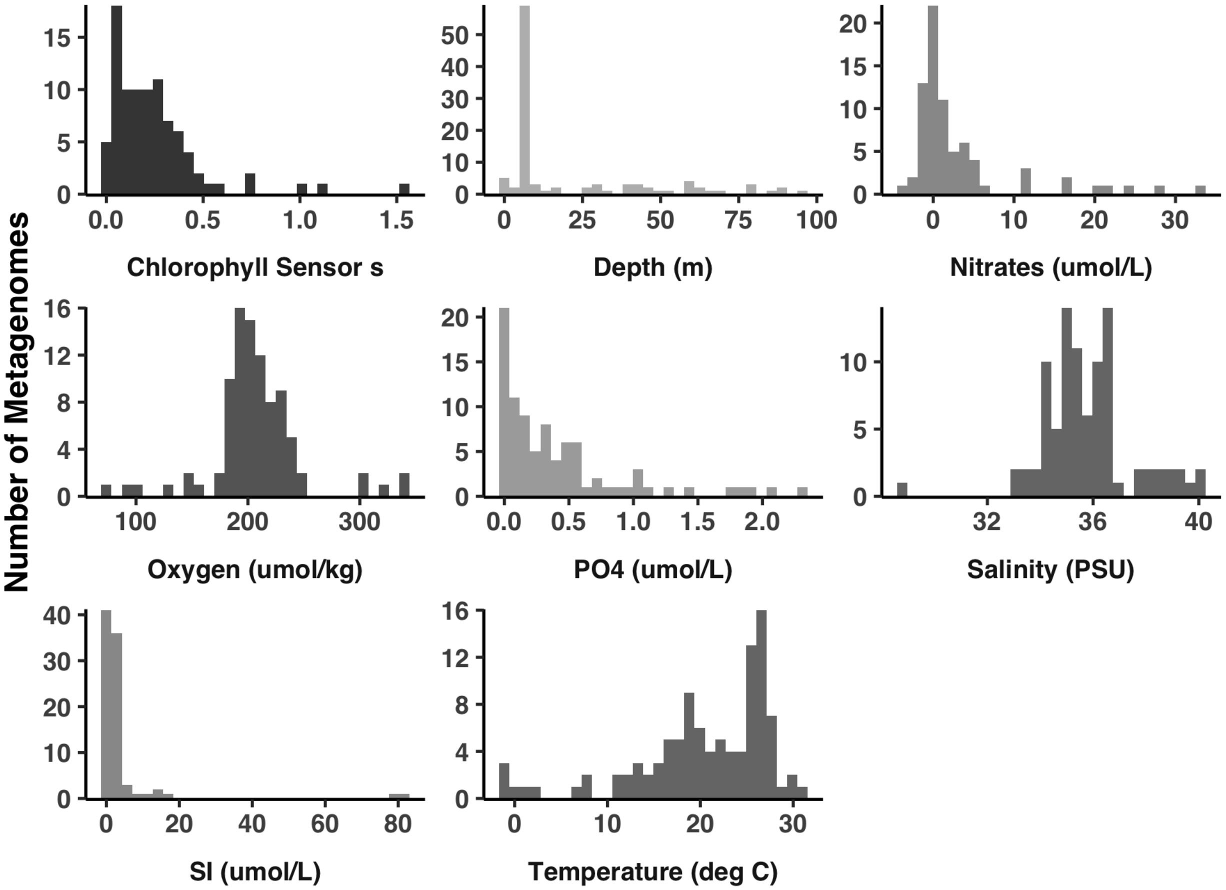


Figure 4.6: Different environments exhibit substantial variation in their environmental parameters. Each subplot shows how the 74 selected metagenomes distribute according to various environmental variables measured by the TARA ocean metagenome project.

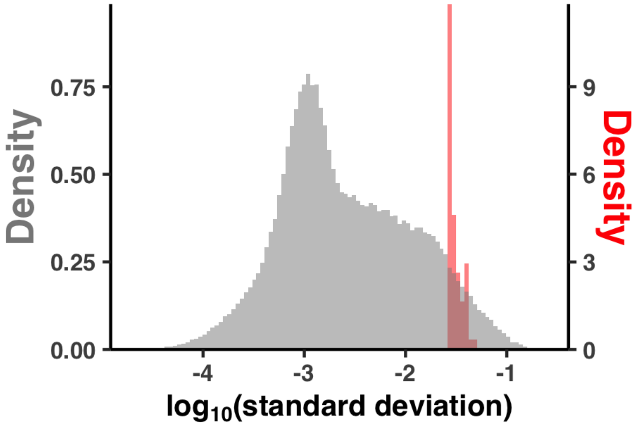


Figure 4.7: pN$^{(\text{site})}$ varies more significantly between sites in a given sample than between samples for a given site. The x-axis is the log-transformed standard deviation of either a sample's pN$^{(\text{site})}$ values observed over many sites (orange), or a site's pN$^{(\text{site})}$ values observed over the 74 samples (gray).

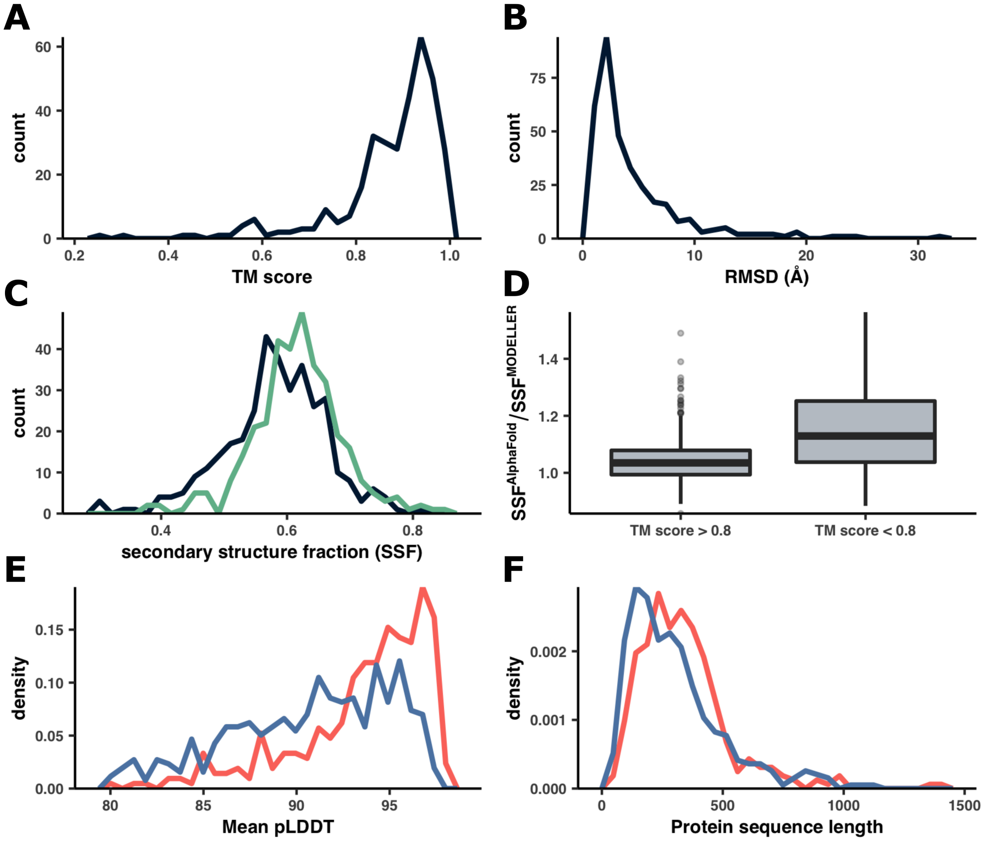


Figure 4.8: Comparisons between structures predicted by AlphaFold and MODELLER. (A-B) Distributions of TM scores and RMSD between structures predicted by both MODELLER and AlphaFold. (C) Distribution of secondary structure fractions, between MODELLER (black) and AlphaFold (green). Secondary structure fraction was defined for each gene as the fraction of sites that DSSP predicted as part of an alpha helix or beta strand. (D) Comparison of secondary structure fractions between MODELLER and AlphaFold for two TM score groups. The y-axis is the secondary structure fraction of AlphaFold divided by the secondary structure fraction of MODELLER. The two groups were defined as having TM scores above or below 0.8, where the >0.8 group corresponded to the 291 best alignments (left) and the <0.8 group corresponded to the 48 worst alignments. (E-F) Distributions describing the mean pLDDT and protein sequence length of AlphaFold structures that either (1) had analog MODELLER structures (red) or (2) did not (blue).

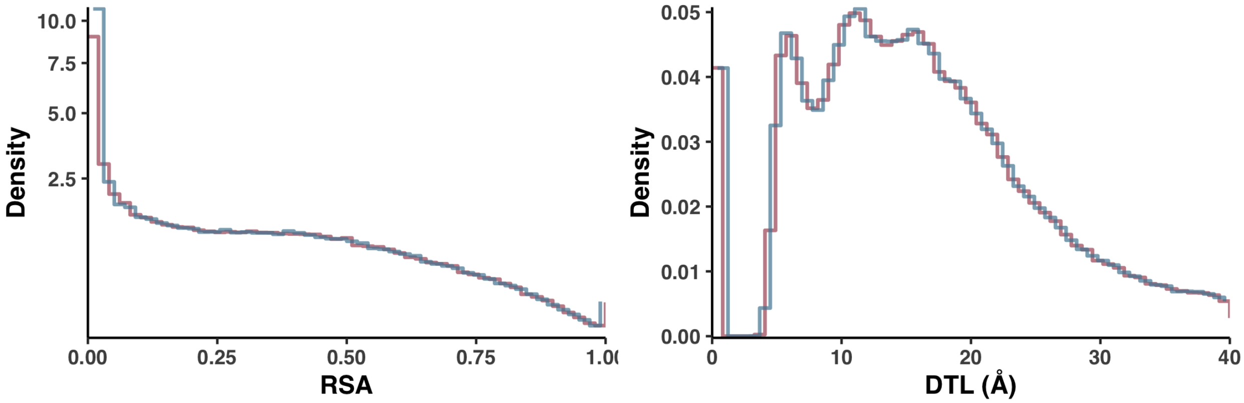


Figure 4.9: Comparison of null distributions for $pN^{(site)}$ and $pS^{(site)}$ for RSA and DTL. Each distribution was calculated by averaging 10 independent, randomly shuffled datasets of either $pN^{(site)}$ (red line) or $pS^{(site)}$ (blue line). To better visualize differences between the null distributions, the blue lines depicting the $pS^{(site)}$ distributions were shifted right by half of a bin's width.

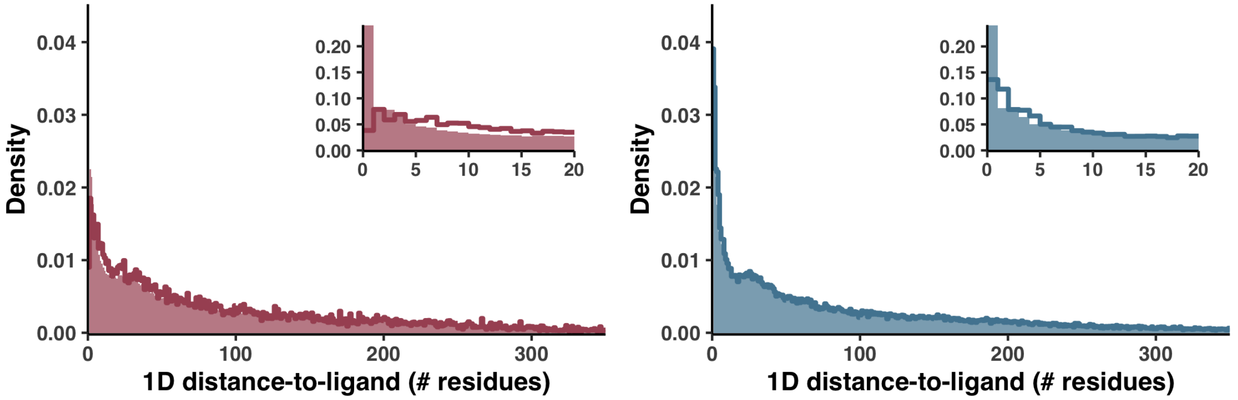


Figure 4.10: Functional constraint is less resolved when using a sequence-distance metric of DTL. $pN^{(site)}$ (left panel) and $pS^{(site)}$ (right panel) distributions with respect to 1D DTL, which we defined as the number of sites in a protein's sequence that separate a given site from a predicted ligand-binding site. Lines represent the observed distributions, and filled regions represent the null distributions, calculated via the shuffling procedure described in Figure 4.2. Insets show the same data zoomed into the 1D DTL range [0, 20].

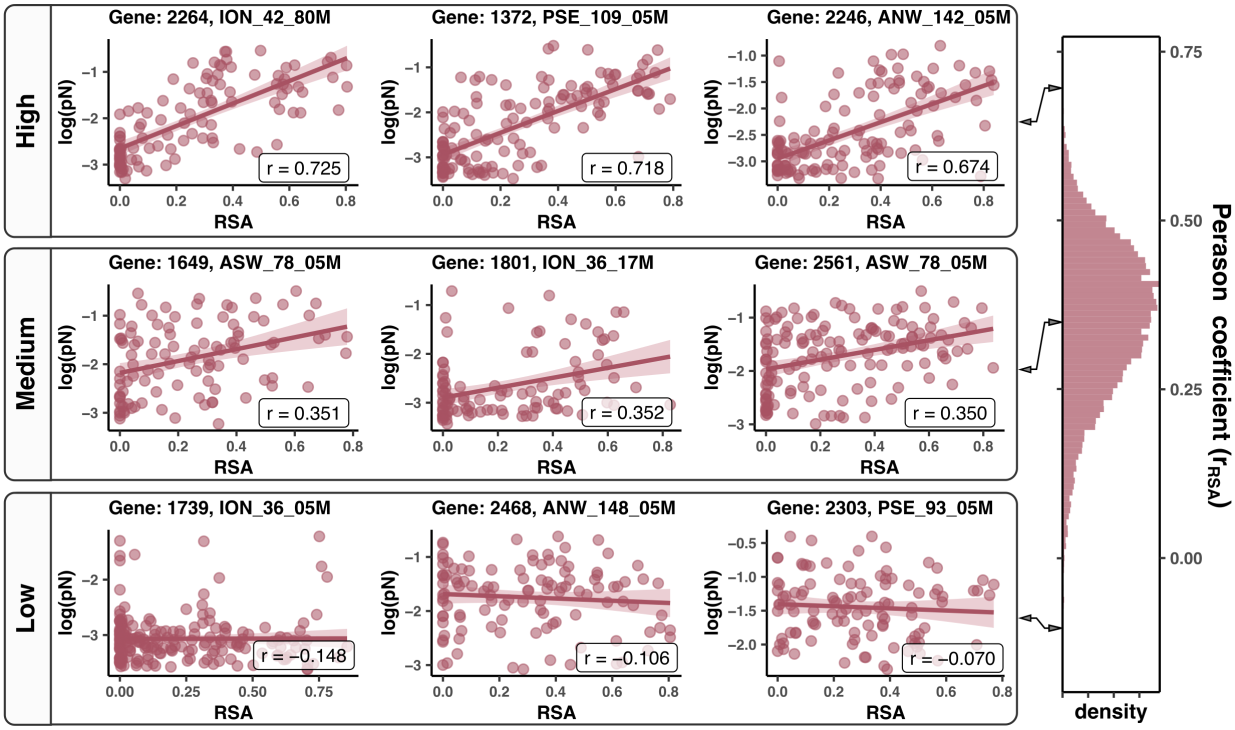


Figure 4.11: Select gene-sample pairs illustrate the diversity with which $\mathrm{pN}^{(\mathrm{site})}$ associates with RSA. Scatterplots for handpicked gene-sample pairs are shown from three regimes of model quality: high (top), mid (middle), and low (bottom). The right panel shows the distribution of Pearson coefficients, and the bin that each example was taken from is highlighted in pink. Each scatter plot is a gene-sample pair, each datapoint is a residue, the x-axis is the RSA of the residue, and the y-axis is the observed $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$. Lines of best fit are shown in red, with 95% confidence intervals visualized translucently. The Pearson coefficients of each fit are labeled on the scatterplot.

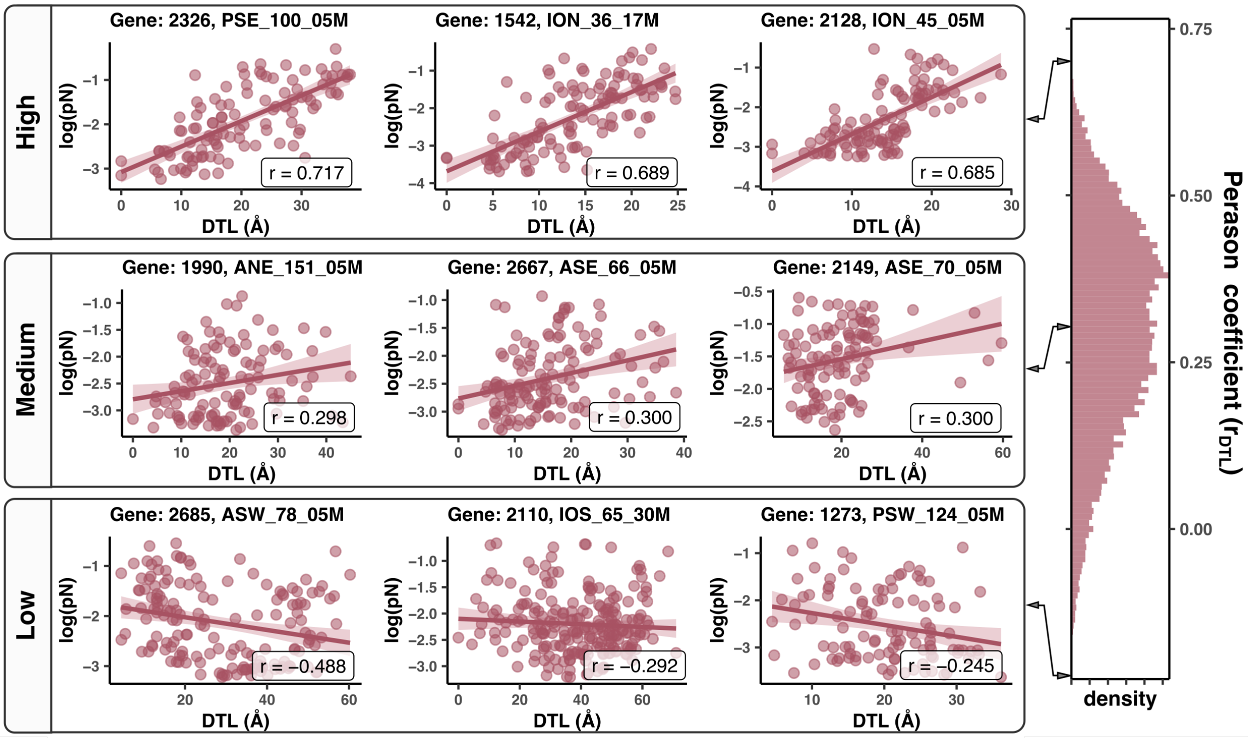


Figure 4.12: Select gene-sample pairs illustrate the diversity with which $\text{pN}^{(\text{site})}$ associates with DTL. Scatterplots for handpicked gene-sample pairs are shown from three regimes of model quality: high (top), mid (middle), and low (bottom). The right panel shows the distribution of Pearson coefficients, and the bin that each example was taken from is highlighted in pink. Each scatter plot is a gene-sample pair, each datapoint is a residue, the x-axis is the DTL of the residue, and the y-axis is the observed $\log_{10}(\text{pN}^{(\text{site})})$. Lines of best fit are shown in red, with 95% confidence intervals visualized translucently. The Pearson coefficients of each fit are labeled on the scatterplot.

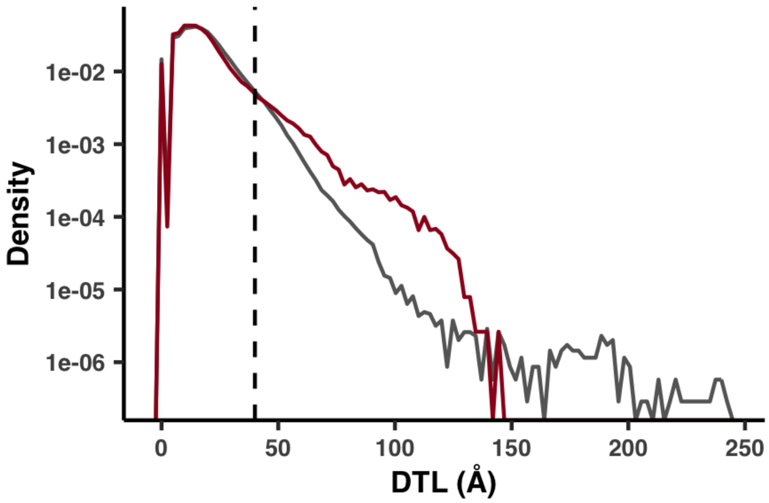


Figure 4.13: Incomplete ligand characterization leads to erroneously high DTL values. A comparison of DTL distributions (semi-log axis) for the 1a.3.V and the BioLiP database. The 1a.3.V core distribution (red) was calculated from all sites in the subset of genes with both a predicted structure and at least one predicted ligand-binding residue. The BioLiP distribution (gray) was calculated from the sites of 5,000 structures in the BioLiP database. For the 1a.3.V core, DTL was calculated as the distance to the closest predicted ligand-binding residue. For BioLiP, it was calculated as the distance to the closest annotated ligand-binding residue. For both methods, distance was calculated between the sites' side chain center of masses. The dashed line marks the 40Å cutoff we used for all analyses besides Figure 4.2b, which excludes 8.0% of the total sites.

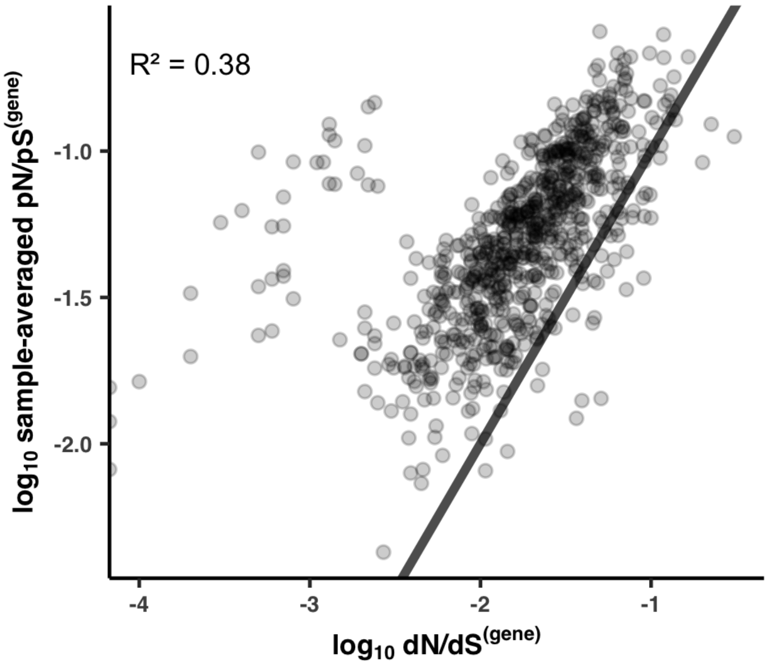


Figure 4.14: Sample-averaged pN/pS$^{(\text{gene})}$ values correlate with dN/dS$^{(\text{gene})}$ values between HIMB83 and HIMB122. The x- and y-axes are the log-transformed dN/dS$^{(\text{gene})}$ and sample-averaged pN/pS$^{(\text{gene})}$ values (respectively) for the 743 genes that (1) belonged to the 1a.3.V core and (2) had HIMB122 homologs. The black line is the equation y = x, meaning that genes above this line maintain sample-averaged pN/pS$^{(\text{gene})}$ values that exceed dN/dS$^{(\text{gene})}$. The R^2 is for a linear regression of the log-transformed variables.

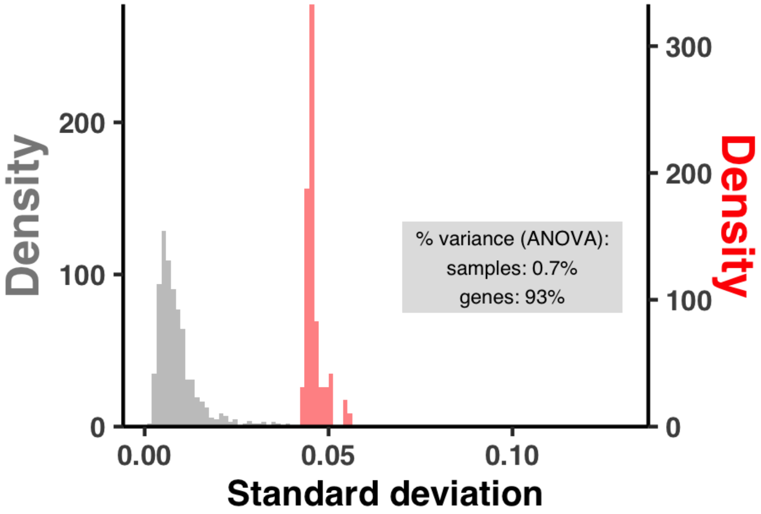


Figure 4.15: pN/pS$^{(gene)}$ varies more significantly between genes in a given sample than between samples for a given gene. The x-axis is the standard deviation of either a sample's pN/pS$^{(gene)}$ values observed over genes (orange), or a gene's pN/pS$^{(gene)}$ values observed over the 74 samples (gray). The gray box denotes the amount of variance explained by genes and samples in an ANOVA from the linear model pN/pS$^{(gene)}$ $\sim$ gene + sample.

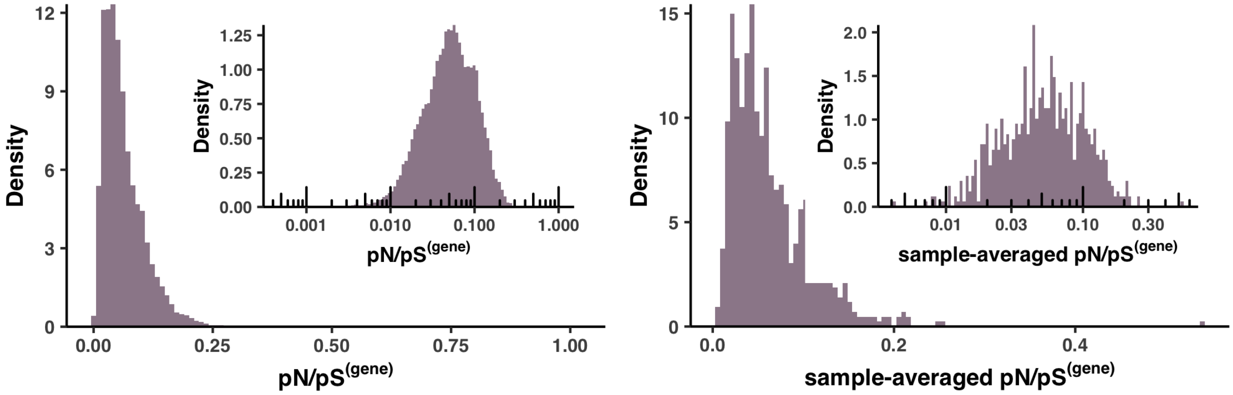


Figure 4.16: Distributions of pN/pS$^{(gene)}$. Left panel shows the distribution of pN/pS$^{(gene)}$, and the right panel shows the distribution of sample-averaged pN/pS$^{(gene)}$. Insets show the same distributions with a $\log_{10}$-transformed x-axis.

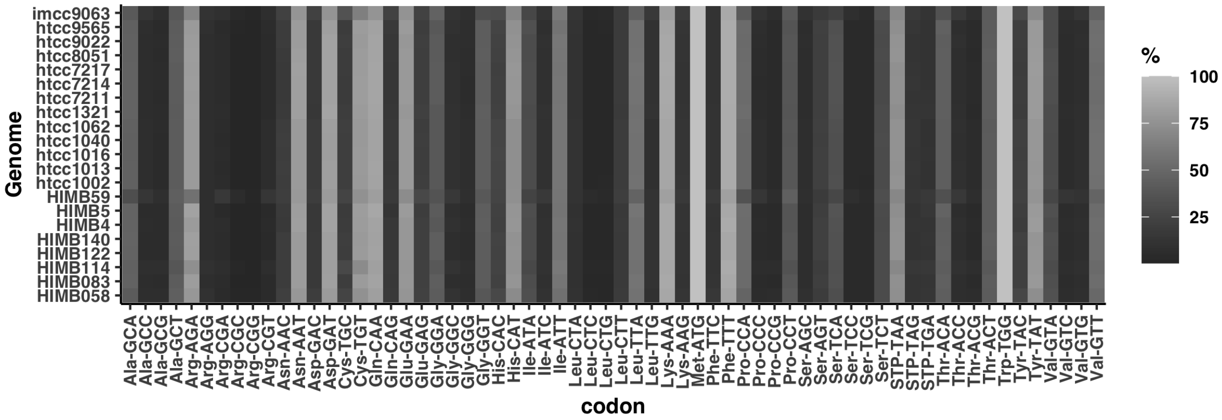


Figure 4.17: Codon usage of HIMB83 and 20 other genomes in the SAR11 clade.

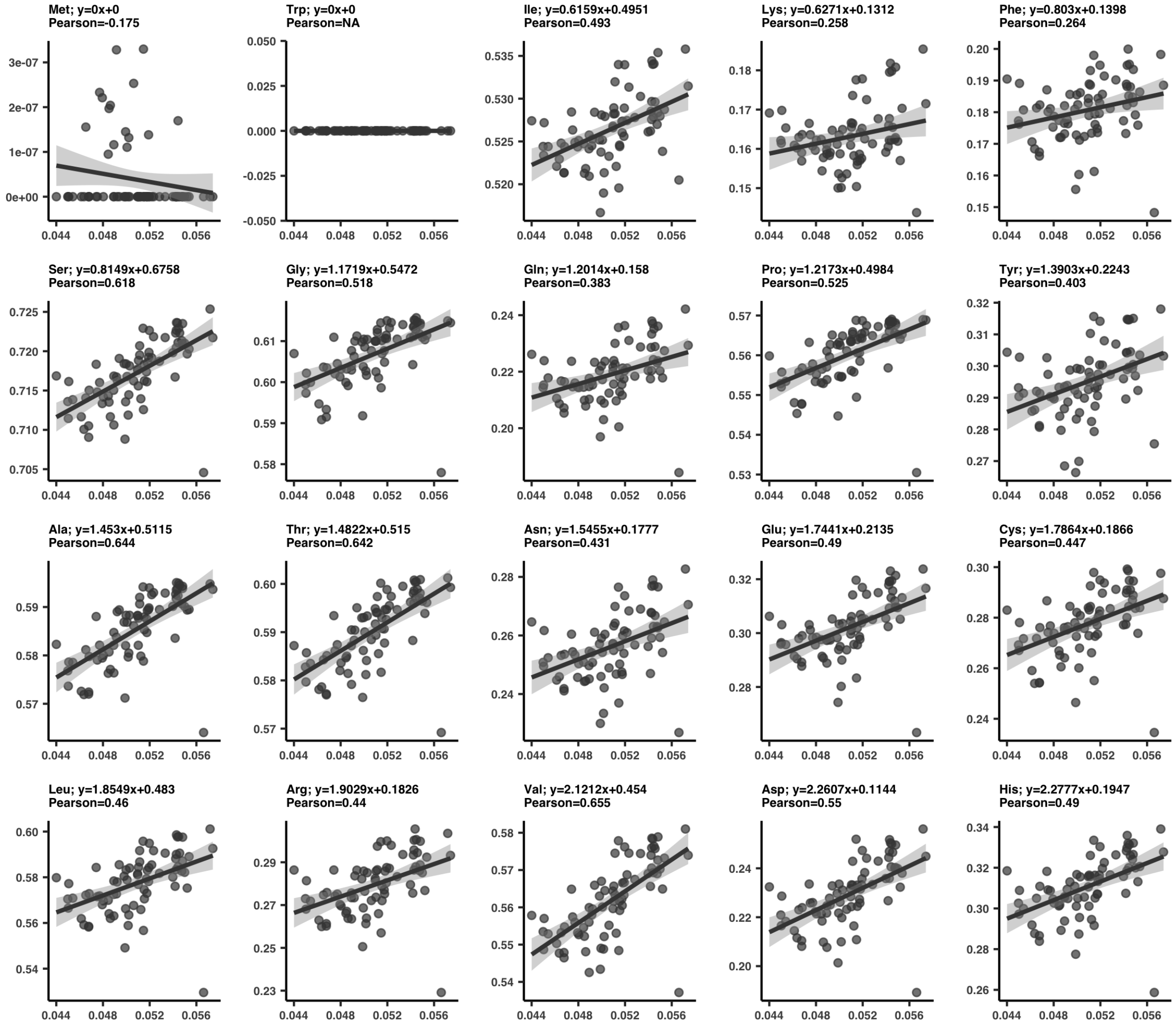


Figure 4.18: Codon rarity measured for each amino acid reveals varied response to selection strength, with most amino acids preferring rare codons in high selection samples. Each plot is a different amino acid, and each datapoint is a sample. The x-axis is $\mathrm{pN/pS}^{(\mathrm{core})}$, *i.e.* the ratio of nonsynonymous to s-polymorphism rates in the 1a.3.V core genome, and is shared between all plots. For a given plot, the y-axis was determined by first subsetting the polymorphism data to only include synonymous sites (in this instance we define synonymous as exhibiting $\mathrm{pN}^{(\mathrm{site})} < 0.0005$) that corresponded to the given amino acid. Using lysine as an example, this led to on average 21,127 sites per sample. For each amino acid in each sample, we then calculated the overall codon rarity (y-axis) by averaging codon rarities across all included positions. A line of best fit (gray line) with 95% confidence intervals (light gray) is shown for each plot, with equation and Pearson correlation coefficient shown above.

4.10 Supplementary Tables

All supplementary tables are available at `https://doi.org/10.6084/m9.figshare.19363997`.

Table 4.1: Read recruitment and coverage statistics of the 21 SAR11 genomes. (A-D) Genome-wide statistics for each genome in each metatranscriptomic and metagenomic sample. (A) is the mean coverage, (B) is the mean coverage, excluding nucleotide coverage values outside the interquartile range (IQR), (C) is the detection, and (D) is the percentage of reads mapping to a genome (sums to 100 for a given sample) (E) The mean coverage of each HIMB83 gene in each metatranscriptomic and metagenomic sample.

Table 4.2: Average percent similarity of recruited reads by HIMB83 for each (A) gene-sample pair, (B) gene (marginalized over samples), and (C) sample (marginalized over genes).

Table 4.3: Mean per-site polymorphism rates ($pN^{(\text{site})}$ and $pS^{(\text{site})}$) of HIMB83 (A) over all sites, genes, and samples, as well as (B) for each gene-sample pair (C) each gene (marginalized over samples), and (D) each sample (marginalized over genes).

Table 4.4: Methodological comparisons between AlphaFold and MODELLER structures. (A) Key metrics for AlphaFold- and MODELLER-predicted structures and their alignments. (B) PDB structures used as templates for MODELLER predictions. (C) Per-residue pLDDT scores for AlphaFold-predicted structures. (D) Gene-averaged pLDDT scores for AlphaFold-predicted structures. (E-F) Genes with AlphaFold and MODELLER structures, respectively, that we determined to be of sufficiently high quality.

Table 4.5: Summary of ligand-binding residue predictions with InteracDome. (A) All predicted ligand-binding sites, the predicted ligand, and the predicted ligand binding score. (B) Characterization of each HMM domain hit. (C) Each match state from the Pfam profile HMMs that contributed to each predicted ligand-binding residue of HIMB83.

Table 4.6: Summary of models used for estimating the explanatory power of RSA and DTL on polymorphism rates (see Methods).

Table 4.7: Summary statistics for the polymorphism models of gene-sample pairs.

Table 4.8: Summary of per-group polymorphism data for (A) $\text{pN}^{(\text{group})}$, (B) $\text{pS}^{(\text{group})}$, (C) $\text{pN/pS}^{(\text{group})}$, and (D) the size of each group.

Table 4.9: Summary of per-gene polymorphism data for (A) $\text{pN/pS}^{(\text{gene})}$, (B) sample-averaged $\text{pN/pS}^{(\text{gene})}$, (C) $\text{pN}^{(\text{gene})}$, (D) $\text{pS}^{(\text{gene})}$ and (E) the number of potential synonymous and nonsynonymous point mutations of each gene.

Table 4.10: Correlations of $\text{pN/pS}^{(\text{gene})}$ for each 1a.3.V core gene with respect to the measured environmental parameters: nitrates, chlorophyll, temperature, salinity, phosphate, silicon, depth, and oxygen.

Table 4.11: Codon metrics, including anti-codon, encoded amino acid, frequency and rarity in genome, and frequency and rarity compared to synonymous codons.

Table 4.12: Comparison between dN/dS between HIMB83 and HIMB122 homologs and sample-averaged $\text{pN/pS}^{(\text{gene})}$ of 1a.3.V genes.

Table 4.13: Per sample and gene measures of transcript abundance (TA) and related quantities.

Table 4.14: Bootstrap estimates of Pearson correlation coefficients and p-values from Figure 4.24.

4.11 Supplementary Information

4.11.1 Regimes of sequence similarity probed by metagenomics, SAR11 cultured genomes, and protein families

We investigated how sequence similarity between HIMB83 and aligned metagenomic reads compares to the traditional methods of sequence comparisons between other SAR11 cultured genomes, as well as between members of associated protein families. To do this, we calculated the percent similarity (PS) between HIMB83 genes and (a) all aligned reads, (b) homologs found in 20 SAR11 ocean isolates, and (c) members of the best matching Pfam protein family.

For (a), PS values for each gene were calculated by considering one metagenome at a time. In each metagenome, the reads that aligned to the gene were captured, trimmed (so there were no reads overhanging the gene), and compared to the aligned segment of HIMB83. The PS was calculated by comparing non-gap positions. This was then averaged to yield a PS value for each gene-metagenome pair. To define a single PS value for each gene, PS values were averaged across metagenomes.

For (b), gene clusters were calculated for HIMB83 and 20 additional SAR11 isolates using the anvi'o pangenomic workflow. An MSA was built from the sequences of each gene cluster using muscle [Edgar, 2004], and then each non-HIMB83 sequence was compared to the HIMB83 sequence. The PS was determined by calculating the fraction of matches in non-gap positions. Each HIMB83 gene was attributed a single PS value by averaging PS values in each pairwise comparison, weighted by the number of non-gap positions in the pairwise alignment. Gene clusters containing multiple HIMB83 genes were ignored.

For (c), HIMB83 genes were matched to Pfam protein families via the anvi'o program `anvi-run-pfams`. Hits that passed the GA gathering threshold were retained, and the best hit (lowest e-value) for each HIMB83 gene was defined as the associated Pfam. For each HIMB83 gene, the associated Pfam seed sequence MSA was downloaded using the Python package prody [Zhang et al., 2021] and the HIMB83 protein sequence was added to the MSA using muscle. PS values were calculated from the MSAs in a manner identical to that outlined in (b). It is important to note that this comparison used protein sequences, whereas (a) and (b) both used nucleotide sequences.

Figure 4.5 shows the distribution of percent similarities for each comparative method, roughly indicating the distinct regimes of evolutionary relatedness that each method probes. Unsurprisingly, protein families are most evolutionarily divergent (mean amino acid PS 28.8%). Relative to SAR11 homologs (mean nucleotide PS 77.3%), the aligned reads are highly related (mean nucleotide PS 94.5%), showing that metagenomics offers a modality of sequence inquiry more highly resolved than sequence comparisons between isolated cultures.

4.11.2 Comparing structure predictions between AlphaFold and MODELLER

The biggest difference between structure prediction methods was the expectedly higher portion of predictions yielded by AlphaFold. While AlphaFold produced 754 structures we deemed trustworthy (see Methods), MODELLER produced 346 due to its reliance on preexisting template structures. In 339 cases both methods procured a structure prediction for a given protein sequence, and it is within this intersection that we drew comparisons between the methods' structures.

We compared the topological similarity between AlphaFold and MODELLER structures using TM score [Zhang and Skolnick, 2004] and alpha carbon RMSD. Overall, the distributions of these metrics (Figures 4.8a, 4.8b) illustrate the overarching similarity between AlphaFold and MODELLER structures. Poor RMSD scores were usually the result of multidomain proteins linked by unstructured chains, and not the result of large structural discrepancies. TM scores better handle these cases. Since a TM score of 0.5 indicates that proteins likely belong to the same fold family [Xu and Zhang, 2010], our average TM score of 0.88 indicates strong overall agreement between AlphaFold and MODELLER.

On average, AlphaFold yielded a higher proportion of secondary structure (Figure 4.8c), and we found this discrepancy to be most pronounced when TM scores were low (<0.8) (Figure 4.8d). In fact, for the worst alignments (TM score <0.6), in 15 of 16 cases AlphaFold yielded more secondary structure.

Next, we turned our attention to proteins that AlphaFold predicted structures for, but that MODELLER did not due to absent templates. These proteins were on average smaller (Figure 4.8e) and yielded lower mean pLDDT scores compared to structures possessing a MODELLER analog. Since AlphaFold is trained on pre-existing structures, this result is expected and lends credence to pLDDT as a metric for fold confidence. Even still, these structures averaged a mean pLDDT score of 90.8, which is considered to be highly accurate

[Jumper et al., 2021].

Overall, our findings suggest that overall similarity between the two methods is high, that AlphaFold may be outperforming MODELLER due to increased fraction of secondary structure, and that proteins modeled by AlphaFold but not MODELLER are still considered highly accurate predictions.

4.11.3 RSA and DTL predict nonsynonymous polymorphism rates

To complement our analyses in which we estimated the percentage of polymorphism data that can be explained by RSA and DTL (Table S6, Methods), we constructed synonymous models (s-models) and nonsynonymous model (ns-models) for each gene in each sample. We excluded monomorphic sites ($\mathrm{pN}^{(\mathrm{site})} = 0$ for ns-models, $\mathrm{pS}^{(\mathrm{site})} = 0$ for s-models), sites with DTL > 40Å (see Methods), and removed gene-sample pairs containing <100 remaining sites, resulting in 16,285 ns-models and 24,553 s-models (Table 4.7).

We fit linear models of $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ and $\log_{10}(\mathrm{pS}^{(\mathrm{site})})$ to RSA. We found that applying a logarithmic function to polymorphism rates yielded better fits than without. We filtered out any genes that did not have a predicted structure and at least one predicted ligand-binding site, which when applied in conjunction with the above filters resulted in 381 genes for the s-models and 342 genes for the ns-models. ns-models yielded consistently positive correlations (average Pearson coefficient of rRSA = 0.353) (Figure 4.2c), whereas s-models exhibited correlations centered around 0 (average rRSA = -0.029). The average R^2 was 0.137 for ns-models, however model quality varied significantly between gene-sample pairs. In fact, we found that R^2 varied from as high as 0.526 (gene 2264 in sample ION_42_80M), to as low as 0.0% (gene 2486 in sample ION_42_80M). Lines of best fit for select gene-sample pairs illustrate the range of correlatedness seen between $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ and RSA (Figure 4.11). Overall, these results show that RSA is a significant predictor that partially explains the differences in polymorphism rates observed between sites in a given gene and sample.

Using the same procedure, we linearly regressed $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ and $\log_{10}(\mathrm{pS}^{(\mathrm{site})})$ with DTL and found that 96% of ns-models yielded positive correlations with DTL with considerable predictive power, where on average 11.5% of per-site ns-polymorphism rate variation could be explained by DTL (Table S7). R^2 values varied significantly, ranging from 0.514 (gene 2326 in sample PSE_100_05M) to 0.0% (gene 2246 in sample PSE_102_05M). Lines of best fit for select gene-sample pairs illustrate the range of relatedness observed between $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ and DTL (Figure 4.12). Interestingly, we found that $\log_{10}(\mathrm{pS}^{(\mathrm{site})})$ on average negatively correlates with DTL (average Pearson coefficient -0.057). The overall positive correlation of DTL with $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ suggests that on a proteomic scale, selection for function imposes a spectrum of per-site selective pressures, where pressure increases with proximity to ligand-binding regions.

Individually, RSA and DTL respectively explain 13.7% and 11.5% of per-site ns-polymorphism rate variance. To quantify their collective explanatory power, we fit a third set of models that linearly regressed $\log_{10}(\mathrm{pN}^{(\mathrm{site})})$ and $\log_{10}(\mathrm{pS}^{(\mathrm{site})})$ with RSA and DTL together (Figure 4.20; Table S7). A Pearson correlation between RSA and DTL revealed the relative independence of each variable from the other (R^2 = 0.082, r = 0.286), precluding effects of multicollinearity (Figure 4.19). The results revealed that including both RSA and DTL yielded a considerably better set of models for ns-polymorphism rates, with an average explained variance of 17.7% (average adjusted $R^2_{\text{RSA-DTL}} = 0.177$).

The predictive power of RSA and DTL illuminates how structural and functional constraints influence polymorphism rates by shaping the confines within which neutral evolution operates [Worth et al., 2009], yet observed rates can also be dominantly driven by stochastic processes of mutagenesis and drift. For example, no site will be polymorphic in the absence of a seeding mutagenesis event, even if under low structural and functional constraints. Thus, polymorphism rates are determined in part by constraints, and in part by random chance, the latter of which diminishes the predictive power of RSA and DTL when modeling polymorphism rates of individual sites.

By averaging across groups of sites, we vastly increased the signal-to-noise ratio of polymorphism rate data and revealed a two parameter model (RSA and DTL) that explains the majority of ns-polymorphism trends. To reduce per-site noise, we first grouped sites sharing similar RSA and DTL values so that each group contained the same order of magnitude of data (axes in Figure 4.2e, Table 4.8). For example, the group (RSA_1, DTL_2) contains the 3,164 sites with RSA values in the 1st RSA range [0.00,0.01) and DTL values in the 2nd DTL range [5.0Å,6.4Å). Then, we calculated per-group polymorphism rates $\mathrm{pN}^{(\mathrm{group})}$ and $\mathrm{pS}^{(\mathrm{group})}$, which are weighted averages of $\mathrm{pN}^{(\mathrm{site})}$ and $\mathrm{pS}^{(\mathrm{site})}$ values found within a group (see Methods). Averaging polymorphism rates across sites that exhibit similar RSA and DTL values has the effect of averaging out per-site and per-sample variance, which we found to reveal impressive proteome-wide trends in polymorphism rates with respect to RSA and DTL. $\mathrm{pN}^{(\mathrm{group})}$ values from each group collectively describe a 2D surface (Figure 4.2e, Table S8), where one axis illustrates how structurally constrained sites tend to be due to RSA and the other axis illustrates how functionally constrained sites tend to be due to DTL. In contrast to the noisy $\mathrm{pN}^{(\mathrm{site})}$ data observed within gene-sample pairs (Figures 4.11, 4.12), the $\mathrm{pN}^{(\mathrm{group})}$ surface is smooth and roughly linear (Figure 4.2e). Nonsynonymous polymorphism rates of groups varied from as low as 0.001 to as high as 0.021. A group's polymorphism rate appeared to be chiefly determined by the overall constraint of its sites, which is a composite of both structural and functional constraints. Structural and functional constraints appeared to be additive, such that sites with both low RSA and DTL (left panel of Figure 4.2e, bottom-left) statistically exhibited the lowest rates of ns-polymorphism, and sites with both high RSA and DTL (left panel of Figure 4.2e, top-right) statistically exhibited the highest rates of polymorphism. Additionally, these constraints are seen to act independently of one another: some groups exhibit low $\mathrm{pN}^{(\mathrm{group})}$ due to structural constraint (top-left) while others exhibit low $\mathrm{pN}^{(\mathrm{group})}$ due to functional constraint (bottom-right), illustrating that selection for structure and selection for function can independently constrain evolution.

Sites exhibited a spectrum of ns-polymorphism rates that is roughly linear. We determined this by fitting a linear model $\mathrm{pN}^{(\mathrm{group})} \sim i + j$, where i refers to the group's RSA and DTL indices (RSA_i, DTL_j), yielded an adjusted R^2 of 0.836, meaning that 83.6% of ns-polymorphism rate variation can be explained by RSA and DTL when averaging over per-site effects (Figure 4.21, Table S8). Increasing the number of groups decreased the number of sites in each group, weakening the efficacy of signal averaging, which expectedly decreased model quality. Even still, R^2 values for nonsynonymous models were robust to group numbers ranging from 4 (2x2) to 1,444 (38x38) (Figure 4.22).

Site averaging yielded an unexpected relationship between s-polymorphism rates and RSA/DTL. $\mathrm{pS}^{(\mathrm{group})}$ is not as strongly affected by RSA or DTL as $\mathrm{pN}^{(\mathrm{group})}$, as indicated by the noisy contour lines of its surface (right panel Figure 4.2e). Even still, the linear model $\mathrm{pS}^{(\mathrm{group})} \sim i + j$ yielded a significant, anti-correlated relationship with both RSA and DTL (adjusted R^2 of 0.206), in which s-polymorphism rates tended to decrease when RSA and DTL were high (Figure 4.21). We have observed this surprising finding through other means as well: in the sample-gene models, (1) the mean Pearson correlation coefficient between $\mathrm{pS}^{(\mathrm{site})}$ and RSA is -0.013 (Figure 4.2c), and (2) the mean Pearson correlation coefficient between $\mathrm{pS}^{(\mathrm{site})}$ and DTL is -0.052 (Figure 4.2d). Signal averaging has revealed the extent of its effect: 20.6% of s-polymorphism rates can be explained by RSA and DTL when averaging over per-site effects, compared to 83.6% for ns-polymorphism rates.

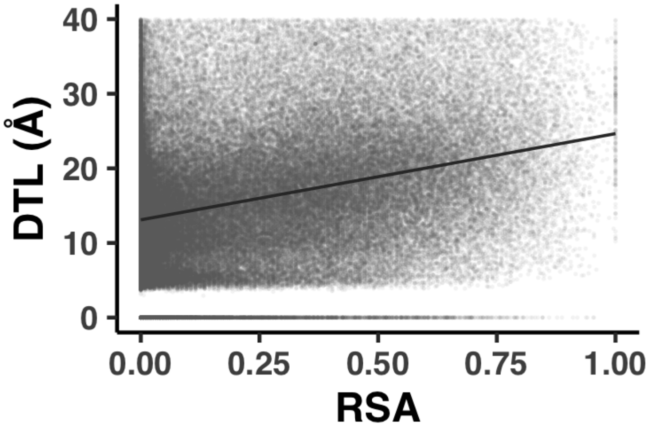


Figure 4.19: RSA and DTL are not problematically correlated. Scatter plot of RSA vs. DTL for the 143,181 sites belonging to genes with a predicted structure and at least one predicted ligand. The line of best fit is shown in black, The Pearson coefficient is 0.313 and the R^2 is 0.098.

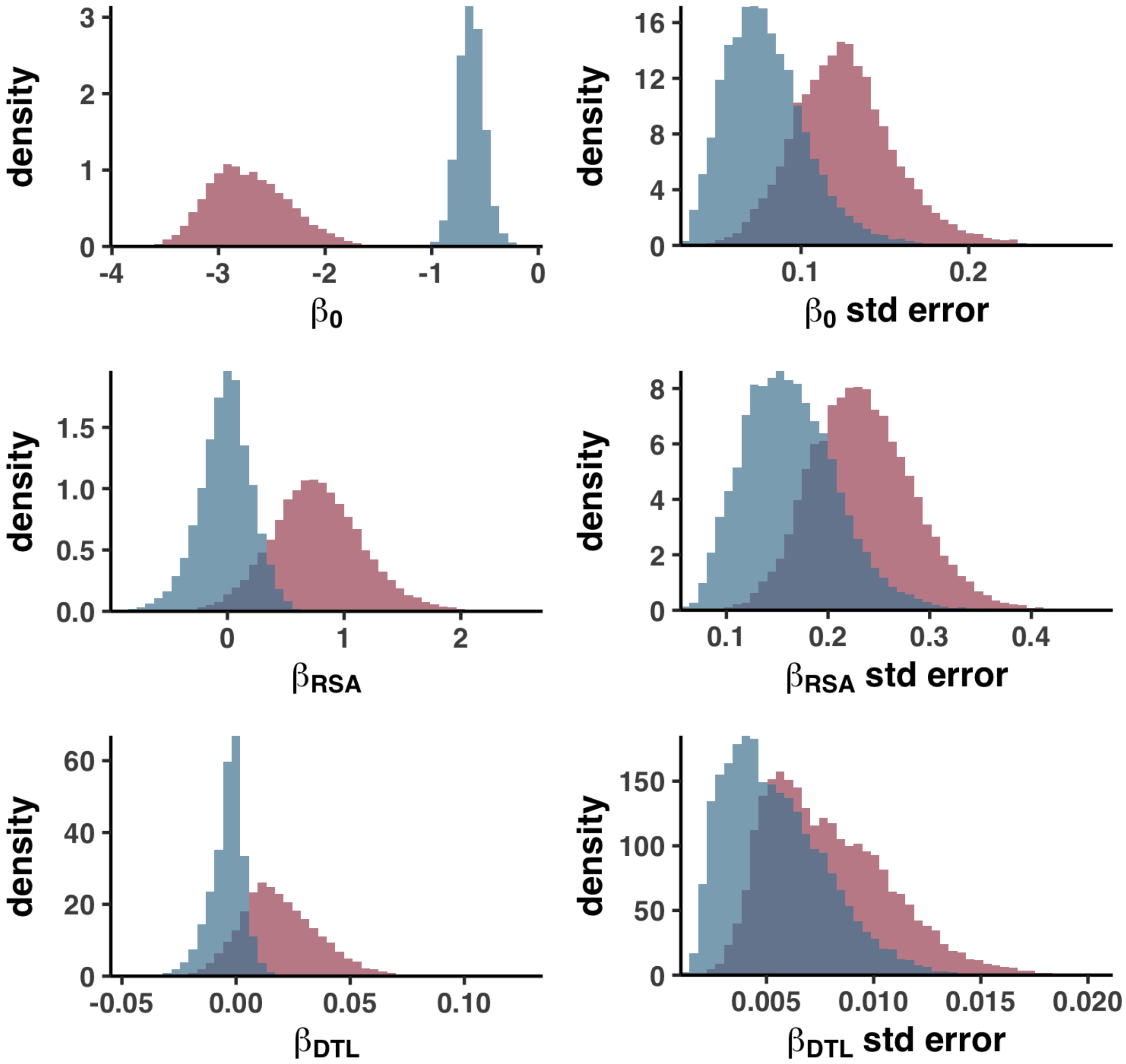


Figure 4.20: Parameter estimate and standard error distributions of the multidimensional linear regression models for $\text{pN}^{(\text{site})}$ and $\text{pS}^{(\text{site})}$. Red denotes parameter/error distributions for the 16,285 nonsynonymous models of the form $\text{pN}^{(\text{site})} = \beta_0 + \beta_{\text{RSA}}\text{RSA} + \beta_{\text{DTL}}\text{DTL}$ and blue denotes parameter/error distributions for the 24,553 models of the form $\text{pS}^{(\text{site})} = \beta_0 + \beta_{\text{RSA}}\text{RSA} + \beta_{\text{DTL}}\text{DTL}$.

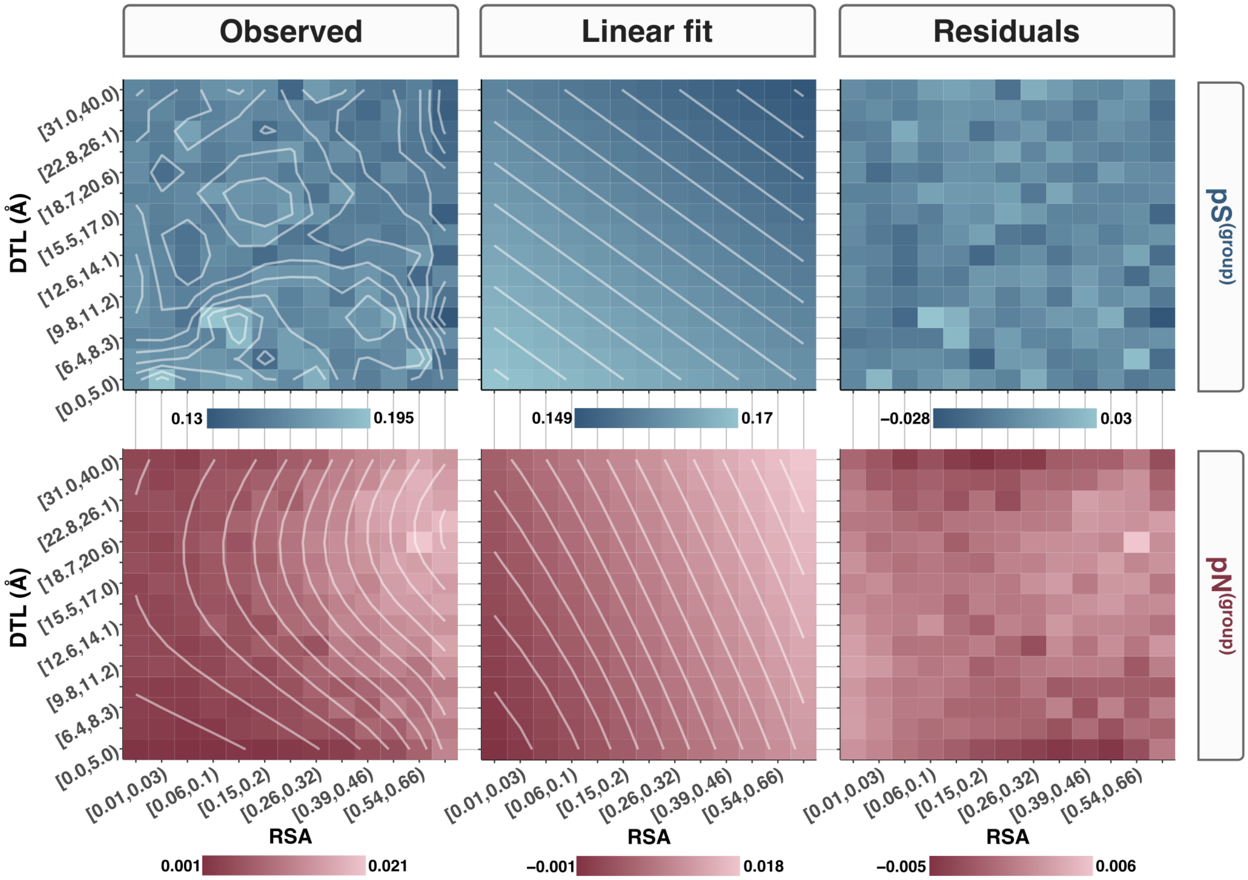


Figure 4.21: Observations, fits, and residuals of linear regressions for $pN^{(\text{group})}$, $pS^{(\text{group})}$, and $pN/pS^{(\text{group})}$. The x-axis and y-axis for each heatmap are RSA and DTL groups, respectively. The first column shows the observed values (those seen in Figure 4.2e), the second column shows the planes of best fit, and the third column shows the residuals. A legend for corresponding colors to values are shown below each heatmap. Contour lines for observed values and planes of best fit are shown as white and are calculated from smoothed data. Note that for the planes of best fit, the contour lines of the underlying data are by definition straight and perpendicular to one another, though due to edge effects of the smoothing procedure, there is a slight bend in the visualization of some contour lines.

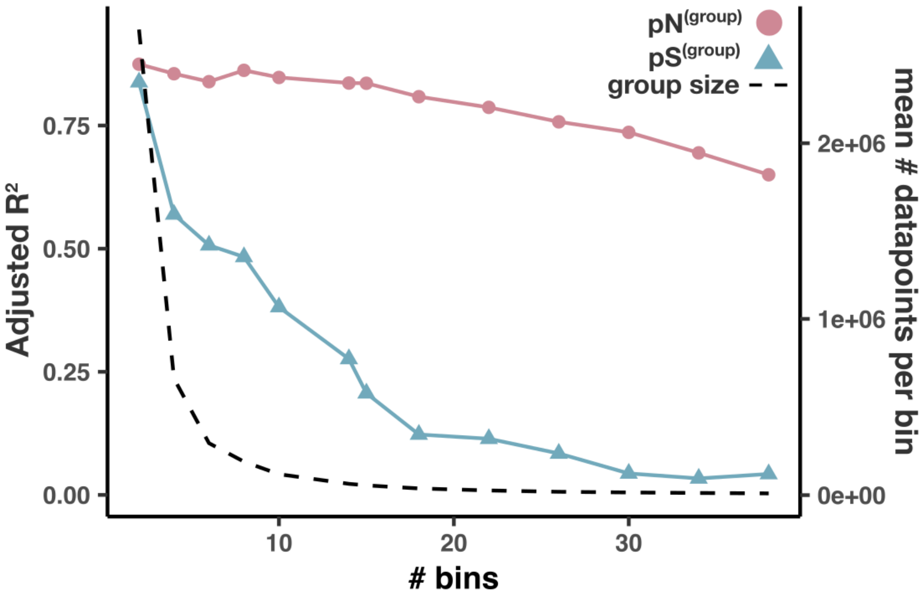


Figure 4.22: Model quality decreases for pN$^{\text{(site)}}$ and pS$^{\text{(group)}}$ as the number of RSA and DTL groups increases. The x-axis represents how many bins RSA and DTL are each split into. For example, the heatmaps in Figure 4.2e correspond to # bins = 15, since RSA and DTL are split into 15 bins, totaling 225 (=15x15) groups. The left y-axis corresponds to the adjusted R^2 value for the models pN$^{\text{(group)}}$ (red) and pS$^{\text{(group)}}$ (blue). The right y-axis corresponds to the average number of data points (# sites multiplied by # samples) found in a group (dashed black line).

4.11.4 dN/dS$^{(gene)}$ and sample-averaged pN/pS$^{(gene)}$ yield consistent results

To validate our pN/pS$^{(\text{gene})}$ calculations, we ascribed a sample-averaged pN/pS$^{(\text{gene})}$ value to each gene and compared the values to dN/dS$^{(\text{gene})}$ (Table 4.12), a more commonly and classically utilized metric that is the ratio of nonsynonymous to synonymous substitutions observed between homologous genes of two or more species. We calculated dN/dS$^{(\text{gene})}$ for 753 homologous gene pairs found between HIMB83 and a closely related cultured representative HIMB122 (see Methods). Importantly, ANI between HIMB83 and HIMB122 was 82.6%, whereas the average ANI between HIMB83 and recruited reads was 94.5%, making it unlikely that sample-averaged pN/pS$^{(\text{gene})}$ and dN/dS$^{(\text{gene})}$ were cross-contaminated due to HIMB83 recruiting significant proportions of reads from HIMB122-like populations. We found that log-transformed sample-averaged pN/pS$^{(\text{gene})}$ highly correlated with log-transformed dN/dS$^{(\text{gene})}$ (Pearson $R^2 = 0.380$), showing that the two metrics are commensurable. Nevertheless, differences were expected and observed. The ratio between sample-averaged pN/pS$^{(\text{gene})}$ and dN/dS$^{(\text{gene})}$ was on average 6.23 (Figure 4.14), matching expectations that slightly deleterious, nonsynonymous mutants commonly drift to observable frequencies, yet far less commonly drift to fixation.

4.11.5 Transcript abundance largely explains genic differences in the strengths of purifying selection

Sample-averaged pN/pS$^{(\text{gene})}$ values varied significantly between genes, varying from 0.004-0.539, with a mean of 0.063 (Figure 4.16, Table 4.9). What causes such variation in purifying selection strengths? Across diverse taxa [Drummond and Wilke, 2008], it has been shown that highly expressed proteins evolve more slowly due to being selectively constrained to be robust to mistranslation in order to safeguard against toxicity of misfolded proteins, whose detrimental fitness costs scale with expression level [Drummond et al., 2005]. We assessed the extent to which expression level may explain purifying selection variation in 1a.3.V by calculating metatranscriptomic coverage values for each 1a.3.V core gene in the 50 of 74 environments that had accompanying metatranscriptomics datasets (see Methods). We defined transcript abundance (TA) as the ratio of metatranscriptomic to metagenomic relative abundances (see Methods), which yielded a widely skewed distribution of values (Figure 4.23a, Table 4.13).

Comparing sample-median TA values to sample-averaged $\mathrm{pN/pS}^{(\mathrm{gene})}$ values yielded a strong, negative correlation (Figure 4.23b, Pearson r = -0.539, $R^2 = 0.290$) according to an inverse power-law relationship. The specific form of the linear model used was $\log_{10}(\mathrm{median}_s(TA) + 0.01) \sim \log_{10}(\mathrm{mean}_s(\mathrm{pN/pS}^{(\mathrm{gene})}))$, where medians and means denote the median and mean across samples for a given gene, respectively. To avoid excluding zeros, we added 0.01 to the log-transformation of medians(TA). These findings indicate that 29.0% of purifying selection variation between genes can be explained via transcript abundance alone, a value in line with what has been observed between yeast homologs [Drummond et al., 2005]. Overall, these results recapitulate a central result in protein evolution, and demonstrate its validity *in situ* using culture-independent approaches that link genetic variation and transcript abundance for a naturally occurring microbe.

Next, we tested whether $\mathrm{pN/pS}^{(\mathrm{gene})}$ values between samples of a given gene also follow an inverse power-law relationship with TA. We found that of the 799 genes tested, 74% exhibited (weak) negative correlations between $\log_{10}(\mathrm{TA}+0.01)$ and $\log_{10}(\mathrm{pN/pS}^{(\mathrm{gene})})$ (Figure 4.23c), yet only 11.5% of genes passed significance tests (one-sided Pearson, 25% Benjamini-Hochberg false discovery rate) (Figure 4.23d). Given the strong correlation observed between genes, the lack of correlation observed between samples is a seemingly contradictory result, yet can be attributed to a difference in timescales: TA fluctuates on the order of minutes, often occurring in 'bursts', whereas $\mathrm{pN/pS}^{(\mathrm{gene})}$ is shaped over time scales orders of magnitude longer than the 2 week replication time of SAR11. Since metagenome-metatranscriptome pairs sample single snapshots in time, measured TAs are unlikely to reflect the time-averaged values that constrain pN/pS$^{(\mathrm{gene})}$. These fluctuations therefore muddy signals that may exist between pN/pS$^{(\mathrm{gene})}$ and TA. Smoothing these fluctuations by averaging across samples thereby reveals the strong negative correlation observed (Figure 4.23b). In other words, TAs that are not averaged across environments are unreliable proxies for overall transcription level.

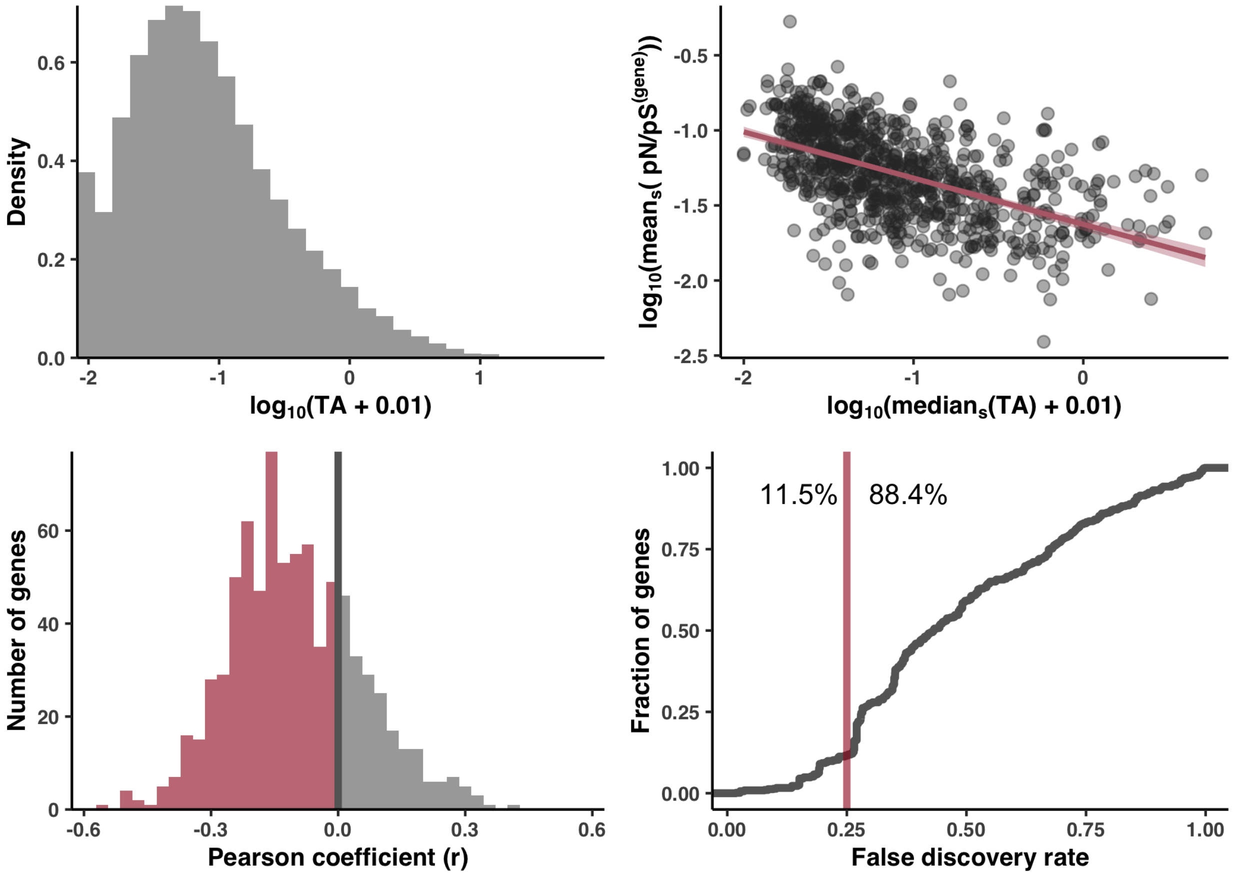


Figure 4.23: Associations of transcript abundance (TA) data with pN/pS$^{(\text{gene})}$. (A) Log-transformed distribution of TA values across genes and samples. See Methods for details on TA calculation. 0.01 has been added to the log-transformation to avoid the exclusion of zeros. (B) TA is a strong predictor of pN/pS$^{(\text{gene})}$ when pooling data across samples. Each datapoint is a gene, where the x-axis is the gene's median TA across samples, the y-axis is the gene's sample-averaged pN/pS$^{(\text{gene})}$, and each axis has been log-transformed. The linear model yielded a Pearson coefficient of 0.539, an R^2 of 0.290 and a line of best fit y = (-0.31 ± 0.02)x + (-1.63 ± 0.02) shown in pink (95% confidence intervals shown in translucent pink). (C) pN/pS$^{(\text{gene})}$ between samples of a given gene weakly correlate (on average) with TA. A one-side Pearson correlation between $\log_{10}$(TA + 0.01) and $\log_{10}$(pN/pS$^{(\text{gene})}$) was calculated separately for 799 genes, resulting in the following distribution of Pearson coefficients, of which 74% were negative (pink). (D) Accounting for multiple testing yields few statistically significant negative correlations. The x-axis is the Benjamini-Hochberg false discovery rate (FDR) and the y-axis is the fraction of genes that have statistically meaningful negative correlations for a given FDR. Allowing a FDR of 25% (pink line), only 11.5% of genes have statistically significant negative correlations of $\log_{10}$(TA + 0.01) with $\log_{10}$(pN/pS$^{(\text{gene})}$).

4.11.6 Stability analysis of polymorphism distributions with respect to $pN/pS^{(core)}$

To assess whether the ‘use it or lose it’ accumulation of ns-polymorphism in low RSA/DTL sites was specific to GS, or a more general feature of 1a.3.V, we performed a comparable procedure where instead of restricting our analysis to GS, we compiled polymorphism rates across all sites in genes with predicted structures and ligand-binding sites, and calculated $\mathrm{pN/pS}^{(\mathrm{core})}$ for each sample, which serves as a proxy for genome-wide selection strength (see Methods). Within this dataset, we observed the same phenomena: in samples with high selection strength (low $\mathrm{pN/pS}^{(\mathrm{core})}$), ns-polymorphism throughout the genome distributed (a) in more solvent-exposed sites (Figure 4.4a) and (b) farther from predicted binding sites (Figure 4.4b). Our bootstrapping stability analysis (Figure 4.24, Table 4.14) showed that in 99.5% of gene resamplings, the mean RSA of ns-polymorphism negatively associated with $\mathrm{pN/pS}^{(\mathrm{core})}$ (one-sided Pearson coefficient p-value <0.05), whereas in only 69.5% of gene resamplings did the mean DTL of ns-polymorphism negatively associate with $\mathrm{pN/pS}^{(\mathrm{core})}$. This latter finding indicates that the signal in Figure 4.4b is driven by an incomplete set of the 1a.3.V core genes. We hypothesized this is due to the many shortcomings of DTL estimation discussed priorly leading to false-positive and/or false-negative ligand predictions that skew DTL distributions, or that not all ligands constraint ns-polymorphism patterns equally.

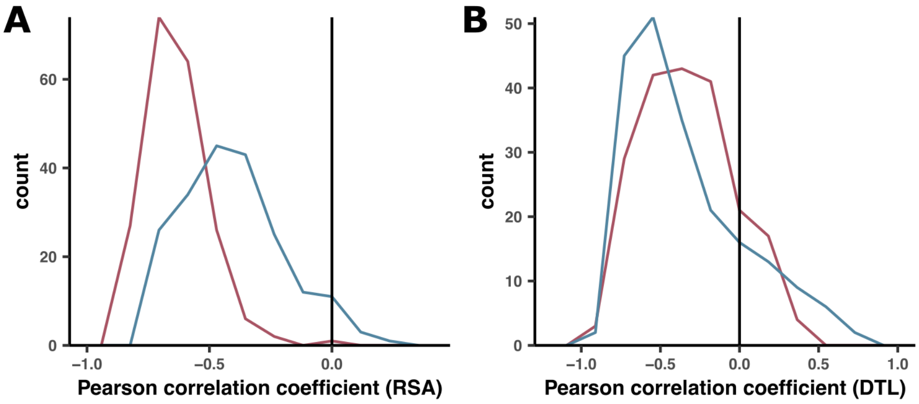


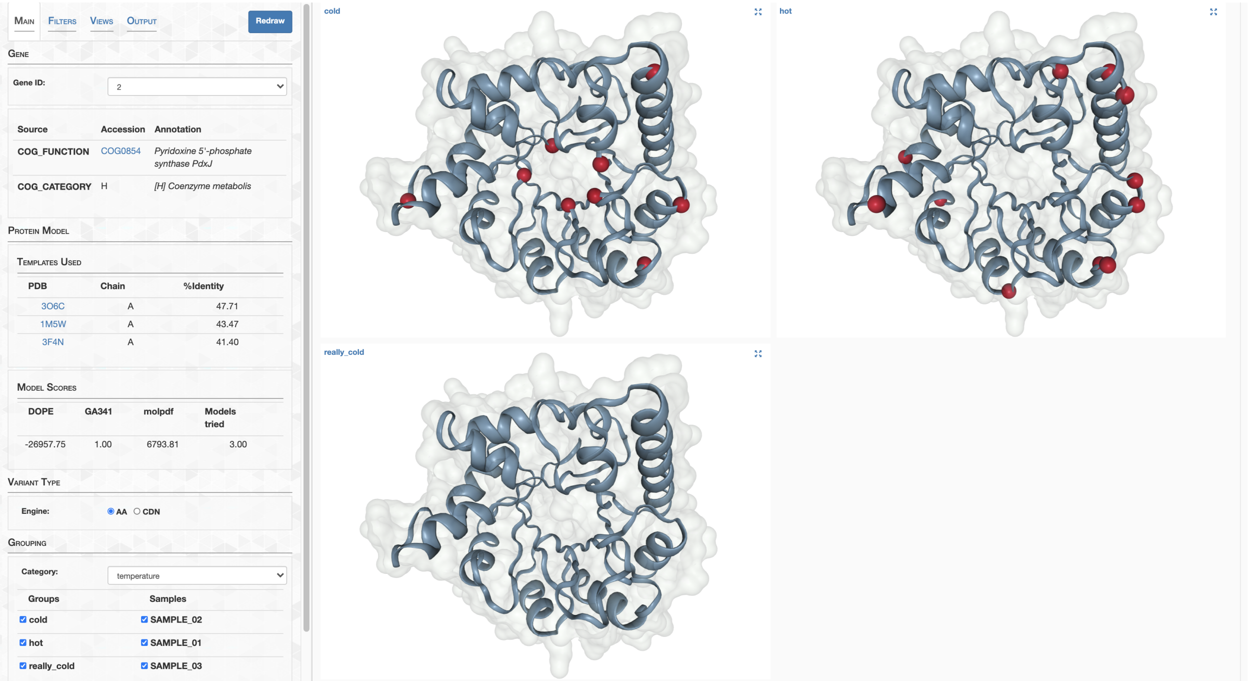
Figure 4.24: Robustness of negative associations between sample selection strength (pN/pS$^{(\mathrm{core})}$) and mean RSA/DTL of polymorphisms. We tested the robustness of results in Figure 4.4 by performing a bootstrapping stability analysis in which we created 200 bootstrapped estimates of the correlation coefficients, where each bootstrap was a resampling of genes. (A) Histograms of the correlation coefficients between the mean RSA of s-polymorphism (blue) and ns-polymorphism (red) versus pN/pS$^{(\mathrm{core})}$. These correspond to Figures 4.4a and 4.4c, respectively. (B) Histograms of the correlation coefficients between the mean DTL of s-polymorphism (blue) and ns-polymorphism (red) versus pN/pS$^{(\mathrm{core})}$. These correspond to Figures 4.4b and 4.4d, respectively.

4.11.7 Enabling interactive, exploratory, structure-informed metagenomic analyses using anvi-display-structure

There is an absence of computational tools that allow researchers to interactively explore metagenomic sequence variance in the context of predicted protein structures and ligand-binding sites. We addressed this gap by developing an interactive interface in which users can visualize, filter, and interact with metagenomic sequence variants in the context of modeled protein structures and predicted binding sites (Figures 4.25, 4.26, 4.27, 4.28). The exploratory analyses enabled by the interface is what has made the current research possible.

We created an interactive interface that dynamically processes data from anvi'o databases, which is done with the program `anvi-display-structure`. Once the interactive interface is initiated, users can select any gene with a modeled structure in their dataset, upon which anvi'o renders the predicted structure of the gene using NGL [Rose and

Hildebrand, 2015, Rose et al., 2016] and overlays sequence variants from metagenomes directly on the structure. By default, all variants across all metagenomes for a given gene are superimposed on a single display, however, the user can subdivide the display into as many as 16 sub-displays to compare and contrast variation across arbitrary groups of metagenomes (Figure 4.25). The interface offers numerous ways to interact with and explore single-codon variants (SCVs) and single-amino acid variants (SAAVs). Hovering the mouse above any variant reveals its allele frequency vector and structural information of the reference residue such as solvent accessibility and secondary structure (Figure 4.26). Interactive sliders filter variants displayed on structures through a suite of continuous, discrete, and categorical variables, including variant-specific parameters such as site entropy, solvent accessibility, BLOSUM scores of the competing alleles, residue number, and secondary structure (Figure 4.26). These same variables can also dynamically change the color and size of individual variants (Figure 4.27). Filters can be combined for exploratory investigations. For example, a user could simultaneously color variants by site entropy, size them by their coverage in metagenomes, and filter out those that exhibit high solvent accessibility (Figure 4.27). The protein surface and backbone can be colored according to arbitrary user-provided data, for example, to visualize predicted binding sites of the protein. `anvi-display-structure` can save and load sessions to preserve filters, export displays as PNG images, and generate rich tabular outputs for allele frequencies and other properties of displayed variants. Finally, users can faithfully migrate the current view into PyMOL [DELANO and W. L, 2002] for further graphical refinement or statistical analyses (Figure 4.28).



Gene

Gene ID: 2

Source	Accession	Annotation
COG_FUNCTION	COG0854	*Pyridoxine 5'-phosphate synthase PdxJ*
COG_CATEGORY	H	*[H] Coenzyme metabolis*

Protein Model

Templates Used

PDB	Chain	%Identity
3O6C	A	47.71
1M5W	A	43.47
3F4N	A	41.40

Model Scores

DOPE	GA341	molpdf	Models tried
-26957.75	1.00	6793.81	3.00

Variant Type

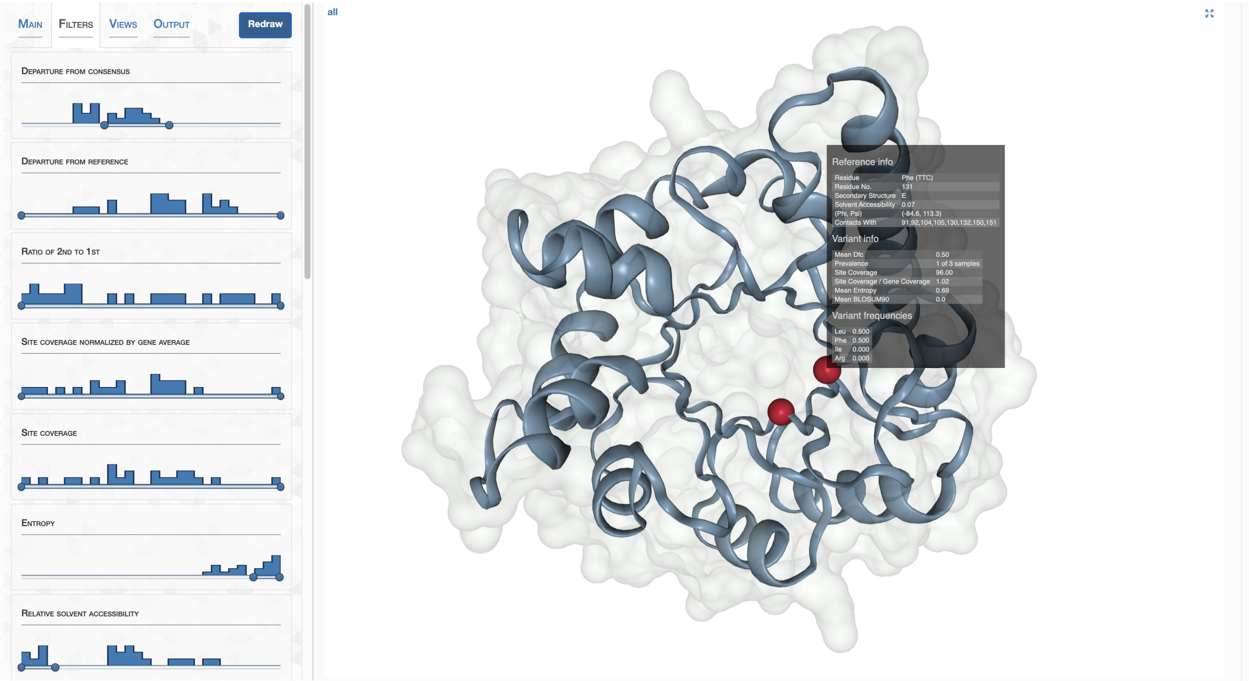
Engine: ● AA ○ CDN

Grouping

Category: temperature

Groups	Samples
☑ cold	☑ SAMPLE_02
☑ hot	☑ SAMPLE_01
☑ really_cold	☑ SAMPLE_03

Figure 4.25: Screenshot of the interface with the "Main" tab active. The user has chosen to visualize Gene ID 2 from the left-hand side panel. Functional annotations from COG indicate this is a Pyridoxine 5'-phosphate synthase, and its structure was modeled using the PDB IDs 3O6C, 1M5W, and 3F4N templates. The resulting structure is visualized on the right-hand side in 3 separate views corresponding to each of the 3 groups of metagenomes specified by the user in the bottom left corner. The spheres overlaid onto the 3 views are the positions of single-amino acid variants found from each group, and can be switched to single-codon variants by switching the Variant Type Engine from “AA” (amino acid) to “CDN” (codon).



Reference info

Residue	Phe (TTC)
Residue No.	131
Secondary Structure	E
Solvent Accessibility	0.07
(Phi, Psi)	(-84.6, 113.3)
Contacts With	91,92,104,105,130,132,150,151

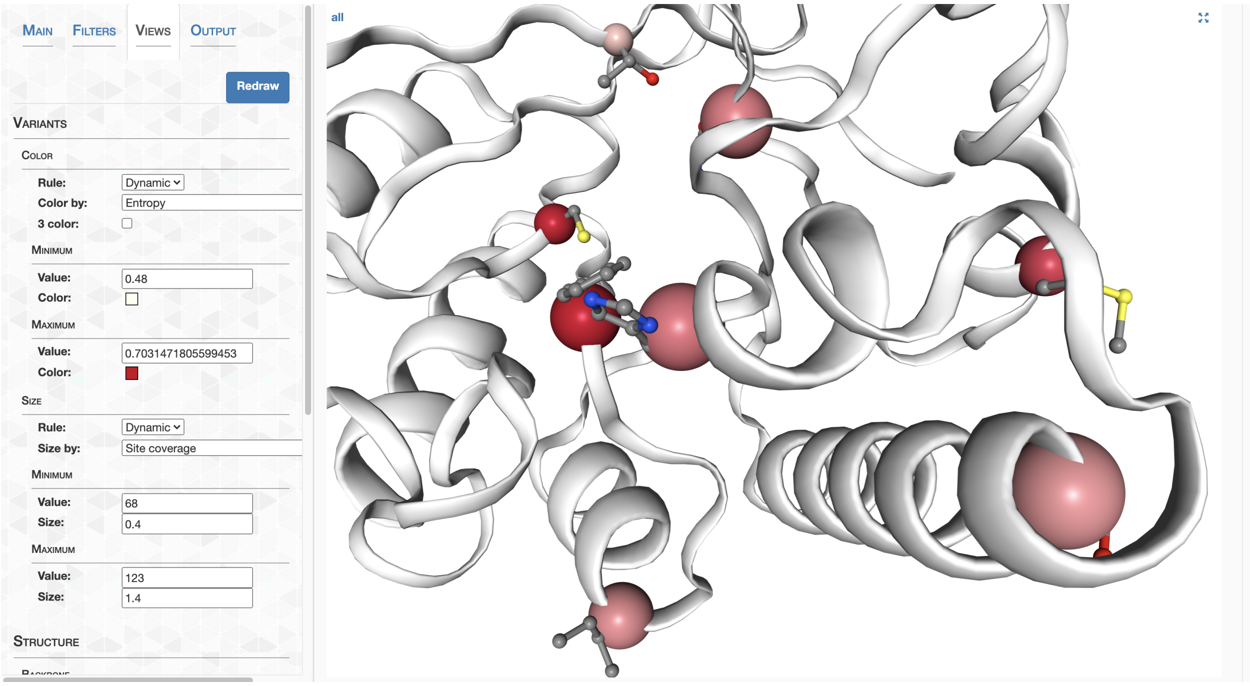
Variant info

Mean Dfc	0.50
Prevalence	1 of 3 samples
Site Coverage	96.00
Site Coverage / Gene Coverage	1.02
Mean Entropy	0.69
Mean BLOSUM90	0.0

Variant frequencies

Leu	0.500
Phe	0.500
Ile	0.000
Arg	0.000

Figure 4.26: Screenshot of the interface with the "Filter" tab active. Variants can be filtered in the "Filters" tab, which shows a suite of filters, each represented as an interactive slider with endpoints that can be clicked and dragged by the user. Above each slider is a histogram detailing how the variants distribute according to the filter. In this screenshot, the user has included variants with mid-range "departure from consensus" values, high "entropy" values, and low "relative solvent accessibility" variants. The right-hand side reveals that two variants (red spheres) match this filter criteria. Hovering the mouse above one of the variants activates a pop-up menu from which relevant statistics can be learned about.



Variants

Color

Rule: Dynamic

Color by: Entropy

3 color: □

Minimum

Value: 0.48

Color: □

Maximum

Value: 0.7031471805599453

Color: 🟥

Size

Rule: Dynamic

Size by: Site coverage

Minimum

Value: 68

Size: 0.4

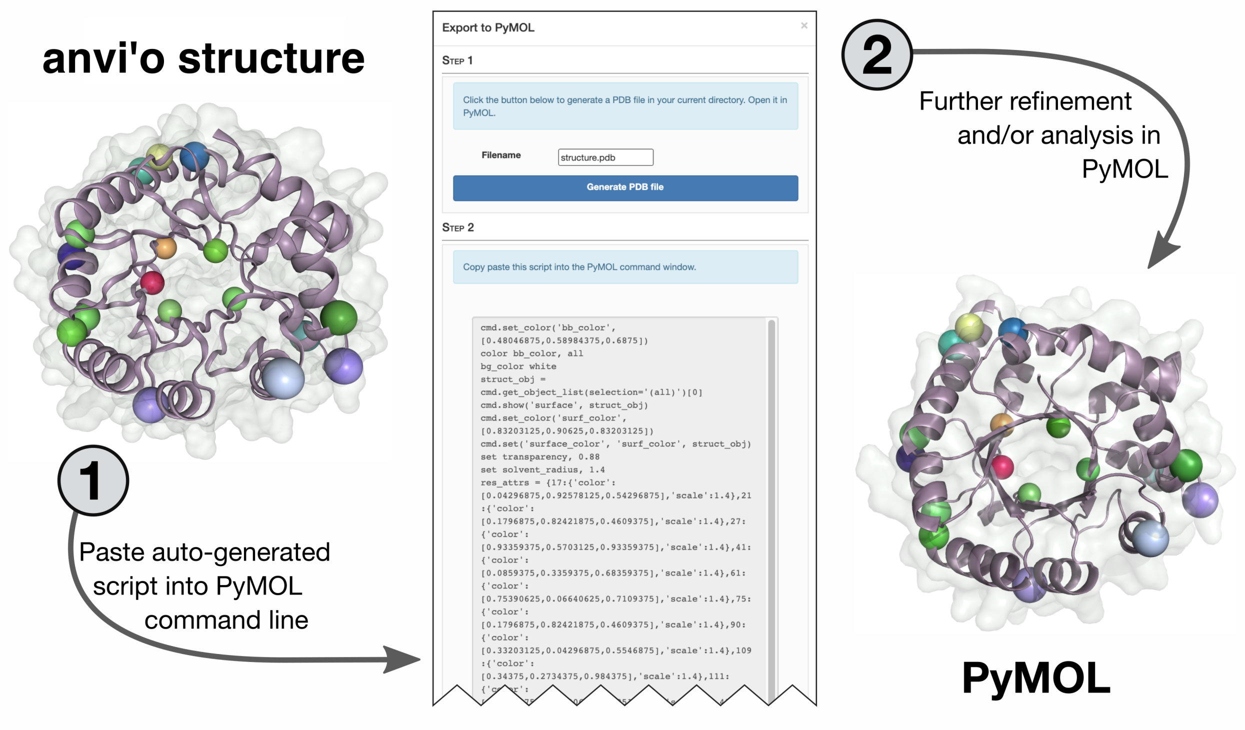
Maximum

Value: 123

Size: 1.4

Structure

Figure 4.27: Screenshot of the interface with the ”Views” tab active. In the “Views” tab, variants can be colored and sized according to variables. In this screenshot, the user has colored variants according to their entropy values on a linear gradient between white and red, and sized them according to their metagenomic coverage values.


anvi'o structure
1
Paste auto-generated script into PyMOL command line
2
Further refinement and/or analysis in PyMOL
PyMOL
Export to PyMOL
×
Step 1
Filename
structure.pdb
Generate PDB file
Step 2

Click the button below to generate a PDB file in your current directory. Open it in PyMOL.

Copy paste this script into the PyMOL command window.

```
cmd.set_color('bb_color',
[0.48046875,0.58984375,0.6875])
color bb_color, all
bg_color white
struct_obj =
cmd.get_object_list(selection='(all)')[0]
cmd.show('surface', struct_obj)
cmd.set_color('surf_color',
[0.83203125,0.90625,0.83203125])
cmd.set('surface_color', 'surf_color', struct_obj)
set transparency, 0.88
set solvent_radius, 1.4
res_attrs = {17:{'color':
[0.04296875,0.92578125,0.54296875],'scale':1.4},21
:{'color':
[0.1796875,0.82421875,0.4609375],'scale':1.4},27:
{'color':
[0.93359375,0.5703125,0.93359375],'scale':1.4},41:
{'color':
[0.0859375,0.3359375,0.68359375],'scale':1.4},61:
{'color':
[0.75390625,0.06640625,0.7109375],'scale':1.4},75:
{'color':
[0.1796875,0.82421875,0.4609375],'scale':1.4},90:
{'color':
[0.33203125,0.04296875,0.5546875],'scale':1.4},109
:{'color':
[0.34375,0.2734375,0.984375],'scale':1.4},111:
{'color':
```

Figure 4.28: The interface can seamlessly migrate user sessions into PyMOL for visual refinement and more sophisticated analysis than is possible with `anvi-display-structure`. Under the "Output" tab, users can select "Generate in PyMOL" to auto-generate a script (middle) that when pasted into the PyMOL command line, reproduces the current interface view directly in PyMOL.

CHAPTER 5
CONCLUSION

5.1 Summary of contributions

The processes governing the emergence and maintenance of genetic diversity are fascinatingly complex and provide a fingerprint for studying how dynamic environments constantly remodel the fitness landscape and how microbial populations consequently cope. My work has furthered this research avenue by including structure-informed analyses into the realm of microbial population genetics. This work showcases how patterns of polymorphism seen in natural microbial populations are much more interpretable when viewed from the context of predicted protein structures and binding sites. Using structural features, I illustrated how polymorphism in natural populations shift in response to environmentally-mediated selection strength. Using an abundant marine microbe (SAR11) as a model system, 17% of nonsynonymous polymorphism data was explainable using just relative solvent accessibility and proximity to active site. In glutamine synthetase, a central gene in nitrogen metabolism, it was found that nitrogen limitation governs selective pressures that define how close nonsynonymous polymorphism is 'allowed' to get to the binding site. This is an *in situ* observation of conditional neutrality, in which alleles are permissible in one environment but selected against in another. Finally, this study illustrates how structure governs polymorphism that is synonymous, but not silent: rare synonymous codons are systematically purified out of important protein sites when genes are under high selection, yet encroach upon these sites when selection is low. Although broadly speaking it is commonplace to utilize protein structures in evolutionary analysis, microbial population genetics has been late to adopt these practices, in part because there are no tools that enable interactive browsing of polymorphism at the scale of data required for metagenomic analyses. To fill this gap so others may increase the interpretability of their data, my thesis work includes the development of anvi'o structure, which is an automated, scalable, and interactive work environment for analyzing and visualizing metagenomic variants with respect to predicted protein structures and ligand binding sites. This software trivializes many tasks that took me years to streamline and provides a straightforward path for researchers to undertake their own structure-informed analyses of genetic variation within natural microbial populations.

5.2 Future directions

My studies have focused exclusively on a single marine microbe, SAR11, which was used as a model system to illustrate how fruitful the intersection between structural bioinformatics and microbial ecology can be. If I was given more time, I think a more extensive investigation spanning more environments and more genomes would be practical and rich with discovery. Having already laid the groundwork, an analysis with thousands of populations rather than just one is within arm's reach. In fact, many already existing analyses could be recycled by referencing the step-by-step reproducible workflow I created for [Kiefl et al., 2022] (see `https://merenlab.org/data/anvio-structure`).

On a separate but related note, microbial ecologists rely on short read alignment software tools such as Bowtie2 [Langmead and Salzberg, 2012] and BWA [Li and Durbin, 2009] in order to identify single nucleotide variants. Although these alignment softwares provide the raw data required for quantifying insertions/deletions (INDELs) relative to the reference sequence, I am not aware of researchers taking advantage of these potentially very important genetic structural changes found in environmental populations. During my studies I redesigned anvi'o to summarize and store the raw INDEL information in metagenomic read recruitment results, however as far as I am aware, no anvi'o users are utilizing this data. I intuit that a proper analysis of this data, perhaps in the project proposed above, would yield low-hanging fruit.

REFERENCES

Richard J Abdill, Elizabeth M Adamowicz, and Ran Blekhman. Public human microbiome data are dominated by highly developed countries. *PLoS Biol.*, 20(2):e3001536, February 2022.

Silvia G Acinas, Vanja Klepac-Ceraj, Dana E Hunt, Chanathip Pharino, Ivica Ceraj, Daniel L Distel, and Martin F Polz. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430(6999):551–554, July 2004.

Deepa Agashe, N Cecilia Martinez-Gomez, D Allan Drummond, and Christopher J Marx. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol. Biol. Evol.*, 30(3):549–560, March 2013.

Eric E Allen, Gene W Tyson, Rachel J Whitaker, John C Detter, Paul M Richardson, and Jillian F Banfield. Genome dynamics in a natural archaeal population. *Proc. Natl. Acad. Sci. U. S. A.*, 104(6):1883–1888, February 2007.

Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S Pollard, Ekaterina Sakharova, Donovan H Parks, Philip Hugenholtz, Nicola Segata, Nikos C Kyrpides, and Robert D Finn. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, 39(1): 105–114, January 2021.

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.

Karthik Anantharaman, Christopher T Brown, Laura A Hug, Itai Sharon, Cindy J Castelle, Alexander J Probst, Brian C Thomas, Andrea Singh, Michael J Wilkins, Ulas Karaoz, Eoin L Brodie, Kenneth H Williams, Susan S Hubbard, and Jillian F Banfield. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.*, 7:13219, October 2016.

Rika E Anderson, Julie Reveillaud, Emily Reddington, Tom O Delmont, A Murat Eren, Jill M McDermott, Jeff S Seewald, and Julie A Huber. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. *Nat. Commun.*, 8(1):1114, October 2017.

C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096): 223–230, July 1973.

Adam P Arkin, Robert W Cottingham, Christopher S Henry, Nomi L Harris, Rick L Stevens, Sergei Maslov, Paramvir Dehal, Doreen Ware, Fernando Perez, Shane Canon, Michael W Sneddon, Matthew L Henderson, William J Riehl, Dan Murphy-Olson, Stephen Y Chan, Roy T Kamimura, Sunita Kumari, Meghan M Drake, Thomas S Brettin, Elizabeth M Glass, Dylan Chivian, Dan Gunter, David J Weston, Benjamin H Allen, Jason Baumohl, Aaron A Best, Ben Bowen, Steven E Brenner, Christopher C Bun, John-Marc Chandonia, Jer-Ming Chia, Ric Colasanti, Neal Conrad, James J Davis, Brian H Davison, Matthew DeJongh, Scott Devoid, Emily Dietrich, Inna Dubchak, Janaka N Edirisinghe, Gang Fang, José P Faria, Paul M Frybarger, Wolfgang Gerlach, Mark Gerstein, Annette Greiner, James Gurtowski, Holly L Haun, Fei He, Rashmi Jain, Marcin P Joachimiak, Kevin P Keegan, Shinnosuke Kondo, Vivek Kumar, Miriam L Land, Folker Meyer, Marissa Mills, Pavel S Novichkov, Taeyun Oh, Gary J Olsen, Robert Olson, Bruce Parrello, Shiran Pasternak, Erik Pearson, Sarah S Poon, Gavin A Price, Srividya Ramakrishnan, Priya Ranjan, Pamela C Ronald, Michael C Schatz, Samuel M D Seaver, Maulik Shukla, Roman A Sutormin, Mustafa H Syed, James Thomason, Nathan L Tintle, Daifeng Wang, Fangfang Xia, Hyunseung Yoo, Shinjae Yoo, and Dantong Yu. KBase: The united states department of energy systems biology knowledgebase. *Nat. Biotechnol.*, 36(7):566–569, July 2018.

L G M Baas Becking. *Geobiologie of inleiding tot de milieukunde.* W.P. Van Stockum & Zoon, Den Haag, 1934.

Yinon M Bar-On, Rob Phillips, and Ron Milo. The biomass distribution on earth. *Proc. Natl. Acad. Sci. U. S. A.*, 115(25):6506–6511, June 2018.

Matthew L Bendall, Sarah Lr Stevens, Leong-Keat Chan, Stephanie Malfatti, Patrick Schwientek, Julien Tremblay, Wendy Schackwitz, Joel Martin, Amrita Pati, Brian Bushnell, Jeff Froula, Dongwan Kang, Susannah G Tringe, Stefan Bertilsson, Mary A Moran, Ashley Shade, Ryan J Newton, Katherine D McMahon, and Rex R Malmstrom. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.*, 10(7):1589–1601, July 2016.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 57(1):289–300, January 1995.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, January 2000.

Stéphanie M Bernard and Dimah Z Habash. The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling. *New Phytol.*, 182(3):608–620, 2009.

F C Bernstein, T F Koetzle, G J Williams, E F Meyer, Jr, M D Brice, J R Rodgers, O Kennard, T Shimanouchi, and M Tasumi. The protein data bank. a computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, 80(2):319–324, November 1977.

D Bordo and P Argos. Suggestions for “safe” residue substitutions in site-directed mutagenesis. *J. Mol. Biol.*, 217(4):721–729, February 1991.

Laura A Bristow, Wiebke Mohr, Soeren Ahmerkamp, and Marcel M M Kuypers. Nutrients that limit growth in the ocean. *Curr. Biol.*, 27(11):R474–R478, June 2017.

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

Mark V Brown, Federico M Lauro, Matthew Z DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin J Riddle, Jed A Fuhrman, Cynthia Andrews-Pfannkoch, Jeffrey M Hoffman, Jeffrey B McQuaid, Andrew Allen, Stephen R Rintoul, and Ricardo Cavicchioli. Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.*, 8:595, July 2012.

Eric S Brucks. Metagenome recruitment of the global ocean survey dataset to four closely-related sar11 genomes. Master's thesis, University of Hawaii at Mānoa, 2014.

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND, 2015.

Molly K Burke, Joseph P Dunham, Parvin Shahrestani, Kevin R Thornton, Michael R Rose, and Anthony D Long. Genome-wide analysis of a long-term evolution experiment with drosophila. *Nature*, 467(7315):587–590, September 2010.

C D Bustamante, J P Townsend, and D L Hartl. Solvent accessibility and purifying selection within proteins of escherichia coli and salmonella enterica. *Mol. Biol. Evol.*, 17(2):301–308, February 2000.

Alison Callahan, Rainer Winnenburg, and Nigam H Shah. U-Index, a dataset and an impact metric for informatics tools and databases. *Sci Data*, 5:180043, March 2018.

Craig A Carlson, Robert Morris, Rachel Parsons, Alexander H Treusch, Stephen J Giovannoni, and Kevin Vergin. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern sargasso sea. *ISME J.*, 3(3):283–295, March 2009.

Alejandro Caro-Quintero and Konstantinos T Konstantinidis. Bacterial species may exist, metagenomics reveal. *Environ. Microbiol.*, 14(2):347–355, February 2012.

Huiling Chen and Huan-Xiang Zhou. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.*, 33(10):3193–3199, June 2005.

K Chen and F H Arnold. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U. S. A.*, 90(12):5618–5622, June 1993.

Lin-Xing Chen, Karthik Anantharaman, Alon Shaiber, A Murat Eren, and Jillian F Banfield. Accurate and complete genomes from metagenomes. *Genome Res.*, 30(3):315–333, March 2020.

Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel

J L de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009.

Frederick M Cohan. Systematics: The cohesive nature of bacterial species taxa. *Curr. Biol.*, 29(5):R169–R172, March 2019.

Frederick M Cohan and Elizabeth B Perry. A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol.*, 17(10):R373–86, May 2007.

Maureen L Coleman, Matthew B Sullivan, Adam C Martiny, Claudia Steglich, Kerrie Barry, Edward F Delong, and Sallie W Chisholm. Genomic islands and the ecology and evolution of prochlorococcus. *Science*, 311(5768):1768–1770, March 2006.

Arolyn Conwill, Anne C Kuan, Ravalika Damerla, Alexandra J Poret, Jacob S Baker, A Delphine Tripp, Eric J Alm, and Tami D Lieberman. Anatomy promotes neutral coexistence of strains in the human skin microbiome. *Cell Host Microbe*, January 2022.

Paul Igor Costea, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork. metaSNV: A tool for metagenomic strain level analysis. *PLoS One*, 12(7): e0182392, July 2017.

T P Curtis and W T Sloan. Microbiology. exploring microbial diversity–a vast below. *Science*, 309(5739):1331–1333, August 2005.

Thomas P Curtis, Ian M Head, Mary Lunn, Stephen Woodcock, Patrick D Schloss, and William T Sloan. What is the extent of prokaryotic diversity? *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 361(1475):2023–2037, November 2006.

Charles Darwin. *On the Origin of Species by Means of Natural Selection.* Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

Antony M Dean, Claudia Neuhauser, Elise Grenier, and G Brian Golding. The pattern of amino acid replacements in alpha/beta-barrels. *Mol. Biol. Evol.*, 19(11):1846–1864, November 2002.

DELANO and W. L. The PyMOL molecular graphics system. *http://www.pymol.org*, 2002.

Tom O Delmont and A Murat Eren. Linking pangenomes and metagenomes: the prochlorococcus metapangenome. *PeerJ*, 6:e4320, January 2018.

Tom O Delmont, Christopher Quince, Alon Shaiber, Özcan C Esen, Sonny Tm Lee, Michael S Rappé, Sandra L McLellan, Sebastian Lücker, and A Murat Eren. Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol*, 3(7):804–813, July 2018.

Tom O Delmont, Evan Kiefl, Ozsel Kilinc, Ozcan C Esen, Ismail Uysal, Michael S Rappé, Steven Giovannoni, and A Murat Eren. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife*, 8, September 2019.

Vincent J Denef. Peering into the genetic makeup of natural microbial populations using metagenomics. In Martin F Polz and Om P Rajora, editors, *Population Genomics: Microorganisms*, pages 49–75. Springer International Publishing, Cham, 2019.

Gregory J Dick, Karthik Anantharaman, Brett J Baker, Meng Li, Daniel C Reed, and Cody S Sheik. The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Front. Microbiol.*, 4:124, May 2013.

Domingo Esteban, Sheldon Julie, and Perales Celia. Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, 76(2):159–216, June 2012.

D Allan Drummond and Claus O Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–352, July 2008.

D Allan Drummond, Jesse D Bloom, Christoph Adami, Claus O Wilke, and Frances H Arnold. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.*, 102 (40):14338–14343, October 2005.

Julian Echave, Stephanie J Spielman, and Claus O Wilke. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.*, 17(2):109–121, February 2016.

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

Robert C Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, August 2004.

Alexander Eiler, Darin H Hayakawa, Matthew J Church, David M Karl, and Michael S Rappé. Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic north pacific subtropical gyre. *Environ. Microbiol.*, 11(9): 2291–2300, September 2009.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The pfam protein families database in 2019. *Nucleic Acids Res.*, 47 (D1):D427–D432, January 2019.

Hans Ellegren, Nick G C Smith, and Matthew T Webster. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.*, 13(6):562–568, December 2003.

A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, April 2002.

A Murat Eren, Loïs Maignien, Woo Jun Sul, Leslie G Murphy, Sharon L Grim, Hilary G Morrison, and Mitchell L Sogin. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.*, 4(12), December 2013.

A Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E Miller, Matthew S Schechter, Isaac Fink, Jessica N Pan, Mahmoud Yousef, Emily C Fogarty, Florian Trigodet, Andrea R Watson, Özcan C Esen, Ryan M Moore, Quentin Clayssen, Michael D Lee, Veronika Kivenson, Elaina D Graham, Bryan D Merrill, Antti Karkman, Daniel Blankenberg, John M Eppley, Andreas Sjödin, Jarrod J Scott, Xabier Vázquez-Campos, Luke J McKay, Elizabeth A McDaniel, Sarah L R Stevens, Rika E Anderson, Jessika Fuessel, Antonio Fernandez-Guerra, Lois Maignien, Tom O Delmont, and Amy D Willis. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*, 6(1):3–6, January 2021.

Yong Fan and Oluf Pedersen. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.*, 19(1):55–71, January 2021.

Scott S Farley, Andria Dawson, Simon J Goring, and John W Williams. Situating ecology as a Big-Data science: Current advances, challenges, and solutions. *Bioscience*, 68(8): 563–576, July 2018.

Matteo P Ferla, J Cameron Thrash, Stephen J Giovannoni, and Wayne M Patrick. New rRNA gene-based phylogenies of the alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. *PLoS One*, 8(12):e83383, December 2013.

C B Field, M J Behrenfeld, J T Randerson, and P Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374):237–240, July 1998.

K G Field, D Gordon, T Wright, M Rappé, E Urback, K Vergin, and S J Giovannoni. Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl. Environ. Microbiol.*, 63(1):63–70, January 1997.

Sarahi L Garcia, Sarah L R Stevens, Benjamin Crary, Manuel Martinez-Garcia, Ramunas Stepanauskas, Tanja Woyke, Susannah G Tringe, Siv G E Andersson, Stefan Bertilsson, Rex R Malmstrom, and Katherine D McMahon. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *ISME J.*, 12(3):742–755, March 2018.

Nandita R Garud and Katherine S Pollard. Population genetics in the human microbiome. *Trends Genet.*, 36(1):53–67, January 2020.

Nandita R Garud, Benjamin H Good, Oskar Hallatschek, and Katherine S Pollard. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.*, 17(1):e3000102, January 2019.

Cedric Ginestet. Ggplot2: Elegant graphics for data analysis. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 174(1):245–246, January 2011a.

Cedric Ginestet. ggplot2: Elegant graphics for data analysis: Book reviews. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 174(1):245–246, January 2011b.

S J Giovannoni, T B Britschgi, C L Moyer, and K G Field. Genetic diversity in sargasso sea bacterioplankton. *Nature*, 345(6270):60–63, May 1990.

Stephen J Giovannoni. SAR11 bacteria: The most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.*, 9:231–255, January 2017.

Stephen J Giovannoni, Lisa Bibbs, Jang-Cheon Cho, Martha D Stapels, Russell Desiderio, Kevin L Vergin, Michael S Rappé, Samuel Laney, Lawrence J Wilhelm, H James Tripp, Eric J Mathur, and Douglas F Barofsky. Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature*, 438(7064):82–85, November 2005.

G B Golding and A M Dean. The structural basis of molecular adaptation. *Mol. Biol. Evol.*, 15(4):355–369, April 1998.

Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678): 45–50, November 2017.

David S Goodsell and and Arthur J Olson. Structural symmetry and protein function. November 2003.

Jana Grote, J Cameron Thrash, Megan J Huggett, Zachary C Landry, Paul Carini, Stephen J Giovannoni, and Michael S Rappé. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio*, 3(5), September 2012.

Michael J Harms and Joseph W Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, 14(8):559–571, August 2013.

Jose M Haro-Moreno, Francisco Rodriguez-Valera, Riccardo Rosselli, Francisco Martinez-Hernandez, Juan J Roda-Garcia, Monica Lluesma Gomez, Oscar Fornas, Manuel Martinez-Garcia, and Mario López-Pérez. Ecogenomics of the SAR11 clade. *Environ. Microbiol.*, 22(5):1748–1763, May 2020.

Ferdi L Hellweger, Erik van Sebille, and Neil D Fredrick. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science*, 345(6202):1346–1349, September 2014.

Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nat Microbiol*, 1: 16048, April 2016.

Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012a.

Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, June 2012b.

Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010.

Michael Imelfort, Donovan Parks, Ben J Woodcroft, Paul Dennis, Philip Hugenholtz, and Gene W Tyson. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603, September 2014.

Benjamin R Jack, Austin G Meyer, Julian Echave, and Claus O Wilke. Functional sites induce Long-Range evolutionary constraints in enzymes. *PLoS Biol.*, 14(5):e1002452, May 2016.

Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, 9(1):5114, November 2018.

Vahid Jalili, Enis Afgan, Qiang Gu, Dave Clements, Daniel Blankenberg, Jeremy Goecks, James Taylor, and Anton Nekrutenko. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.*, 48(W1):W395–W402, July 2020.

H W Jannasch and M J Mottl. Geomicrobiology of deep-sea hydrothermal vents. *Science*, 229(4715):717–725, August 1985.

B Jesse Shapiro. What microbial population genomics has taught us about speciation. In *Population Genomics: Microorganisms*. unknown, January 2018.

Bror F Jönsson and James R Watson. The timescales of global surface-ocean connectivity. *Nat. Commun.*, 7:11239, April 2016.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

Lukas Käll, Anders Krogh, and Erik L L Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, 338(5):1027–1036, May 2004.

Lukas Käll, Anders Krogh, and Erik L L Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction–the phobius web server. *Nucleic Acids Res.*, 35(Web Server issue):W429–32, July 2007.

Morten Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, 7(8):1511–1522, July 2012.

Nadav Kashtan, Sara E Roggensack, Sébastien Rodrigue, Jessie W Thompson, Steven J Biller, Allison Coe, Huiming Ding, Pekka Marttinen, Rex R Malmstrom, Roman Stocker, Michael J Follows, Ramunas Stepanauskas, and Sallie W Chisholm. Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science*, 344(6182): 416–420, April 2014.

Zhi Y Kho and Sunil K Lal. The human gut microbiome - a potential controller of wellness and disease. *Front. Microbiol.*, 9:1835, August 2018.

Evan Kiefl, Ozcan C Esen, Samuel E Miller, Kourtney L Kroll, Amy D Willis, Michael S Rappé, Tao Pan, and A Murat Eren. Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. March 2022.

Ozsel Kilinc and Ismail Uysal. Auto-clustering output layer: Automatic learning of latent annotations in neural networks. February 2017.

Ozsel Kilinc and Ismail Uysal. Learning latent representations in neural networks for clustering through pseudo supervision and graph-based activity regularization. February 2018.

Alan K Knapp, Melinda D Smith, Sarah E Hobbie, Scott L Collins, Timothy J Fahey, Gretchen J A Hansen, Douglas A Landis, Kimberly J La Pierre, Jerry M Melillo, Timothy R Seastedt, Gaius R Shaver, and Jackson R Webster. Past, present, and future roles of Long-Term experiments in the LTER network. *Bioscience*, 62(4):377–389, April 2012.

Shilpa Nadimpalli Kobren and Mona Singh. Systematic domain-based aggregation of protein structures highlights DNA-, RNA- and other ligand-binding positions. *Nucleic Acids Res.*, 47(2):582–593, January 2019.

Anton A Komar. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.*, 34(1):16–24, January 2009.

Konstantinos T Konstantinidis and Edward F DeLong. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.*, 2(10):1052–1065, October 2008.

Konstantinos T Konstantinidis and James M Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, 102(7):2567–2572, February 2005.

Anna Kopf, Mesude Bicak, Renzo Kottmann, Julia Schnetzer, Ivaylo Kostadinov, Katja Lehmann, Antonio Fernandez-Guerra, Christian Jeanthon, Eyal Rahav, Matthias Ullrich, Antje Wichels, Gunnar Gerdts, Paraskevi Polymenakou, Giorgos Kotoulas, Rania Siam, Rehab Z Abdallah, Eva C Sonnenschein, Thierry Cariou, Fergal O'Gara, Stephen Jackson, Sandi Orlic, Michael Steinke, Julia Busch, Bernardo Duarte, Isabel Caçador, João Canning-Clode, Oleksandra Bobrova, Viggo Marteinsson, Eyjolfur Reynisson, Clara Magalhães Loureiro, Gian Marco Luna, Grazia Marina Quero, Carolin R Löscher, Anke Kremp, Marie E DeLorenzo, Lise Øvreås, Jennifer Tolman, Julie LaRoche, Antonella Penna, Marc Frischer, Timothy Davis, Barker Katherine, Christopher P Meyer, Sandra Ramos, Catarina Magalhães, Florence Jude-Lemeilleur, Ma Leopoldina Aguirre-Macedo, Shiao Wang, Nicole Poulton, Scott Jones, Rachel Collin, Jed A Fuhrman, Pascal Conan, Cecilia Alonso, Noga Stambler, Kelly Goodwin, Michael M Yakimov, Federico Baltar, Levente Bodrossy, Jodie Van De Kamp, Dion Mf Frampton, Martin Ostrowski, Paul Van Ruth, Paul Malthouse, Simon Claus, Klaas Deneudt, Jonas Mortelmans, Sophie Pitois, David Wallom, Ian Salter, Rodrigo Costa, Declan C Schroeder, Mahrous M Kandil, Valentina Amaral, Florencia Biancalana, Rafael Santana, Maria Luiza Pedrotti, Takashi Yoshida, Hiroyuki Ogata, Tim Ingleton, Kate Munnik, Naiara Rodriguez-Ezpeleta, Veronique Berteaux-Lecellier, Patricia Wecker, Ibon Cancio, Daniel Vaulot, Christina Bienhold, Hassan Ghazal, Bouchra Chaouni, Soumya Essayeh, Sara Ettamimi, El Houcine Zaid, Noureddine Boukhatem, Abderrahim Bouali, Rajaa Chahboune, Said Barrijal, Mohammed Timinouni, Fatima El Otmani, Mohamed Bennani, Marianna Mea, Nadezhda Todorova, Ventzislav Karamfilov, Petra Ten Hoopen, Guy Cochrane, Stephane L'Haridon, Kemal Can Bizsel, Alessandro Vezzi, Federico M Lauro, Patrick Martin, Rachelle M Jensen, Jamie Hinks, Susan Gebbels, Riccardo Rosselli, Fabio De Pascale, Riccardo Schiavon, Antonina Dos Santos, Emilie Villar, Stéphane Pesant, Bruno Cataletto, Francesca Malfatti, Ranjith Edirisinghe, Jorge A Herrera Silveira, Michele Barbier, Valentina Turk, Tinkara Tinta, Wayne J Fuller, Ilkay Salihoglu, Nedime Serakinci, Mahmut Cerkez Ergoren, Eileen Bresnan, Juan Iriberri, Paul Anders Fronth Nyhus, Edvardsen Bente, Hans Erik Karlsen, Peter N Golyshin, Josep M Gasol, Snejana Moncheva, Nina Dzhembekova, Zackary Johnson, Christopher David Sinigalliano, Maribeth Louise Gidley, Adriana Zingone, Roberto Danovaro, George Tsiamis, Melody S Clark, Ana Cristina Costa, Monia El Bour, Ana M Martins, R Eric Collins, Anne-Lise Ducluzeau, Jonathan Martinez, Mark J Costello, Linda A Amaral-Zettler, Jack A Gilbert, Neil Davies, Dawn Field, and Frank Oliver Glöckner. The ocean sampling day consortium. *Gigascience*, 4:27, June 2015.

Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012.

Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.*, 20(11):681–697, November 2019.

Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: a LLVM-based python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, number Article 7 in LLVM '15, pages 1–6, New York, NY, USA, November 2015. Association for Computing Machinery.

E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

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

Richard E Lenski, Michael R Rose, Suzanne C Simpson, and Scott C Tadler. Long-Term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138(6):1315–1341, December 1991.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

Timothy W Lyons, Christopher T Reinhard, and Noah J Planavsky. The rise of oxygen in earth's early ocean and atmosphere. *Nature*, 506(7488):307–315, February 2014.

Julieta M Manrique and Leandro R Jones. Are ocean currents too slow to counteract SAR11 evolution? a next-generation sequencing, phylogeographic analysis. *Mol. Phylogenet. Evol.*, 107:324–337, February 2017.

Lauren J McIver, Galeb Abu-Ali, Eric A Franzosa, Randall Schwager, Xochitl C Morgan, Levi Waldron, Nicola Segata, and Curtis Huttenhower. biobakery: a meta'omic analysis environment, 2018.

Ted H M Mes. Microbial diversity–insights from population genetics. *Environ. Microbiol.*, 10(1):251–264, January 2008.

Alexandra Meziti, Despina Tsementzi, Luis M Rodriguez-R, Janet K Hatt, Hera Karayanni, Konstantinos A Kormas, and Konstantinos T Konstantinidis. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. *ISME J.*, 13(3):767–779, March 2019.

André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biol.*, 12(11):R112, November 2011.

Robert M Morris, Michael S Rappé, Stephanie A Connon, Kevin L Vergin, William A Siebold, Craig A Carlson, and Stephen J Giovannoni. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420(6917):806–810, 2002.

A Murat Eren, Özcan C Esen, Christopher Quince, Joseph H Vineis, Hilary G Morrison, Mitchell L Sogin, and Tom O Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, October 2015.

Stephen Nayfach, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S Pollard. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.*, 26(11):1612–1625, November 2016.

NIH. DNA sequencing costs: Data. `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`. Accessed: 2022-3-22.

Stephen E Noell and Stephen J Giovannoni. SAR11 bacteria have a high affinity and multifunctional glycine betaine transporter. *Environ. Microbiol.*, 21(7):2559–2575, July 2019.

Howard Ochman. Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.*, 20(12):2091–2096, December 2003.

Hyun-Myung Oh, Ilnam Kang, Kiyoung Lee, Yoonra Jang, Seung-Il Lim, and Jang-Cheon Cho. Complete genome sequence of strain IMCC9063, belonging to SAR11 subgroup 3, isolated from the arctic ocean. *J. Bacteriol.*, 193(13):3379–3380, July 2011a.

Seungdae Oh, Alejandro Caro-Quintero, Despina Tsementzi, Natasha DeLeon-Rodriguez, Chengwei Luo, Rachel Poretsky, and Konstantinos T Konstantinidis. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of lake lanier, a temperate freshwater ecosystem. *Appl. Environ. Microbiol.*, 77(17):6000–6011, September 2011b.

T Ohta and J H Gillespie. Development of neutral and nearly neutral theories. *Theor. Popul. Biol.*, 49(2):128–142, April 1996.

Matthew R Olm, Alexander Crits-Christoph, Keith Bouma-Gregson, Brian A Firek, Michael J Morowitz, and Jillian F Banfield. instrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.*, 39(6):727–736, June 2021.

G J Olsen, D J Lane, S J Giovannoni, N R Pace, and D A Stahl. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.*, 40:337–365, 1986.

Maria G Pachiadaki, Julia M Brown, Joseph Brown, Oliver Bezuidt, Paul M Berube, Steven J Biller, Nicole J Poulton, Michael D Burkart, James J La Clair, Sallie W Chisholm, and Ramunas Stepanauskas. Charting the complexity of the marine microbiome through Single-Cell genomics. *Cell*, 179(7):1623–1635.e11, December 2019.

Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew T G Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, November 2015.

C Pál, B Papp, and L D Hurst. Highly expressed genes in yeast evolve slowly. *Genetics*, 158 (2):927–931, June 2001.

Lucas Paoli, Hans-Joachim Ruscheweyh, Clarissa C Forneris, Satria Kautsar, Quentin Clayssen, Guillem Salazar, Alessio Milanese, Daniel Gehrig, Martin Larralde, Laura M Carroll, Pablo Sánchez, Ahmed A Zayed, Dylan R Cronin, Silvia G Acinas, Peer Bork, Chris Bowler, Tom O Delmont, Matthew B Sullivan, Patrick Wincker, Georg Zeller, Serina L Robinson, Jörn Piel, and Shinichi Sunagawa. Uncharted biosynthetic potential of the ocean microbiome. March 2021.

Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7):1043–1055, July 2015.

Joshua B Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, 12(1):32–42, January 2011.

Leighton Pritchard, Rachel H Glover, Sonia Humphris, John G Elphinstone, and Ian K Toth. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods*, 8(1):12–24, December 2015.

Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(9):833–844, September 2017.

R R Development Core Team. R: A language and environment for statistical computing, 2011.

Michael S Rappé, Stephanie A Connon, Kevin L Vergin, and Stephen J Giovannoni. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature*, 418(6898):630–633, August 2002.

Julie Reveillaud, Sarah R Bordenstein, Corinne Cruaud, Alon Shaiber, Özcan C Esen, Mylène Weill, Patrick Makoundou, Karen Lolans, Andrea R Watson, Ignace Rakotoarivony, Seth R Bordenstein, and A Murat Eren. The wolbachia mobilome in culex pipiens includes a putative plasmid. *Nat. Commun.*, 10(1):1051, March 2019.

Eduardo P C Rocha. Neutral theory, microbial practice: Challenges in bacterial population genetics. *Mol. Biol. Evol.*, 35(6):1338–1347, June 2018.

Naiara Rodríguez-Ezpeleta and T Martin Embley. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PLoS One*, 7(1):e30520, January 2012.

Alexander S Rose and Peter W Hildebrand. NGL viewer: a web application for molecular visualization. *Nucleic Acids Res.*, 43(W1):W576–9, July 2015.

Alexander S Rose, Anthony R Bradley, Yana Valasatava, Jose M Duarte, Andreas Prlić, and Peter W Rose. Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology*, Web3D '16, pages 185–186, New York, NY, USA, July 2016. Association for Computing Machinery.

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

Guillem Salazar, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Christopher M Field, Luis Pedro Coelho, Corinne Cruaud, Stefan Engelen, Ann C Gregory, Karine Labadie, Claudie Marec, Eric Pelletier, Marta Royo-Llonch, Simon Roux, Pablo Sánchez, Hideya Uehara, Ahmed A Zayed, Georg Zeller, Margaux Carmichael, Céline Dimier, Joannie Ferland, Stefanie Kandels, Marc Picheral, Sergey Pisarev, Julie Poulain, Tara Oceans Coordinators, Silvia G Acinas, Marcel Babin, Peer Bork, Chris Bowler, Colomban de Vargas, Lionel Guidi, Pascal Hingamp, Daniele Iudicone, Lee Karp-Boss, Eric Karsenti, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Matthew B Sullivan, Patrick Wincker, and Shinichi Sunagawa. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179(5):1068–1083.e21, November 2019.

Siegfried Schloissnig, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, Daniel R Mende, Jens Roat Kultima, John Martin, Karthik Kota, Shamil R Sunyaev, George M Weinstock, and Peer Bork. Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50, January 2013.

Matthias Scholz, Doyle V Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L Morrow, and Nicola Segata. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods*, 13(5):435–438, May 2016.

Daniel R Schrider, Jonathan N Hourmozdi, and Matthew W Hahn. Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.*, 21(12):1051–1054, June 2011.

Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.*, 14(8):e1002533, August 2016.

Alon Shaiber, Amy D Willis, Tom O Delmont, Simon Roux, Lin-Xing Chen, Abigail C Schmid, Mahmoud Yousef, Andrea R Watson, Karen Lolans, Özcan C Esen, Sonny T M Lee, Nora Downey, Hilary G Morrison, Floyd E Dewhirst, Jessica L Mark Welch, and A Murat Eren. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, 21(1):292, December 2020.

B Jesse Shapiro and Martin F Polz. Microbial speciation. *Cold Spring Harb. Perspect. Biol.*, 7(10):a018143, September 2015.

Avital Sharir-Ivry and Yu Xia. Non-catalytic binding sites induce weaker Long-Range evolutionary rate gradients than catalytic sites in enzymes. *J. Mol. Biol.*, 431(19):3860–3870, September 2019.

Avital Sharir-Ivry and Yu Xia. Quantifying evolutionary importance of protein sites: A tale of two measures. *PLoS Genet.*, 17(4):e1009476, April 2021.

Liat Shenhav and David Zeevi. Resource conservation manifests in the genetic code. *Science*, 370(6517):683–687, November 2020.

Tobias Sikosek and Hue Sun Chan. Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface*, 11(100):20140419, November 2014.

Jessica Siltberg-Liberles, Johan A Grahnen, and David A Liberles. The evolution of protein structures and structural ensembles under functional constraint. *Genes*, 2(4):748–762, October 2011.

Sheri L Simmons, Genevieve Dibartolo, Vincent J Denef, Daniela S Aliaga Goltsman, Michael P Thelen, and Jillian F Banfield. Population genomic analysis of strain variation in leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol.*, 6(7):e177, July 2008.

Smith Daniel P., Thrash J. Cameron, Nicora Carrie D., Lipton Mary S., Burnum-Johnson Kristin E., Carini Paul, Smith Richard D., Giovannoni Stephen J., and McFall-Ngai Margaret J. Proteomic and transcriptomic analyses of "candidatus pelagibacter ubique" describe the first PII-Independent response to nitrogen limitation in a Free-Living alphaproteobacterium. *MBio*, 4(6):e00133–12.

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". *Proc. Natl. Acad. Sci. U. S. A.*, 103(32): 12115–12120, August 2006.

Victor Sojo, Christophe Dessimoz, Andrew Pomiankowski, and Nick Lane. Membrane proteins are dramatically less conserved than Water-Soluble proteins across the tree of life. *Mol. Biol. Evol.*, 33(11):2874–2884, November 2016.

M A Sørensen, C G Kurland, and S Pedersen. Codon usage determines translation rate in escherichia coli. *J. Mol. Biol.*, 207(2):365–377, May 1989.

Anja Spang, Jimmy H Saw, Steffen L Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs J G Ettema. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551):173–179, May 2015.

Ulrich Stingl, Harry James Tripp, and Stephen J Giovannoni. Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the oregon coast and the bermuda atlantic time series study site. *ISME J.*, 1(4): 361–371, August 2007.

Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, May 2015.

S Sunyaev, W Lathe, 3rd, and P Bork. Integration of genome data and protein structures: prediction of protein folds, protein interactions and “molecular phenotypes” of single nucleotide polymorphisms. *Curr. Opin. Struct. Biol.*, 11(1):125–130, February 2001.

Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, September 2003.

J Cameron Thrash, Alex Boyd, Megan J Huggett, Jana Grote, Paul Carini, Ryan J Yoder, Barbara Robbertse, Joseph W Spatafora, Michael S Rappé, and Stephen J Giovannoni. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.*, 1:13, June 2011.

Alexander H Treusch, Kevin L Vergin, Liam A Finlay, Michael G Donatz, Robert M Burton, Craig A Carlson, and Stephen J Giovannoni. Seasonality and vertical structure of microbial communities in an ocean gyre. *ISME J.*, 3(10):1148–1163, October 2009.

Duy Tin Truong, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.*, 27(4):626–638, April 2017.

Despina Tsementzi, Jieying Wu, Samuel Deutsch, Sangeeta Nath, Luis M Rodriguez-R, Andrew S Burns, Piyush Ranjan, Neha Sarode, Rex R Malmstrom, Cory C Padilla, Benjamin K Stone, Laura A Bristow, Morten Larsen, Jennifer B Glass, Bo Thamdrup, Tanja Woyke, Konstantinos T Konstantinidis, and Frank J Stewart. SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature*, 536(7615):179–183, August 2016.

Benjamin J Tully, Elaina D Graham, and John F Heidelberg. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data*, 5:170203, January 2018.

Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, March 2004.

Thea Van Rossum, Pamela Ferretti, Oleksandr M Maistrenko, and Peer Bork. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.*, 18(9):491–506, September 2020.

Neha J Varghese, Supratim Mukherjee, Natalia Ivanova, Konstantinos T Konstantinidis, Kostas Mavrommatis, Nikos C Kyrpides, and Amrita Pati. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, 43(14):6761–6771, August 2015.

Kevin L Vergin, H James Tripp, Larry J Wilhelm, Dee R Denver, Michael S Rappé, and Stephen J Giovannoni. High intraspecific recombination rate in a native population of candidatus pelagibacter ubique (SAR11). *Environ. Microbiol.*, 9(10):2430–2440, October 2007.

Johan Viklund, Joran Martijn, Thijs J G Ettema, and Siv G E Andersson. Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. *PLoS One*, 8(11):e78858, November 2013.

Ian M Walsh, Micayla A Bowman, Iker F Soto Santarriaga, Anabel Rodriguez, and Patricia L Clark. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. U. S. A.*, 117(7):3528–3534, February 2020.

Benjamin Webb and Andrej Sali. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics*, 54:5.6.1–5.6.37, June 2016.

Bruce S Weir. Estimating f-statistics: A historical view. *Philos. Sci.*, 79(5):637–643, December 2012.

John H Werren, Laura Baldo, and Michael E Clark. Wolbachia: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.*, 6(10):741–751, October 2008.

Rachel J Whitaker and Jillian F Banfield. Population genomics in natural microbial communities. *Trends Ecol. Evol.*, 21(9):508–516, September 2006.

Angelicque E White, Stephen J Giovannoni, Yanlin Zhao, Kevin Vergin, and Craig A Carlson. Elemental content and stoichiometry of SAR11 chemoheterotrophic marine bacteria. *Limnol. Oceanogr. Lett.*, 4(2):44–51, April 2019.

Richard Allen White, Stephen J Callister, Ronald J Moore, Erin S Baker, and Janet K Jansson. The past, present and future of microbiome analyses. *Nat. Protoc.*, 11(11): 2049–2053, November 2016.

Kerry A Whittaker and Tatiana A Rynearson. Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. *Proc. Natl. Acad. Sci. U. S. A.*, 114(10):2651–2656, March 2017.

Claus O Wilke. Bringing molecules back into molecular evolution. *PLoS Comput. Biol.*, 8 (6):e1002572, June 2012.

Catherine L Worth, Sungsam Gong, and Tom L Blundell. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.*, 10(10):709–720, October 2009.

Tanja Woyke, Devin F R Doud, and Frederik Schulz. The trajectory of microbial single-cell sequencing. *Nat. Methods*, 14(11):1045–1054, October 2017.

Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7):889–895, April 2010.

Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, Fumie Hosoda, Hirofumi Rokutan, Minori Matsumoto, Hiroyuki Takamaru, Masayoshi Yamada, Takahisa Matsuda, Motoki Iwasaki, Taiki Yamaji, Tatsuo Yachida, Tomoyoshi Soga, Ken Kurokawa, Atsushi Toyoda, Yoshitoshi Ogura, Tetsuya Hayashi, Masanori Hatakeyama, Hitoshi Nakagama, Yutaka Saito, Shinji Fukuda, Tatsuhiro Shibata, and Takuji Yamada. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.*, 25(6):968–976, June 2019.

Jianyi Yang, Ambrish Roy, and Yang Zhang. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, 41(Database issue): D1096–103, January 2013.

Carl J Yeoman, Laura M Brutscher, Özcan C Esen, Furkan Ibaoglu, Curtis Fowler, A Murat Eren, Kevin Wanner, and David K Weaver. Genome-resolved insights into a novel spiroplasma symbiont of the wheat stem sawfly (cephus cinctus). *PeerJ*, 7:e7548, August 2019.

Katarzyna Zaremba-Niedzwiedzka, Johan Viklund, Weizhou Zhao, Jennifer Ast, Alexander Sczyrba, Tanja Woyke, Katherina McMahon, Stefan Bertilsson, Ramunas Stepanauskas, and Siv G E Andersson. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol.*, 14(11):R130, November 2013.

E M Zdobnov and R Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848, September 2001.

David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, Jotham Suez, Jemal Ali Mahdi, Elad Matot, Gal Malka, Noa Kosower, Michal Rein, Gili Zilberman-Schapira, Lenka Dohnalová, Meirav Pevsner-Fischer, Rony Bikovsky, Zamir Halpern, Eran Elinav, and Eran Segal. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, November 2015.

She Zhang, James M Krieger, Yan Zhang, Cihan Kaya, Burak Kaynak, Karolina Mikulska-Ruminska, Pemra Doruker, Hongchun Li, and Ivet Bahar. ProDy 2.0: Increased scale and scope after 10 years of protein dynamics modelling with python. *Bioinformatics*, April 2021.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, December 2004.

Shijie Zhao, Tami D Lieberman, Mathilde Poyet, Kathryn M Kauffman, Sean M Gibbons, Mathieu Groussin, Ramnik J Xavier, and Eric J Alm. Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe*, 25(5):656–667.e8, May 2019.

Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, Jun Wang, Floris Imhann, Eelke Brandsma, Soesma A Jankipersadsing, Marie Joossens, Maria Carmen Cenit, Patrick Deelen, Morris A Swertz, LifeLines cohort study, Rinse K Weersma, Edith J M Feskens, Mihai G Netea, Dirk Gevers, Daisy Jonkers, Lude Franke, Yurii S Aulchenko, Curtis Huttenhower, Jeroen Raes, Marten H Hofker, Ramnik J Xavier, Cisca Wijmenga, and Jingyuan Fu. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, April 2016.

Lucie Zinger, Linda A Amaral-Zettler, Jed A Fuhrman, M Claire Horner-Devine, Susan M Huse, David B Mark Welch, Jennifer B H Martiny, Mitchell Sogin, Antje Boetius, and Alban Ramette. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One*, 6(9):e24570, September 2011.