EVOLUTIONARY BIOLOGY

Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution

Evan Kiefl[1,2]*, Ozcan C. Esen[1], Samuel E. Miller[1,3], Kourtney L. Kroll[2], Amy D. Willis[4], Michael S. Rappé[5], Tao Pan[6], A. Murat Eren[1,3,7,8,9]*

Comprehensive sampling of natural genetic diversity with metagenomics enables highly resolved insights into the interplay between ecology and evolution. However, resolving adaptive, neutral, or purifying processes of evolution from intrapopulation genomic variation remains a challenge, partly due to the sole reliance on gene sequences to interpret variants. Here, we describe an approach to analyze genetic variation in the context of predicted protein structures and apply it to a marine microbial population within the SAR11 subclade 1a.3.V, which dominates low-latitude surface oceans. Our analyses reveal a tight association between genetic variation and protein structure. In a central gene in nitrogen metabolism, we observe decreased occurrence of nonsynonymous variants from ligand-binding sites as a function of nitrate concentrations, revealing genetic targets of distinct evolutionary pressures maintained by nutrient availability. Our work yields insights into the governing principles of evolution and enables structure-aware investigations of microbial population genetics.

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 (*1*, *2*). Surveys of microbial communities within natural habitats through phylogenetic marker genes (*3*–*5*) and metagenomics (*6*, *7*) have revealed a tremendous amount of genetic variation within environmental populations (*8*, *9*), 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 (*10*), have hindered our ability to determine which variants are under the influence of which pressures (*11*, *12*).

Inferring selective pressures through the isolation of microbial strains and comparative genomics has been widely successful. More recently, metagenome-assembled genomes (*13*) and single-amplified genomes (*14*) have markedly increased the number (*15*–*17*) and diversity (*18*) of microbial clades represented in genomic databases, offering an even larger sampling of environmental microbes to study the emergence and maintenance of genetic variation (*19*). 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 (*20*). This shortcoming is partially addressed by shotgun metagenomics (*21*) 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 (*22*, *23*). In particular, using genomes to recruit reads from metagenomes enables a comprehensive sampling of all genetic variants within environmental populations (*6*). Because of 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 (*24*–*32*). Although quantification and analysis of sequence variants derived from metagenomic data have improved markedly, 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" (*33*), including protein structures as a dimension of analysis is commonplace in studies of protein evolution (*34*–*36*), and it is appreciated that the accuracy of evolutionary models improves with combined analyses of protein structures and the evolution of underlying sequences (*37*). In contrast, the state-of-the-art approaches that quantify genetic variants in environmental microbial populations typically treat genes as strings of nucleotides (*24*, *31*, *38*–*40*). While this strategy enables rapid surveys of population dynamics through single-nucleotide variants (SNVs), it disregards the physical properties of three-dimensional (3D) gene products that selection acts upon and thus misses a critical intermediate to understand the relationship between selection and fitness (*41*, *42*). The importance of mapping sequence variants on predicted protein structures to

[1]Department of Medicine, University of Chicago, Chicago, IL 60637, USA. [2]Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL 60637, USA. [3]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA. [4]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA. [5]Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, Kāne'ohe, HI 96822, USA. [6]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA. [7]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany. [8]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany. [9]Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany.
*Corresponding author. Email: ekiefl@uchicago.edu (E.F.); meren@hifmb.de (A.M. E.)

identify genetic determinants of phenotypic variation has been noted more than two decades ago (*43*), yet the limited availability of protein structures has historically rendered protein structure-informed microbial population genetics impractical. Given marked advances in both solving and predicting protein structures in recent years (*44*), most notably deep learning approaches such as AlphaFold (*45*) that offer highly accurate protein structure predictions, this constraint is likely a problem of the past. Together, 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 advanced software solutions that integrate protein structures with omics data 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 (i) demonstrate the shortcomings of purely sequence-based approaches to interpret patterns of polymorphism observed within complex microbial populations; (ii) propose two structural features to interpret genetic variation, relative solvent accessibility (RSA) and distance to ligand (DTL); and (iii) 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.

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 (*46*), and Tara Oceans Project metagenomes (*47*), a massive collection of deeply sequenced marine samples from oceans and seas across the globe. SAR11 is divided into multiple subclades with distinct ecology (*48*). 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 (*39*) and most diverse (*27*) microbial lineages in marine systems, which recruits as much as 1.5% of all metagenomic short reads in surface ocean metagenomes (*27*).

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 S1), resulting in 390 million reads that were 94.5% identical to HIMB83 on average (fig. S1). As an individual member of a diverse subclade, HIMB83 has a genomic context that is insufficient for resolving the extent of genetic diversity within 1a.3.V. Regardless, HIMB83 has 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 1470 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 50× (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 (fig. S1). 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 (fig. S2).

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 SNVs, which quantify the frequency that each nucleotide allele (A, C, G, and 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 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 nt (nucleotides) out of 100 (table S2), precluding the possibility that this diversity is generated from excessive nonspecific mapping. While puzzling, this level of diversity is expected as it agrees with numerous studies that have pointed out the astonishing complexity of the SAR11 subclade 1a.3.V (*27*, *39*, *49*) that could not be further divided into sequence-discrete populations (*27*).

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 $pS^{(site)}$ and $pN^{(site)}$, 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, $pN^{(site)}$ and $pS^{(site)}$ can (i) resolve shorter evolutionary time scales than the characteristic fixation rate, (ii) be calculated from metagenomic read recruitment data without complete haplotypes, and (iii) define rates on a per-sample basis, thus enabling intersample comparisons. Overall, we found that the average $pS^{(site)}$ outweighed $pN^{(site)}$ by 19:1 (table S3), 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. While this is generally assumed to be true in general, SAR11 clades have been shown to exhibit particularly high enrichment of synonymous polymorphism relative to other marine-dwelling clades (*50*).

Nonsynonymous polymorphism avoids buried sites

$pN^{(site)}$ values varied significantly from site to site and from sample to sample, but overall, site-to-site variance was more explanatory

than sample-to-sample variance [79.74% ± 0.11% versus 0.42% ± 0.01% of total variance, ANOVA (analysis of variance)] (fig. S3). 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 (Fig. 1).

We used two independent methods to predict protein structures for the 799 core genes of 1a.3.V: (i) a template-based homology modeling approach with MODELLER (*51*), which predicted 346 structures, and (ii) a transformer-like deep learning approach with AlphaFold (*45*), 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 (fig. S4 and table S4). Thus, we opted to use AlphaFold structures for all downstream analyses due to its higher structural coverage. AlphaFold-predicted protein structures covered more than 90% of the core genes, highlighting the emerging opportunities afforded by recent advances in de novo structure prediction.

Aligning SCVs 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 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 (*52*). By calculating RSA for each site in the predicted structures, and then weighting every site by the $pN^{(site)}$ and $pS^{(site)}$ across all samples, we established proteome-wide distributions for $pN^{(site)}$ and $pS^{(site)}$ relative to RSA (Fig. 2A and fig. S5). These data showed that $pS^{(site)}$ closely resembled the null distribution (two-sample Kolmogorov-Smirnov statistic = 0.016), which illustrates the lack of influence of RSA on s-polymorphism, while $pN^{(site)}$ deviated significantly and instead exhibited strong preference for sites with higher RSA (two-sample Kolmogorov-Smirnov statistic = 0.235). 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.

Nonsynonymous polymorphism avoids active sites

While structural constraints ensure that 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 (*53*) 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 outward from the active site with a strength inversely correlated with distance from active site (*54*, *55*). More generally, it has been observed that conserved sites induce "conservation gradients" that surround them, leading to increased conservation among neighboring sites (*56*). On the basis of these ideas, we conceptualized the metric "distance to ligand" (DTL) as the distance of a given site to the closest active

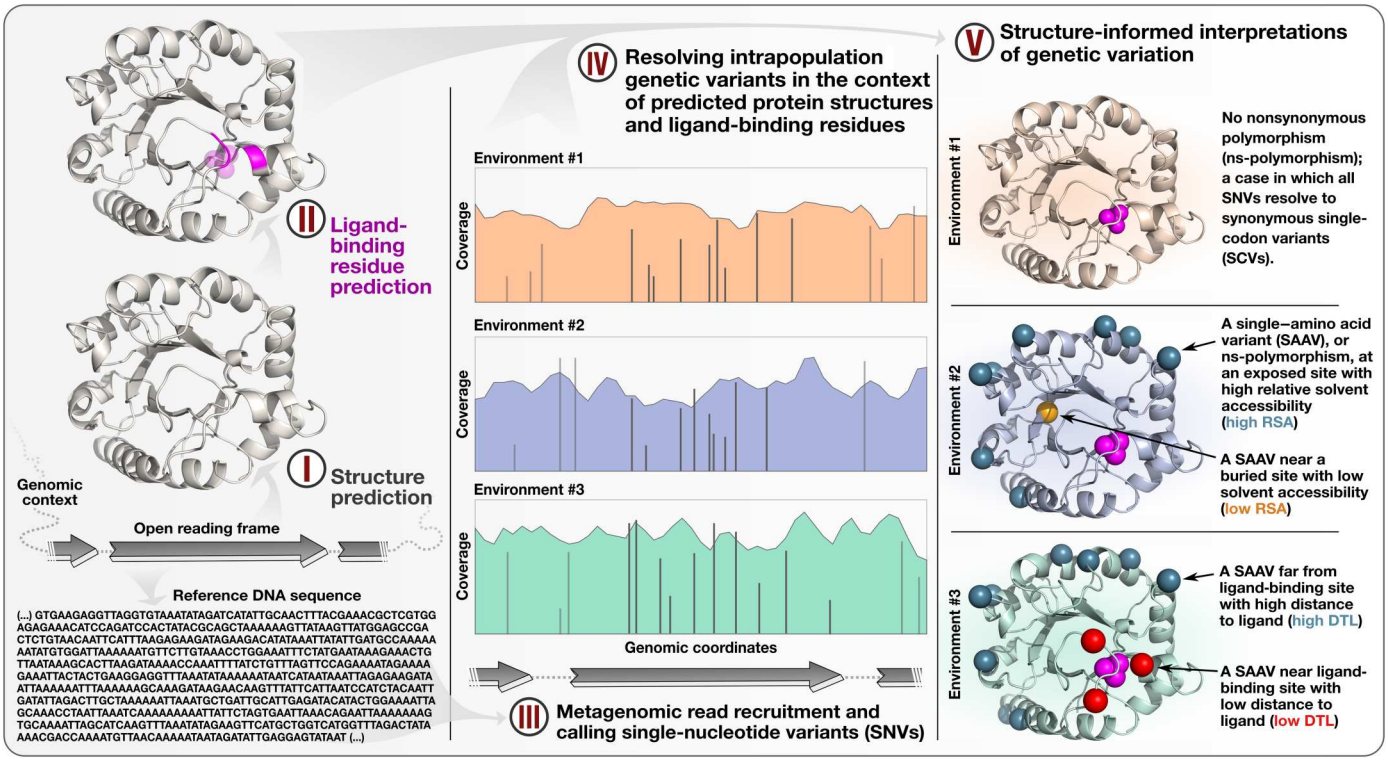

Fig. 1. The Anvi'o workflow for structure-informed population genetics. The proposed workflow combines predicted structures and metagenomic read recruitment results to interpret intrapopulation genetic variants in the context of protein structural properties.

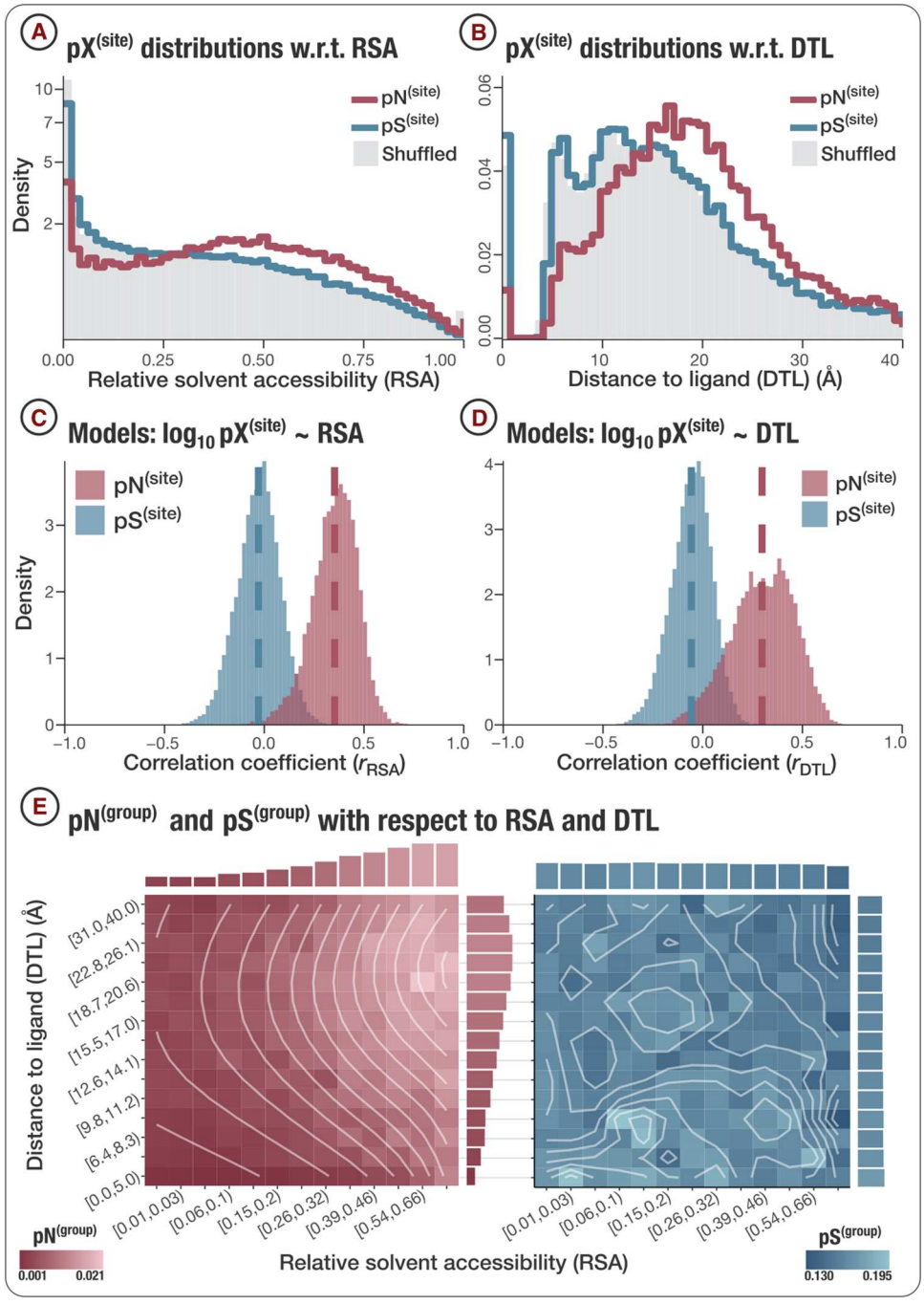

Fig. 2. Synonymous and nonsynonymous population genetic variants in the context of protein structures. (**A**) Structural constraints shift the $pN^{(site)}$ distribution toward high RSA. The $pN^{(site)}$ distribution (red line) and $pS^{(site)}$ distribution (blue line) were created by weighting the RSA values of 239,528 variable sites by the $pN^{(site)}$ and $pS^{(site)}$ values observed in each of the 74 metagenomes. The average distribution of 10 independent, randomly shuffled datasets of $pN^{(site)}$, is depicted by the gray regions for $pN^{(site)}$ and represents the null distribution expected if no association between $pN^{(site)}$ and RSA existed. (**B**) Functional constraint shifts the $pN^{(site)}$ distribution toward high 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 from 415 genes that had predicted structures and at least one predicted ligand by the $pN^{(site)}$ and $pS^{(site)}$ values observed in 74 metagenomes. (**C**) $pN^{(site)}$ 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); dashed lines visualize distribution means (see Supplementary Information). (**D**) $pN^{(site)}$ and DTL. The two distributions show Pearson correlation coefficients produced by linear models of the form $\log_{10}[pN^{(site)}] \sim$ DTL (red-filled region) and $\log_{10}[pS^{(site)}] \sim$ DTL (blue-filled region). (**E**) Per-group polymorphism rates with respect to (w.r.t.) RSA and DTL. Left and right: 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.

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 $pN^{(site)}$ and $pS^{(site)}$ as a function of DTL for each predicted structure by first predicting sites implicated in ligand binding using InteracDome (*53*), and then calculating a DTL for each site, given the closest predicted ligand-binding site (table S5).

The average per-site ns-polymorphism rate throughout the 1a.3.V core genome was 0.0088; however, we observed a nearly fourfold reduction in this rate to just 0.0024 at predicted ligand-binding sites (DTL = 0), indicating significantly (left-tailed Z test, $P < 1 \times 10^{-300}$) stronger purifying selection at ligand-binding sites (Fig. 2B and fig. S5). Sites neighboring ligand-binding regions also harbored disproportionately low rates of ns-polymorphism, as indicated by the significant deviation toward larger DTL values (two-sample Kolmogorov-Smirnov statistic = 0.157). 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 computed DTL in sequence space rather than Euclidean space afforded by protein structures, this effect was no longer observable beyond sequence distances of ~5 to 10 amino acids (fig. S6). Comparatively, $pS^{(site)}$ deviated minimally from the null distribution (two-sample Kolmogorov-Smirnov statistic = 0.013). 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 (i) structural constraints bias $pN^{(site)}$ distributions toward solvent-exposed sites (i.e., high RSA) (Fig. 2A) and (ii) functional constraints bias $pN^{(site)}$ distributions toward sites that are distant from ligand-binding sites (i.e., high DTL) (Fig. 2B).

Proteomic trends in purifying selection are explained by RSA and DTL

Given the clear shift in ns-polymorphism rates toward 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 $pN^{(site)}$ variation, respectively. On the basis of these models, we estimate that, for any given gene in any given sample, (i) a 1% increase in RSA corresponds to a 0.98% increase in $pN^{(site}$ and (ii) a 1% increase in DTL (normalized by the maximum DTL in the gene) corresponds to a 0.90% increase in $pN^{(site)}$. In a combined model, RSA and DTL jointly explained 14.12% of $pN^{(site)}$ 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 $pS^{(site)}$ 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 $pN^{(site)}$ can be modeled by RSA and DTL relative to $pS^{(site)}$ (Fig. 2, C and D). 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 (figs. S7 and S8). Last, 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 (Fig. 2E and table S8). 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, although oligomeric proteins are common, and they represent the majority of proteins in some organisms (*57*). 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 (i) ligands that are proteins, (ii) ligand-protein complexes that have not cocrystallized with each other, (iii) ligands of proteins with no shared homology in the InteracDome database, and (iv) 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 (fig. S9). 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. In addition, studies have shown that conservation gradients are stronger for catalytic versus noncatalytic binding sites (*58*), yet we do not distinguish between these ligand classes. Last, 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 (Fig. 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. The application of RSA and DTL to SAR11 demonstrates 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 remain difficult due to a multitude of factors. However, by providing structure-informed means to partition the total intrapopulation variation into distinct

pools, RSA and DTL offer a quantitative framework that enables new opportunities to study distinct evolutionary processes.

Differences in selection strength acting on individual genes can be measured across environments

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. $pN/pS^{(gene)}$ is a metric that quantifies the overall direction and magnitude of selection acting on a single gene (*24*, *30*), where $pN/pS^{(gene)} < 1$ indicates the presence of purifying selection, the intensity of which increases as the ratio decreases. Since $pN/pS^{(gene)}$ is defined for a given gene in a given sample, $pN/pS^{(gene)}$ values for a single gene can be compiled from multiple samples, enabling the tracking of selective pressures across environments (*30*). Taking advantage of the large number of metagenomes in which 1a.3.V was present, we calculated $pN/pS^{(gene)}$ for all 799 protein-coding core genes across 74 samples (see Methods), resulting in 59,126 gene/

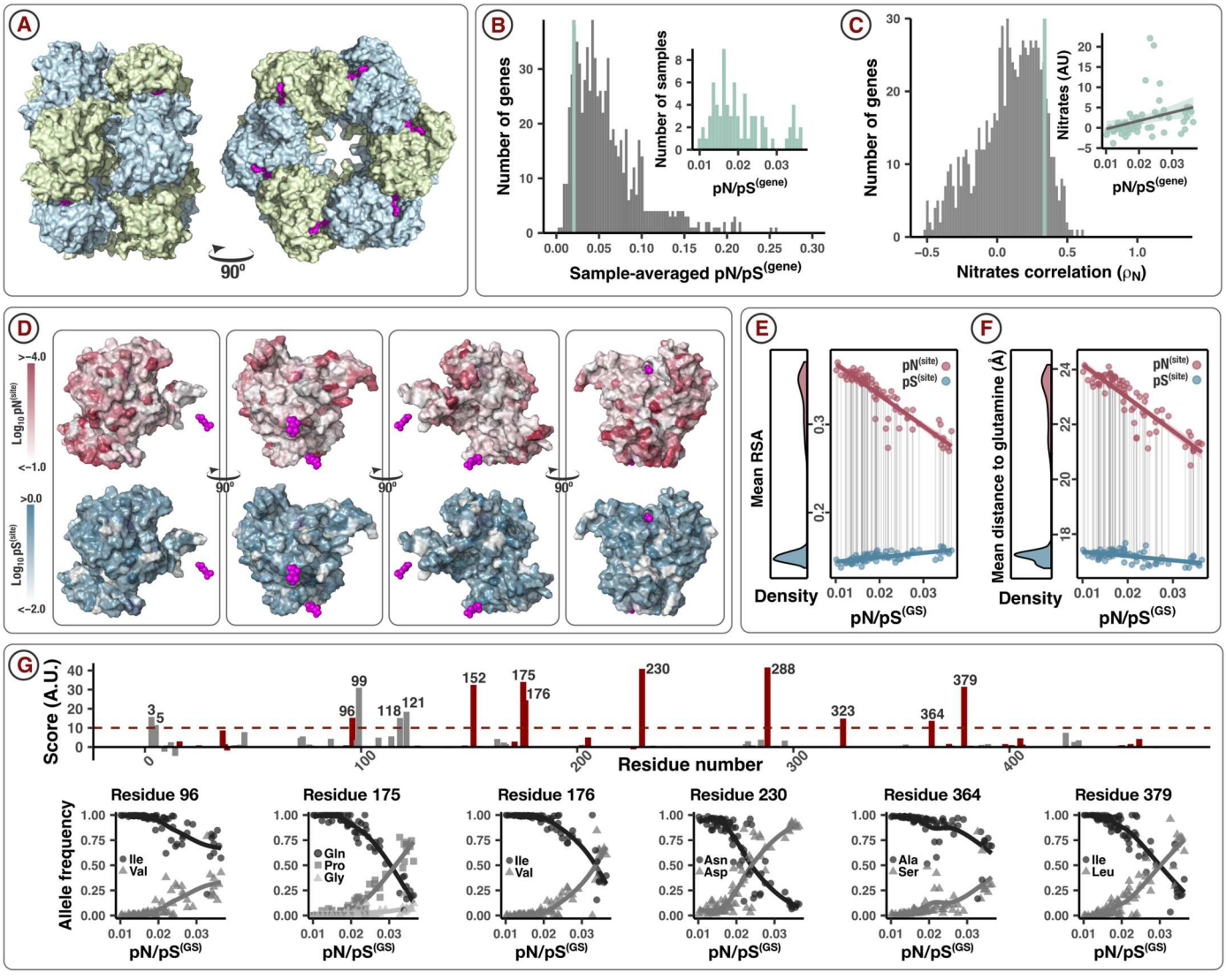

Fig. 3. Polymorphism distribution patterns in glutamine synthetase. (**A**) Glutamine synthetase (GS) dodecameric complex. Pink molecules are adenosine diphosphate and phosphinothricin (steric inhibitor of glutamate), situated within the active site of GS. (**B**) Sample-averaged $pN/pS^{(gene)}$ for GS (at 0.020, vertical green line) versus sample-averaged $pN/pS^{(gene)}$ for all 799 genes in the 1a.3.V core (truncated at 0.30). Inset: Distribution of $pN/pS^{(gene)}$ value for GS as seen across all metagenomes. (**C**) Pearson correlation coefficients for GS pN/pS and measured concentration of nitrates in each sample (0.34, vertical green line) versus $pN/pS^{(gene)}$ for all genes. Inset: Scatterplot of $pN/pS^{(gene)}$ versus nitrate concentrations from which the GS correlation coefficient was calculated. (**D**) 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 (A). Red surfaces are colored according to the sample-averaged $\log_{10}pN^{(site)}$ value of each residue, and blue surfaces are colored according to the sample-averaged $\log_{10}pS^{(site)}$ value of each residue. Darker colors refer to higher rates. (**E**) Correlation between average RSA values (*y* axis) with $pN/pS^{(gene)}$ across samples (*x* axis). (**F**) RSA in (E) is replaced with the distance-to-glutamate substrate (DTL). (**G**) Top: Site covariation with $pN/pS^{(GS)}$, where the *x* axis is the residue number and the *y* axis is the linear regression slope estimate between the sum of minor allele frequencies and $pN/pS^{(GS)}$. Sites with DTL values less than the average are indicated in red. All sites above the dashed horizontal line are annotated with their residue number. Scatterplots below show the allele frequency trajectories for a select number of these sites. A.U., arbitrary units.

sample pairs (table S9). We validated our calculations by comparing sample-averaged $pN/pS^{(gene)}$ to $dN/dS^{(gene)}$ 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 (see fig. S10, table S11, and Supplementary Information).

We found significantly more $pN/pS^{(gene)}$ variation between genes of a given sample ("gene-to-gene" variation) than between samples of a given gene ("sample-to-sample" variation) (ANOVA; fig. S11). All but one gene (gene #2031, unknown function) maintained $pN/pS^{(gene)} < 1$ in every sample, whereby 95% of values were less than 0.15 (fig. S12 and table S9), 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 $pS^{(site)}$ outweighed $pN^{(site)}$ by 19:1 within the aggregated data across genes and samples. However, the magnitude of purifying selection was not uniform across all genes. Gene-to-gene variance, as opposed to sample-to-sample variance, explained 93% of $pN/pS^{(gene)}$ variation (ANOVA; fig. S11). By analyzing the companion metatranscriptomic data (*59*) that were available for 50 of the 74 metagenomes, we were able to explain 29% of gene-to-gene variance with gene transcript abundance (TA) (see table S12 and Supplementary Information), a known predictor of evolutionary rate (*60*). Overall, these data demonstrate the utility of $pN/pS^{(gene)}$ as a metric to understand the overall extent of selection acting on genes.

The amount of $pN/pS^{(gene)}$ variation attributable to sample-to-sample variance was only 0.7% (fig. S11). While it represents a small proportion of the total variance, the sample-to-sample variance in $pN/pS^{(gene)}$ 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 manifest from differences in environmental parameters (table S10), which we focused on next.

Nitrogen availability governs rates of nonideal 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 (*61*), a limiting nutrient for microbial productivity in surface oceans (*62*). GS yields glutamine and adenosine diphosphate from glutamate, ammonia, and adenosine triphosphate, 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. The sample-averaged $pN/pS^{(GS)}$ was 0.02, ranking GS among the top 11% most purified genes (Fig. 3B and table S9). Although highly purified, we observed significant sample-to-sample variation in $pN/pS^{(GS)}$ (min = 0.010, max = 0.036) suggesting that the strength of purifying selection on GS varies from sample to sample (Fig. 3B, inset), perhaps due to unique environmental conditions (e.g., nutrient compositions) that differentially affect the need for glutamine synthesis. Since previous work has shown that SAR11 up-regulates its transcriptional and translational production of GS in response to nitrogen limitation (*63*), we hypothesized that purifying selection should be highest in nitrogen-limited environments, and lowest in nitrogen-replete environments. We used 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 $pN/pS^{(GS)}$ values across samples (Pearson correlation $P = 0.009$, $R^2 = 0.11$) (Fig. 3C), which ranked among the top 12% of positive correlations between $pN/pS^{(gene)}$ and nitrate concentration (Fig. 3C, inset, and table S10). 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* [Protein Data Bank (PDB) ID: 1FPY], which HIMB83 GS shares 61% amino acid similarity with (Fig. 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 17.8 Å versus 21.2 Å for DTL). With these quaternary estimates of RSA and DTL, we found that ns-polymorphism was 30× 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 (Fig. 3D). In comparison, s-polymorphism distributed relatively homogeneously throughout the protein, whereby 17% of s-polymorphism occurred within 10 Å 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 (Fig. 3E, left). Similarly, the mean DTL was 17.2 Å for s-polymorphism and 22.9 Å for ns-polymorphism (Fig. 3F, left). 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 (Fig. 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 $pN/pS^{(GS)}$. While the mean RSA of s-polymorphism remained relatively invariant (SD 0.005) (Fig. 3E, right), the mean RSA of ns-polymorphism varied markedly from 0.27 to 0.37 and was profoundly influenced by sample $pN/pS^{(GS)}$; samples exhibiting low selection of GS harbored lower mean RSA and samples exhibiting high selection of GS harbored higher mean RSA (Fig. 3E, right). In fact, 82.9% of mean RSA ns-polymorphism variance could be explained by $pN/pS^{(GS)}$ alone (Pearson correlation, $P < 1 \times 10^{-16}$, $R^2 = 0.829$). This correlation disappeared when the sites were shuffled (Pearson correlation, $R^2 = 0.014$, standard error 0.006 from 10 trials). ns-polymorphism distributions with respect to DTL were equally governed by selection strength, where 80.4% of variance could be explained by $pN/pS^{(GS)}$ (Pearson correlation, $P < 1 \times 10^{-16}$, $R^2 = 0.804$; Fig. 3F). This correlation disappeared when the sites were shuffled (Pearson correlation, $R^2 = 0.011$, standard error 0.004 from 10 trials).

When selection is low, we observe high nitrate concentrations (Fig. 3C, inset) and ns-polymorphism distributions toward lower RSA/DTL (Fig. 3, E and F). When selection is high, we observe low environmental nitrate concentrations (Fig. 3C, inset) and ns-polymorphism distributions toward higher RSA/DTL (Fig. 3, E and F). 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 nonideal 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 that 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 in low $pN/pS^{(GS)}$ samples and to reach increasingly higher frequencies in higher $pN/pS^{(GS)}$ samples. We identified putative sites fitting this description by scoring sites based on the extent that their amino acid minor allele frequencies covaried with $pN/pS^{(GS)}$, including only sites with DTL less than the mean DTL of ns-polymorphisms (22.9 Å). Using an arbitrary cutoff, we identified nine top-scoring polymorphisms that covaried with $pN/pS^{(GS)}$ (Fig. 3G): I96V, L152I, Q175P/G, I176V, N230D, S288A/D, I323V, A364S, and I379L. Although each of these sites exhibited DTL lower than the average ns-polymorphism, the closest site (residue number 323) was still 9 Å away from the glutamate substrate. This suggests that 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 $pN/pS^{(GS)}$ is 0.036, which is just more than half the genome-wide average $pN/pS^{(gene)}$ of 0.063. Our data therefore represent a subtle, yet resolvable signal of minute decreases in selection strength manifesting as minute shifts in the distribution of ns-polymorphism toward the active site.

While identifying signatures of positive selection is typically the primary pursuit in evolutionary analysis, our data instead illustrate 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 that 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 (see Fig. 4, A and B, and Supplementary Information), suggesting that this signal is not specific to GS, but rather a general feature of the 1a.3.V core genes. Although highly significant (one-sided Pearson $P = 9 \times 10^{-12}$ for RSA and $P = 2 \times 10^{-4}$ for DTL), the magnitude that ns-polymorphism distributions shift with respect to DTL and RSA was subtle: Across samples, the mean DTL of ns-polymorphism varied by less than 1 Å, and the mean RSA varied between 0.34 and 0.38. We performed the same analysis on s-polymorphism (Fig. 4, C and D) and observed similar trends (one-sided Pearson $P = 1 \times 10^{-5}$ for RSA and $P = 3 \times 10^{-7}$ for DTL), suggesting that differences in selection between samples may subtly drive the distribution of synonymous variants. That said, the effect size was even smaller for ns-polymorphism, with the mean DTL of s-polymorphism varying less than 0.2 Å and the mean RSA varying between 0.230 and 0.236. This is congruent with the observation that s-polymorphism distributes more uniformly throughout protein structure than does ns-polymorphism (Fig. 2, A and B). Resolving such a minute signal with such robust statistical power is attributed to the immense quantities of sequence data afforded by metagenomics.

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. By leveraging structure and ligand-binding predictions, we were able to describe notable patterns of nucleotide polymorphism in an environmental microbial population that we could ascribe to evolutionary constraints that preserve protein structure (folding and 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 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.

METHODS

Overview

The URL https://merenlab.org/data/anvio-structure/ provides a complete reproducible workflow for all analysis steps detailed below, including (i) downloading the publicly available metagenomes and genomes, (ii) recruiting reads from metagenomes, (iii) calculating single–amino acid variants (SAAVs) and single-codon variants (SCVs), (iv) predicting protein structures and ligand-binding sites, and (v) visualizing metagenomic sequence variants and binding sites onto protein structures.

Metagenomic and metatranscriptomic read recruitment and processing

To study the population structure of the environmental SAR11 population 1a.3.V defined previously (*27*), we used anvi'o v7.1 (*64*) and its metagenomics workflow (*65*), which uses snakemake v5.10 (*66*) to automate gene calling, gene function annotation, and metagenomic and metatranscriptomic read recruitment steps. The compendium of anvi'o programs the metagenomics workflow called upon used Prodigal v2.6.3 (*67*) for gene calling, National Center for Biotechnology Information's Clusters of Orthologous Groups database (*68*) and Pfams (*69*) for gene function annotation, HMMER v3.3 (*70*) for profile hidden Markov model (HMM) searches, DIAMOND v2.0.6 (*71*) for sequence searches, Bowtie2

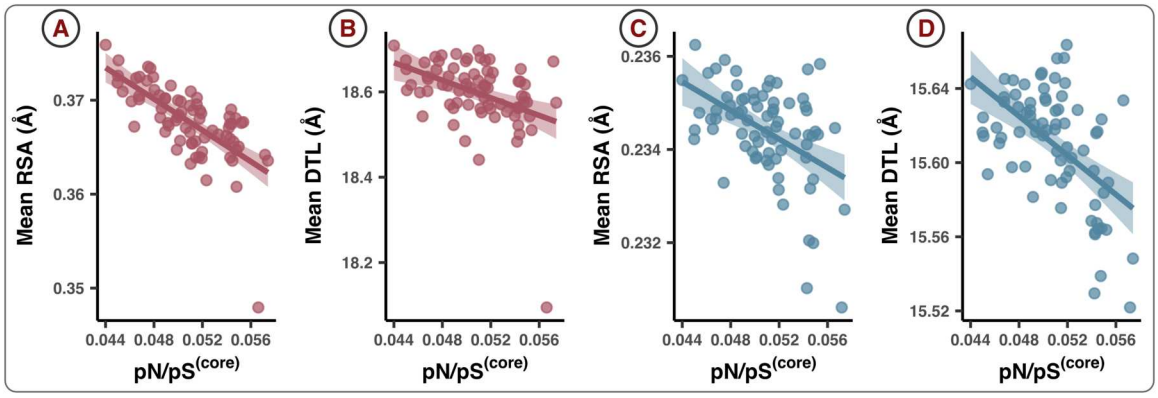

Fig. 4. Polymorphism distribution patterns with respect to genome-wide selection strength. Each data point is a sample (metagenome). Lines represent lines of best fit and corresponding translucent areas represent 95% confidence intervals. The *x* axis is $pN/pS^{(core)}$, which is calculated across the 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 $pN/pS^{(core)}$ (one-sided Pearson $P = 9 \times 10^{-12}$). (**B**) The ns-polymorphism distribution mean with respect to DTL is negatively associated with $pN/pS^{(core)}$ (one-sided Pearson $P = 2 \times 10^{-4}$). (**C**) The s-polymorphism distribution mean with respect to RSA is negatively associated with $pN/pS^{(core)}$ (one-sided Pearson $P = 1 \times 10^{-5}$). (**D**) The s-polymorphism distribution mean with respect to RSA is negatively associated with $pN/pS^{(core)}$ (one-sided Pearson $P = 3 \times 10^{-7}$).

v2.4 (*72*) for read recruitment, and SAMtools v1.9 (*73*) to generate BAM files. The metagenomic workflow resulted in two anvi'o artifacts, a "contigs database" (https://anvio.org/m/contigs-db) and a "merged profile database" (https://anvio.org/m/profile-db), which give access to gene functions, gene and genome coverages (with metagenomic or metatranscriptomic short reads), and 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 nonspecific read recruitment. In all subsequent analyses, we focused on the core genes of the 1a.3.V subclade by only considering (i) reads that mapped to HIMB83, (ii) the 74 metagenomes in which HIMB83 was found above 50×, and (iii) the 799 HIMB83 genes that were previously found to maintain consistent coverage patterns (*27*).

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 SCVs, from which 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 nt, which can happen either if the read terminates within the codon position or if 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 an SCV (or SAAV) and the columns specify (i) identifying information such as the corresponding gene, codon position, and sample id; (ii) the number of mapped reads corresponding to each of the 64 codons (or 20 amino acids); and (iii) numerous miscellaneous statistics, all of which can be explored at https://merenlab.org/analyzing-genetic-varaibility/.

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 (*30*), where a given codon allele contributes [to either $pN^{(site)}$ or $pS^{(site)}$] 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 that is 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 $pN^{(site)}$ and $pS^{(site)}$ by the number of nonsynonymous and synonymous sites of the popular consensus, respectively. For example, if the popular consensus is "ACC" (Thr), there are nine possible single-point mutations, three synonymous and six nonsynonymous; therefore, $pS^{(site)}$ will be divided by 3/3 = 1 and $pN^{(site)}$ will be divided by 6/3 = 2. This procedure can be mathematically expressed as

$$p_N{}^{(site)} = \frac{1}{n_n} \sum_{c \in C \backslash r} f_c N(c, r), p_S{}^{(site)} = \frac{1}{n_s} \sum_{c \in C \backslash r} f_c S(c, r)$$

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; and $N(c, r)$ 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 $pN^{(site)}$ and $pS^{(site)}$ values as additional columns to the tabular output after calculating them for three different choices of the reference codon *r*: (i) the popular consensus (as used here), (ii) the consensus (the allele with the highest frequency), and (iii) the reference sequence (the sequence used for read recruitment). For efficient computation, this calculation uses the Python package numba (*74*) for just-in-time compilation. For a dataset with 12,583,626 SCVs, the current implementation computes $pN^{(site)}$ and $pS^{(site)}$ terms in less than a minute on a laptop computer.

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 $pN^{(group)}$ and $pS^{(group)}$ as

$$p_N{}^{(group)} = \frac{\sum_{g=1}^{G} \sum_{c \in C \setminus r} f_c^{(g)} N[c, r^{(g)}]}{\sum_{g=1}^{G} n_n^{(g)}},$$

$$p_S{}^{(group)} = \frac{\sum_{g=1}^{G} \sum_{c \in C \setminus r} f_c^{(g)} S[c, r^{(g)}]}{\sum_{g=1}^{G} n_s^{(g)}}$$

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

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

Last, $pN/pS^{(group)}$ is defined as

$$pN/pS^{(group)} = pN^{(group)}/pS^{(group)}$$

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 $pN^{(group)}$, $pS^{(group)}$, and $pN/pS^{(group)}$. For example, $pN^{(gene)}$ refers to the ns-polymorphism rate of all sites in a given gene, and $pS^{(core)}$ refers to the s-polymorphism rate of all sites in the 1a.3.V core genome.

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 (*45*) and MODELLER (*51*). 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 RSA, secondary structure, and phi and psi angles calculated using DSSP (*75*, *76*). For AlphaFold predictions, we used a version of the codebase that closely resembles v2.0.1 (https://github.com/johnaparker/alphafold/tree/3829f4e0ba01aa1b4f01916c83e9ca5de771d98a gives access to its exact state) and ran predictions using six graphics processing units, 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 per-residue confidence metric score (pLDDT) <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, (i) searches the Research Collaboratory for Structural Bioinformatics PDB (*77*) for homologs using DIAMOND (*71*), then downloads tertiary structures for matching entries, and (ii) uses these homologs as templates to predict the gene's structure with MODELLER (*51*). 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 (*78*), we used single-domain templates that 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 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 (*79*).

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 HMMs with ligand-binding frequencies predicted from experimentally determined structural data (*53*). To associate match state binding frequencies of the profile HMMs to the sites of HIMB83 genes, we applied a protocol similar to that described by Kobren and 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 2375 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 noise cutoffs defined in Pfam models, 940 of the 1,470 HIMB83 coding genes had at least one domain hit, with a total of 1770 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 multithreaded 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 s on a laptop computer using a single thread.

Calculating RSA

We calculated RSA for each residue of each predicted structure, where RSA was defined as the accessible surface area (ASA) probed by a 1.4-Å-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 (*80*).

Calculating DTL

DTL was calculated for all sites that belonged to genes with (i) a predicted structure and (ii) 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 the "Proteomic trends in purifying selection are explained by RSA and DTL" section, 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 semimanually curated ligand-protein complexes (*81*). We found that the 1a.3.V DTL distribution had a much higher proportion of values >40 Å, suggesting that these likely result from incomplete characterization of binding sites (fig. S9). To mitigate the influence of this inevitable error source, we conservatively excluded DTL values >40 Å (8.0% of sites) in all analyses after Fig. 2B.

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.

Proportion of polymorphism rate variance explained by RSA and DTL

To calculate the extent that RSA and DTL can explain polymorphism rates, we constructed three synonymous models (s-models) and three nonsynonymous models (ns-models) (table S6). s-models fit linear regressions of $\log_{10}[\mathrm{pS}^{(\mathrm{site})}]$ to RSA (s #1), DTL (s #2), and both RSA and 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 and DTL (ns #3). In addition, each model included the gene and sample of the corresponding polymorphism as independent variables 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 marked 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).

Calculating TA

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

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

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

Statistical data analysis and visualization

We used R v3.5.1 (*82*) for the analysis of numerical data reported from anvi'o. For data visualization, we used ggplot2 (*83*) library in R and anvi'o, and finalized images for publication using Inkscape v1.1 (https://inkscape.org/).

Supplementary Materials

This PDF file includes:

Supplementary Information
Figs. S1 to S22
Legends for tables S1 to S13
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S13

REFERENCES AND NOTES

1. M. K. Burke, J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose, A. D. Long, Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature* **467**, 587–590 (2010).
2. R. E. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341 (1991).
3. G. J. Olsen, D. J. Lane, S. J. Giovannoni, N. R. Pace, D. A. Stahl, Microbial ecology and evolution: A ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337–365 (1986).
4. S. G. Acinas, V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, M. F. Polz, Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–554 (2004).
5. M. L. Sogin, H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, G. J. Herndl, Microbial diversity in the deep sea and the underexplored rare biosphere. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12115–12120 (2006).
6. S. L. Simmons, G. Dibartolo, V. J. Denef, D. S. A. Goltsman, M. P. Thelen, J. F. Banfield, Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLOS Biol.* **6**, e177 (2008).
7. E. E. Allen, G. W. Tyson, R. J. Whitaker, J. C. Detter, P. M. Richardson, J. F. Banfield, Genome dynamics in a natural archaeal population. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1883–1888 (2007).
8. T. P. Curtis, W. T. Sloan, Microbiology. Exploring microbial diversity—A vast below. *Science.* **309**, 1331–1333 (2005).
9. T. P. Curtis, I. M. Head, M. Lunn, S. Woodcock, P. D. Schloss, W. T. Sloan, What is the extent of prokaryotic diversity? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2023–2037 (2006).
10. B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski, M. M. Desai, The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
11. H. Ochman, Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.* **20**, 2091–2096 (2003).
12. T. H. M. Mes, Microbial diversity—Insights from population genetics. *Environ. Microbiol.* **10**, 251–264 (2008).
13. L.-X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, J. F. Banfield, Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
14. T. Woyke, D. F. R. Doud, F. Schulz, The trajectory of microbial single-cell sequencing. *Nat. Methods* **14**, 1045–1054 (2017).
15. A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, R. D. Finn, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
16. M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, R. Stepanauskas, Charting the complexity of the marine microbiome through single-cell genomics. *Cell* **179**, 1623–1635.e11 (2019).
17. L. Paoli, H.-J. Ruscheweyh, C. C. Forneris, S. Kautsar, Q. Clayssen, G. Salazar, A. Milanese, D. Gehrig, M. Larralde, L. M. Carroll, P. Sánchez, A. A. Zayed, D. R. Cronin, S. G. Acinas, P. Bork, C. Bowler, T. O. Delmont, M. B. Sullivan, P. Wincker, G. Zeller, S. L. Robinson, J. Piel, S. Sunagawa, Biosynthetic potential of the global ocean microbiome. Nature. 607, 11–118 (2022) .
18. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
19. N. R. Garud, K. S. Pollard, Population genetics in the human microbiome. *Trends Genet.* **36**, 53–67 (2020).
20. T. Van Rossum, P. Ferretti, O. M. Maistrenko, P. Bork, Diversity within species: Interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
21. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
22. R. J. Whitaker, J. F. Banfield, Population genomics in natural microbial communities. *Trends Ecol. Evol.* **21**, 508–516 (2006).
23. V. J. Denef, Peering into the genetic makeup of natural microbial populations using metagenomics, in *Population Genomics: Microorganisms*, M. F. Polz, O. P. Rajora, Eds. (Springer International Publishing, 2018), pp. 49–75.
24. S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, K. Kota, S. R. Sunyaev, G. M. Weinstock, P. Bork, Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
25. M. L. Bendall, S. L. Stevens, L.-K. Chan, S. Malfatti, P. Schwientek, J. Tremblay, W. Schackwitz, J. Martin, A. Pati, B. Bushnell, J. Froula, D. Kang, S. G. Tringe, S. Bertilsson, M. A. Moran, A. Shade, R. J. Newton, K. D. McMahon, R. R. Malmstrom, Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).
26. R. E. Anderson, J. Reveillaud, E. Reddington, T. O. Delmont, A. M. Eren, J. M. McDermott, J. S. Seewald, J. A. Huber, Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. *Nat. Commun.* **8**, 1114 (2017).
27. T. O. Delmont, E. Kiefl, O. Kilinc, O. C. Esen, I. Uysal, M. S. Rappé, S. Giovannoni, A. M. Eren, Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* **8**, e46497 (2019).
28. N. R. Garud, B. H. Good, O. Hallatschek, K. S. Pollard, Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biol.* **17**, e3000102 (2019).
29. S. Zhao, T. D. Lieberman, M. Poyet, K. M. Kauffman, S. M. Gibbons, M. Groussin, R. J. Xavier, E. J. Alm, Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe* **25**, 656–667.e8 (2019).
30. L. Shenhav, D. Zeevi, Resource conservation manifests in the genetic code. *Science* **370**, 683–687 (2020).
31. M. R. Olm, A. Crits-Christoph, K. Bouma-Gregson, B. A. Firek, M. J. Morowitz, J. F. Banfield, inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
32. A. Conwill, A. C. Kuan, R. Damerla, A. J. Poret, J. S. Baker, A. D. Tripp, E. J. Alm, T. D. Lieberman, Anatomy promotes neutral coexistence of strains in the human skin microbiome. *Cell Host Microbe* **30**, 171–182.e7 (2022).
33. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
34. J. Siltberg-Liberles, J. A. Grahnen, D. A. Liberles, The evolution of protein structures and structural ensembles under functional constraint. *Genes* **2**, 748–762 (2011).
35. M. J. Harms, J. W. Thornton, Evolutionary biochemistry: Revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).
36. T. Sikosek, H. S. Chan, Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface* **11**, 20140419 (2014).
37. C. O. Wilke, Bringing molecules back into molecular evolution. *PLOS Comput. Biol.* **8**, e1002572 (2012).
38. A. M. Eren, Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, T. O. Delmont, Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
39. S. Nayfach, B. Rodriguez-Mueller, N. Garud, K. S. Pollard, An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
40. P. I. Costea, R. Munch, L. P. Coelho, L. Paoli, S. Sunagawa, P. Bork, metaSNV: A tool for metagenomic strain level analysis. *PLOS ONE* **12**, e0182392 (2017).
41. G. B. Golding, A. M. Dean, The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15**, 355–369 (1998).
42. K. Chen, 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**, 5618–5622 (1993).
43. S. Sunyaev, W. Lathe III, 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**, 125–130 (2001).
44. B. Kuhlman, P. Bradley, Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
45. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
46. R. M. Morris, M. S. Rappé, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, S. J. Giovannoni, SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).

47. S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis; Tara Oceans coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
48. S. J. Giovannoni, SAR11 bacteria: The most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.* **9**, 231–255 (2017).
49. J. M. Haro-Moreno, F. Rodriguez-Valera, R. Rosselli, F. Martinez-Hernandez, J. J. Roda-Garcia, M. L. Gomez, O. Fornas, M. Martinez-Garcia, M. López-Pérez, Ecogenomics of the SAR11 clade. *Environ. Microbiol.* **22**, 1748–1763 (2020).
50. M. López-Pérez, J. M. Haro-Moreno, F. H. Coutinho, M. Martinez-Garcia, F. Rodriguez-Valera, The evolutionary success of the marine bacterium SAR11 analyzed through a metagenomic perspective. *mSystems* **5**, e00605-20 (2020).
51. B. Webb, A. Sali, Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* **54**, 5.6.1–5.6.37 (2016).
52. J. Echave, S. J. Spielman, C. O. Wilke, Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17**, 109–121 (2016).
53. S. N. Kobren, M. Singh, Systematic domain-based aggregation of protein structures highlights DNA-, RNA- and other ligand-binding positions. *Nucleic Acids Res.* **47**, 582–593 (2019).
54. A. M. Dean, C. Neuhauser, E. Grenier, G. B. Golding, The pattern of amino acid replacements in α/β-barrels. *Mol. Biol. Evol.* **19**, 1846–1864 (2002).
55. B. R. Jack, A. G. Meyer, J. Echave, C. O. Wilke, Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biol.* **14**, e1002452 (2016).
56. A. Sharir-Ivry, Y. Xia, Quantifying evolutionary importance of protein sites: A tale of two measures. *PLOS Genet.* **17**, e1009476 (2021).
57. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
58. A. Sharir-Ivry, Y. Xia, Non-catalytic binding sites induce weaker long-range evolutionary rate gradients than catalytic sites in enzymes. *J. Mol. Biol.* **431**, 3860–3870 (2019).
59. G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain; Tara Oceans Coordinators, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).
60. C. Pál, B. Papp, L. D. Hurst, Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
61. S. M. Bernard, D. Z. Habash, The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling. *New Phytol.* **182**, 608–620 (2009).
62. L. A. Bristow, W. Mohr, S. Ahmerkamp, M. M. M. Kuypers, Nutrients that limit growth in the ocean. *Curr. Biol.* **27**, R474–R478 (2017).
63. D. P. Smith, J. C. Thrash, C. D. Nicora, M. S. Lipton, K. E. Burnum-Johnson, P. Carini, R. D. Smith, S. J. Giovannoni, Proteomic and transcriptomic analyses of "Candidatus Pelagibacter ubique" describe the first PII-independent response to nitrogen limitation in a free-living Alphaproteobacterium. *MBio* **4**, e00133-12 (2013).
64. A. M. Eren, E. Kiefl, A. Shaiber, I. Veseli, S. E. Miller, M. S. Schechter, I. Fink, J. N. Pan, M. Yousef, E. C. Fogarty, F. Trigodet, A. R. Watson, Ö. C. Esen, R. M. Moore, Q. Clayssen, M. D. Lee, V. Kivenson, E. D. Graham, B. D. Merrill, A. Karkman, D. Blankenberg, J. M. Eppley, A. Sjödin, J. J. Scott, X. Vázquez-Campos, L. J. McKay, E. A. McDaniel, S. L. R. Stevens, R. E. Anderson, J. Fuessel, A. Fernandez-Guerra, L. Maignien, T. O. Delmont, A. D. Willis, Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
65. A. Shaiber, A. D. Willis, T. O. Delmont, S. Roux, L.-X. Chen, A. C. Schmid, M. Yousef, A. R. Watson, K. Lolans, Ö. C. Esen, S. T. M. Lee, N. Downey, H. G. Morrison, F. E. Dewhirst, J. L. Mark Welch, A. M. Eren, Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.* **21**, 292 (2020).
66. J. Köster, S. Rahmann, Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
67. D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* **11**, 119 (2010).
68. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
69. S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, R. D. Finn, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
70. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
71. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
72. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
73. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
74. S. K. Lam, A. Pitrou, S. Seibert, Numba: A LLVM-based Python JIT compiler, in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (Association for Computing Machinery, 2015), pp. 1–6.
75. W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, G. Vriend, A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2015).
76. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
77. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
78. M. Källberg, H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu, J. Xu, Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **7**, 1511–1522 (2012).
79. B. Rost, Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
80. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
81. J. Yang, A. Roy, Y. Zhang, BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103 (2013).
82. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Development Core Team, 2011); www.r-project.org.
83. C. Ginestet, ggplot2: Elegant graphics for data analysis: Book reviews. *J. R. Stat. Soc. Ser. A Stat. Soc.* **174**, 245–246 (2011).
84. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
85. S. Zhang, J. M. Krieger, Y. Zhang, C. Kaya, B. Kaynak, K. Mikulska-Ruminska, P. Doruker, H. Li, I. Bahar, ProDy 2.0: Increased scale and scope after 10 years of protein dynamics modelling with python. *Bioinformatics* **37**, 3657–3659 (2021).
86. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
87. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
88. C. L. Worth, S. Gong, T. L. Blundell, Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **10**, 709–720 (2009).
89. D. A. Drummond, C. O. Wilke, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
90. D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, F. H. Arnold, Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14338–14343 (2005).
91. A. S. Rose, P. W. Hildebrand, NGL viewer: A web application for molecular visualization. *Nucleic Acids Res.* **43**, W576–W579 (2015).
92. A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić, P. W. Rose, Web-based molecular graphics for large complexes, in *Proceedings of the 21st International Conference on Web3D Technology* (Association for Computing Machinery, 2016), pp. 185–186.

Acknowledgments: We thank S. Giovannoni (https://microbiology.oregonstate.edu/dr-stephen-giovannoni), S. N. Kobren (http://shilpakobren.com/), M. K. Yu (https://github.com/michaelkyu), T. Sosnick (http://sosnick.uchicago.edu/), C. DeValk, and the members of our laboratory (https://merenlab.org/people/) for helpful discussions. **Funding:** E.K. acknowledges support from the Natural Sciences and Engineering Research Council of Canada. A.D.W. acknowledges support from the National Institute of General Medical Sciences (R35 GM133420). The authors are thankful for the Open Access Publication Funds by Alfred Wegener Institute for Polar and Marine Research. This work was supported by Alfred P. Sloan Foundation Fellowship in Ocean Sciences to A.M.E. and by a Simons Foundation grant (#687269) to A.M.E.
Author contributions: E.K. and A.M.E. conceptualized the study and interpreted findings. E.K. curated data, developed software tools, and performed primary analyses. O.C.E. and A.M.E. contributed software. E.K. and A.M.E. wrote the paper. S.E.M., K.L.K., and A.D.W. helped with data

analyses and interpretation. M.S.R. and T.P. helped with project management and funding acquisition. A.M.E. supervised the project. All authors commented on the drafts of the study. All authors read and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All code and data needed to evaluate the conclusions in the paper are present in the paper, in our reproducible bioinformatics workflow at https://merenlab.org/data/anvio-structure/, and/or the Supplementary Materials.

Submitted 9 April 2022
Accepted 18 January 2023
Published 22 February 2023
10.1126/sciadv.abq4632