Science Advances
AAAS

Supplementary Materials for

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

Evan Kiefl *et al.*

Corresponding author: Evan Kiefl, ekiefl@uchicago.edu; A. Murat Eren, meren@hifmb.de

Sci. Adv. **9**, eabq4632 (2023)
DOI: 10.1126/sciadv.abq4632

The 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

Supplementary Figures

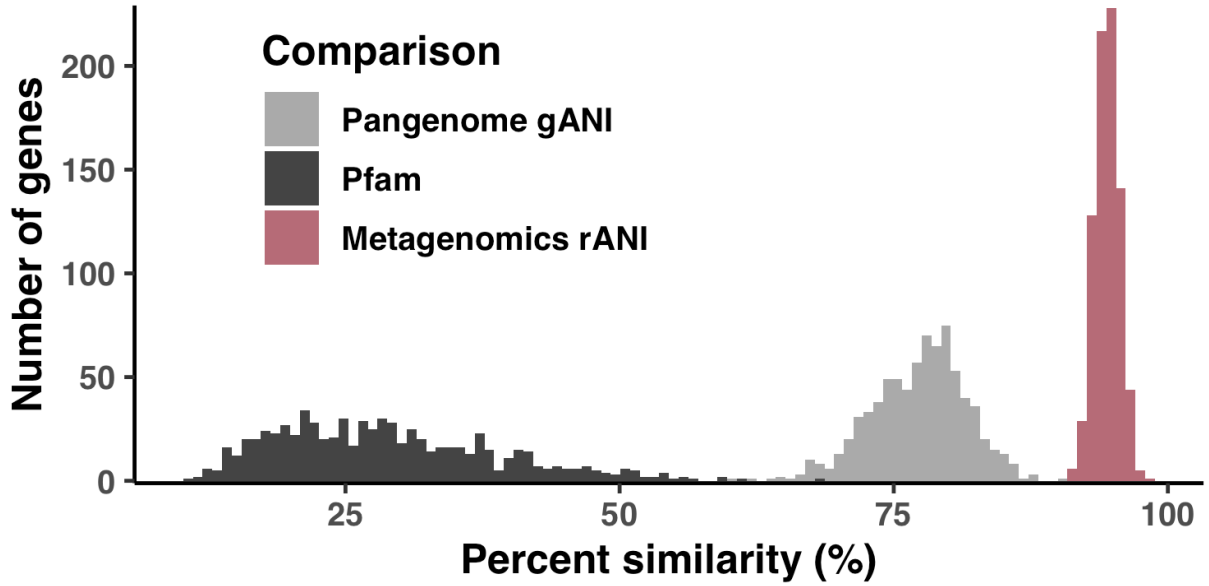

Figure S1. Regimes of sequence similarity probed by metagenomics, SAR11 cultured genomes, and protein families. Empirical distributions of gene-level percent similarity for HIMB83 compared with recruited metagenomic reads (red), homologous SAR11 genomes (grey), and homologous Pfams (black). For calculation details, see Supplementary Information.

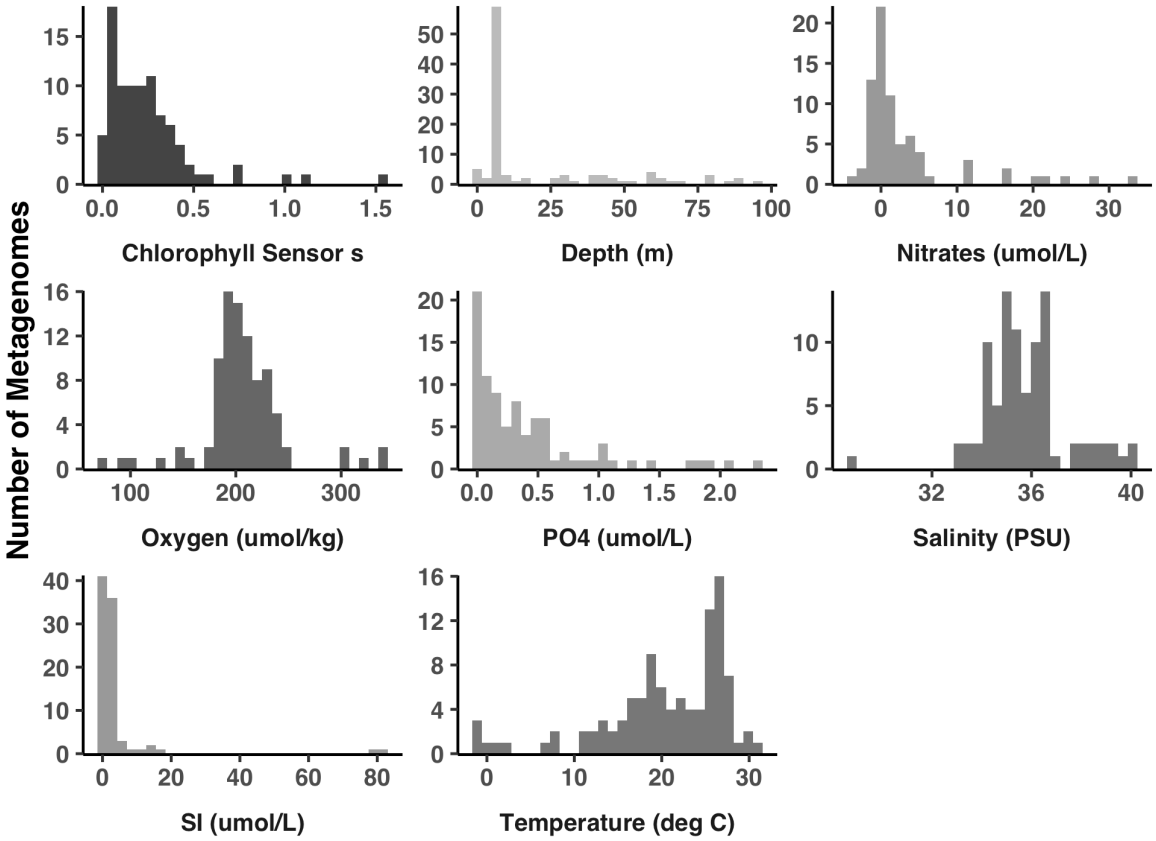

Figure S2. Different environments exhibit substantial variation in their environmental parameters. Each subplot shows how the 74 selected metagenomes distribute according to various environmental variables measured by the TARA ocean metagenome project.

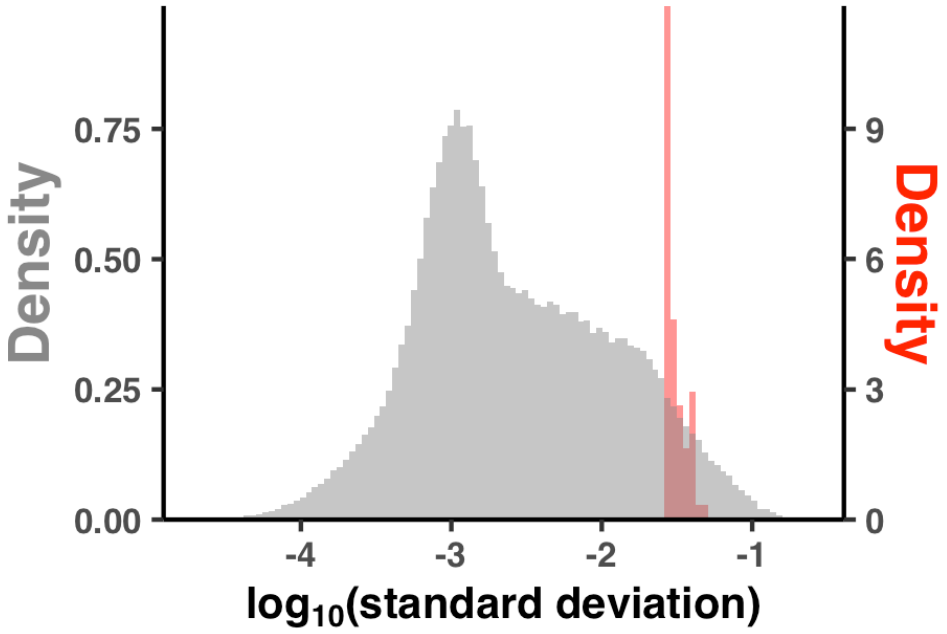

Figure S3. $pN^{(site)}$ varies more significantly between sites in a given sample than between samples for a given site. The x-axis is the log-transformed standard deviation of either a sample's $pN^{(site)}$ values observed over many sites (orange), or a site's $pN^{(site)}$ values observed over the 74 samples (gray).

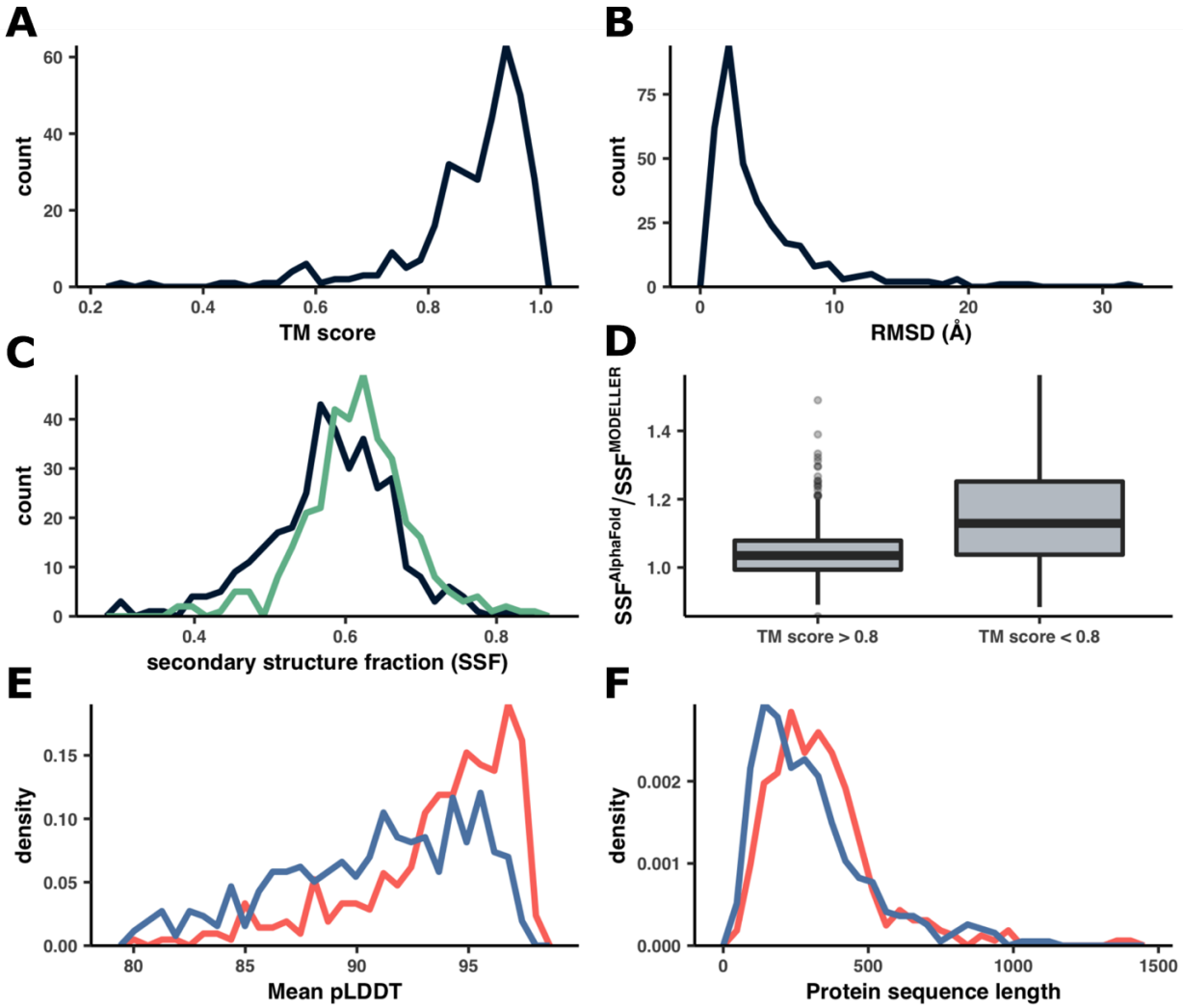

Figure S4. Comparisons between structures predicted by AlphaFold and MODELLER. (A-B) Distributions of TM scores and RMSD between structures predicted by both MODELLER and AlphaFold. **(C)** Distribution of secondary structure fractions, between MODELLER (black) and AlphaFold (green). Secondary structure 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).

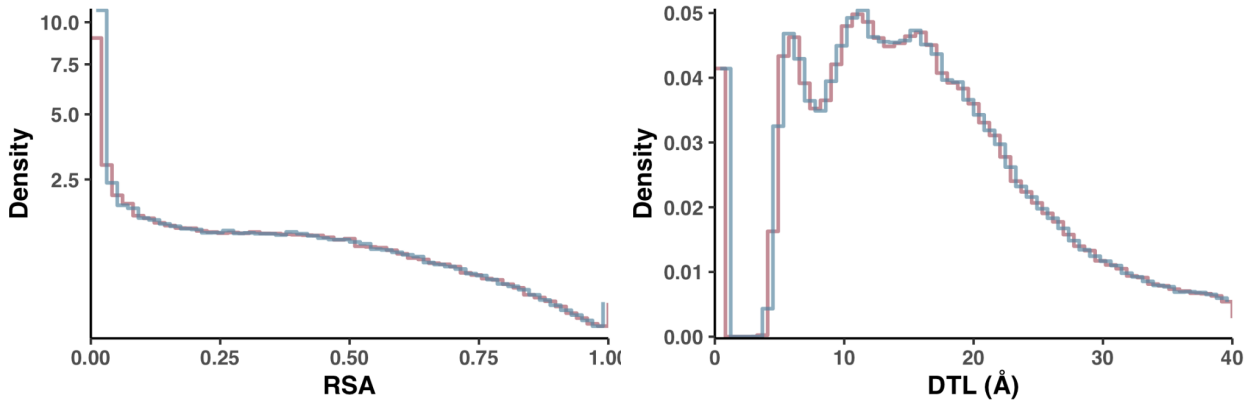

Figure S5. Comparison of null distributions for $pN^{(site)}$ and $pS^{(site)}$ for RSA and DTL. Each distribution was calculated by averaging 10 independent, randomly shuffled datasets of either $pN^{(site)}$ (red line) or $pS^{(site)}$ (blue line). To better visualize differences between the null distributions, the blue lines depicting the $pS^{(site)}$ distributions were shifted right by half of a bin's width.

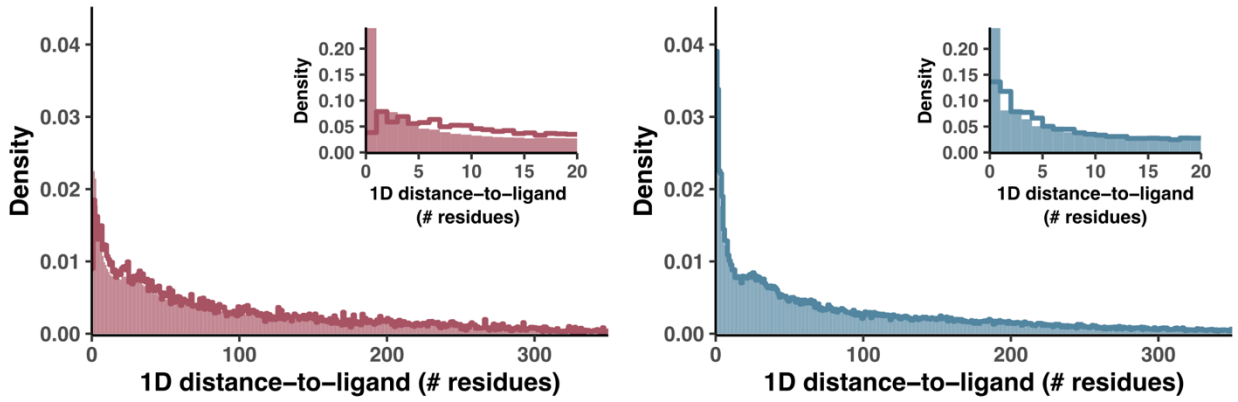

Figure S6. Functional constraint is less resolved when using a sequence-distance metric of DTL. $pN^{(site)}$ (left panel) and $pS^{(site)}$ (right panel) distributions with respect to 1D DTL, which we defined as the number of sites in a protein's sequence that separate a given site from a predicted ligand-binding site. Lines represent the observed distributions, and filled regions represent the null distributions, calculated via the shuffling procedure described in Figure 2. Insets show the same data zoomed into the 1D DTL range [0, 20].

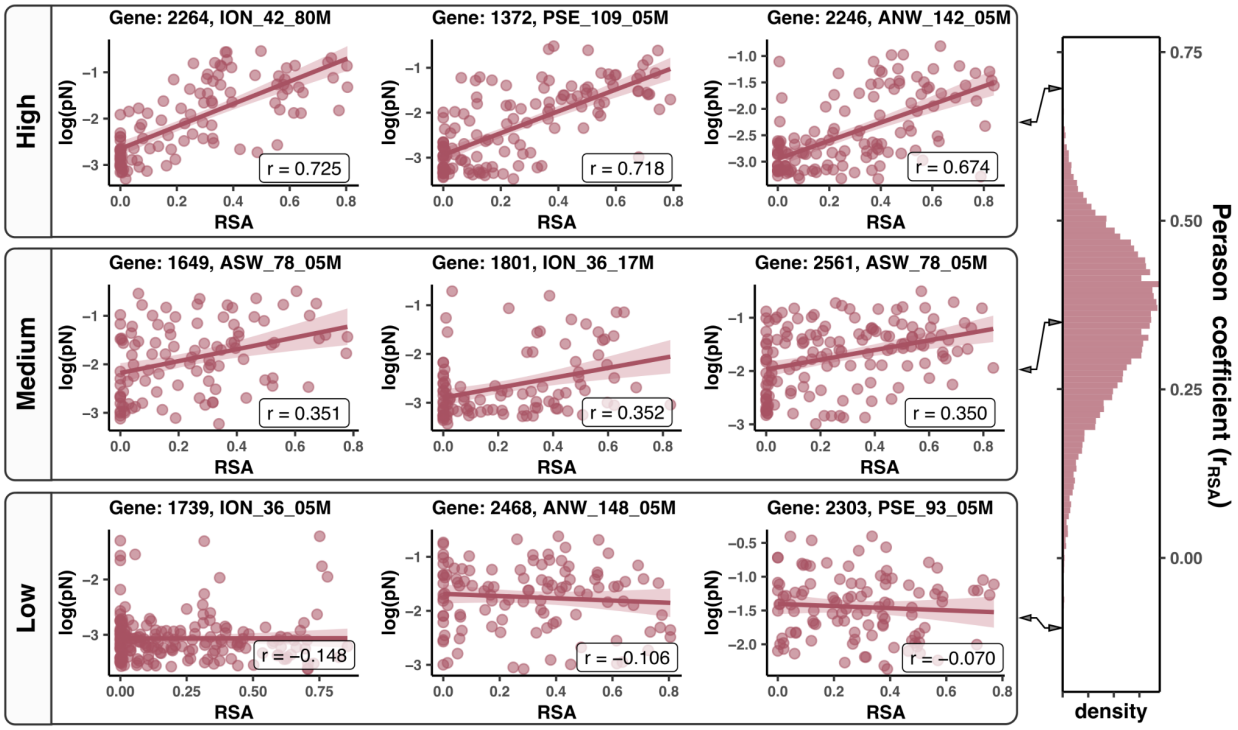

Figure S7. Select gene-sample pairs illustrate the diversity with which $pN^{(site)}$ associates with RSA. Scatterplots for handpicked gene-sample pairs are shown from three regimes of model quality: high (top), mid (middle), and low (bottom). The right panel shows the distribution of Pearson coefficients, 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 $\log_{10}(pN^{(site)})$. 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.

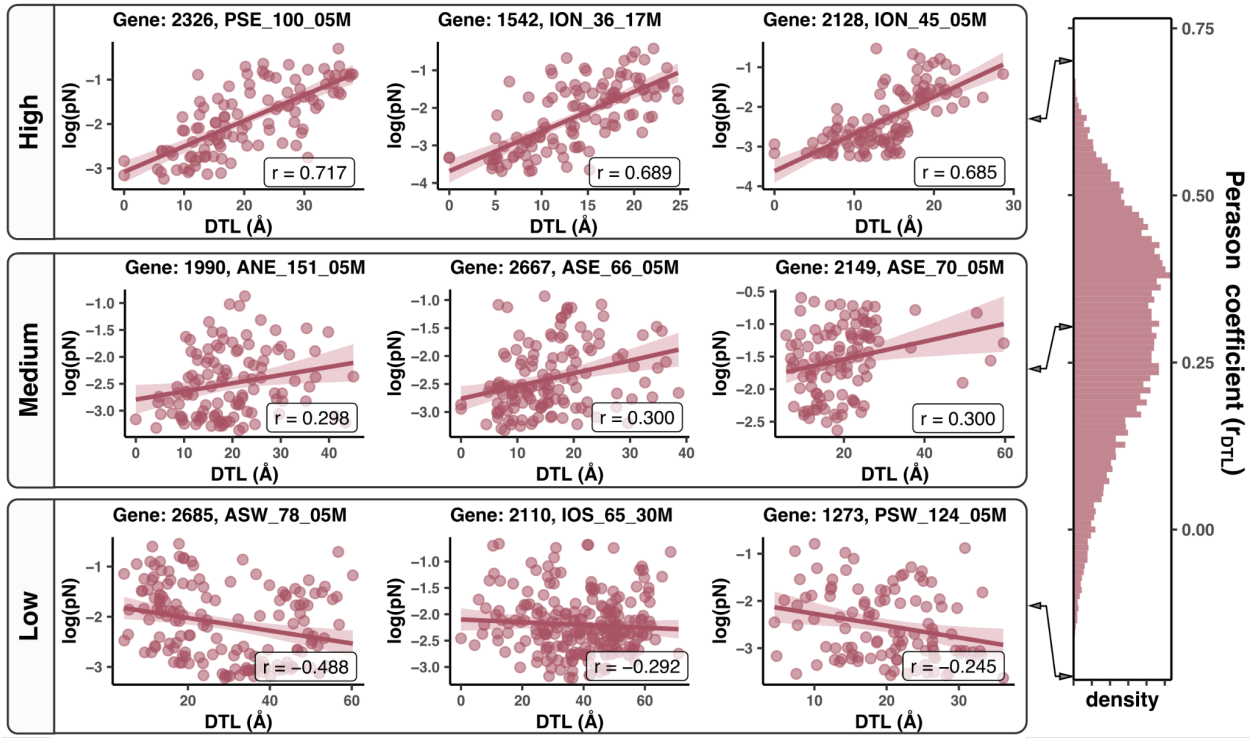

Figure S8. Select gene-sample pairs illustrate the diversity with which $pN^{(site)}$ associates with DTL. Scatterplots for handpicked gene-sample pairs are shown from three regimes of model quality: high (top), mid (middle), and low (bottom). The right panel shows the distribution of Pearson coefficients, 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 $\log_{10}(pN^{(site)})$. 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.

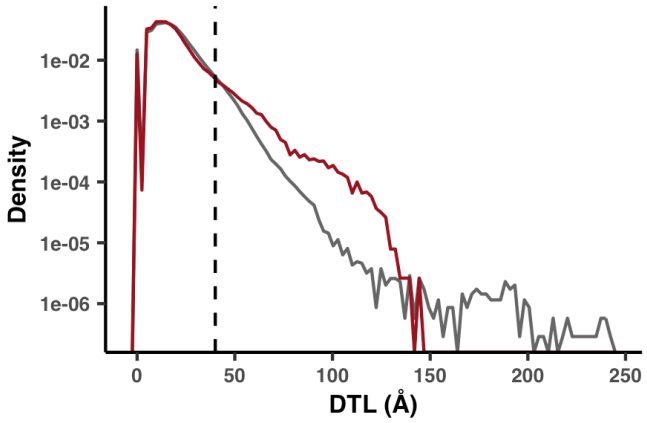

Figure S9. Incomplete ligand characterization leads to erroneously high DTL values. A comparison of DTL distributions (semi-log axis) for the 1a.3.V and the BioLiP database. The 1a.3.V core distribution (red) was calculated from all sites in the subset of genes with both a predicted structure and at 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 2b, which excludes 8.0% of the total sites.

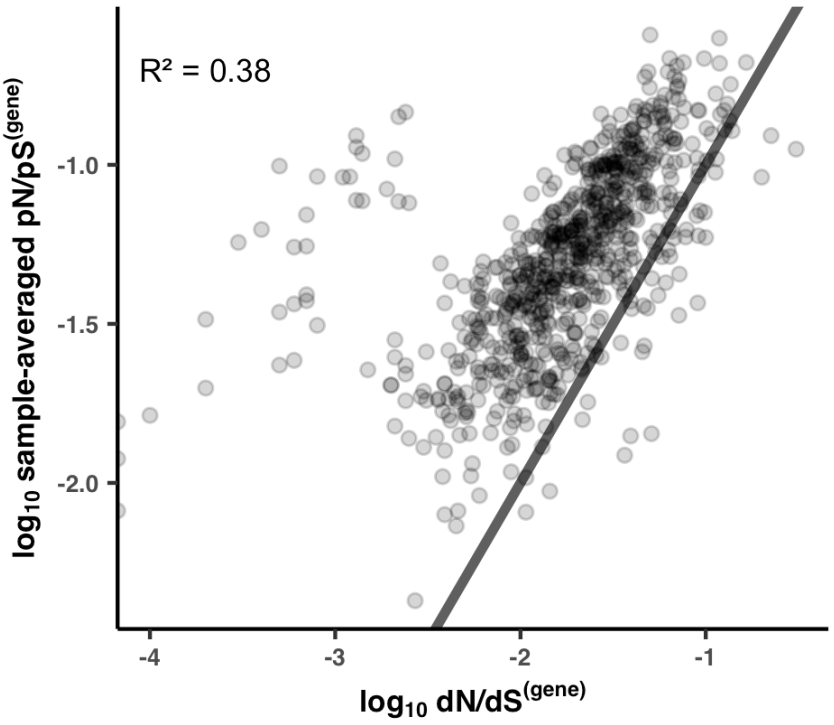

Figure S10. Sample-averaged pN/pS$^{(gene)}$ values correlate with dN/dS$^{(gene)}$ values between HIMB83 and HIMB122. The x- and y-axes are the log-transformed dN/dS$^{(gene)}$ and sample-averaged pN/pS$^{(gene)}$ values (respectively) for the 743 genes that (1) belonged to the 1a.3.V core and (2) had HIMB122 homologs. The black line is the equation y = x, meaning that genes above this line maintain sample-averaged pN/pS$^{(gene)}$ values that exceed dN/dS$^{(gene)}$. The R^2 is for a linear regression of the log-transformed variables.

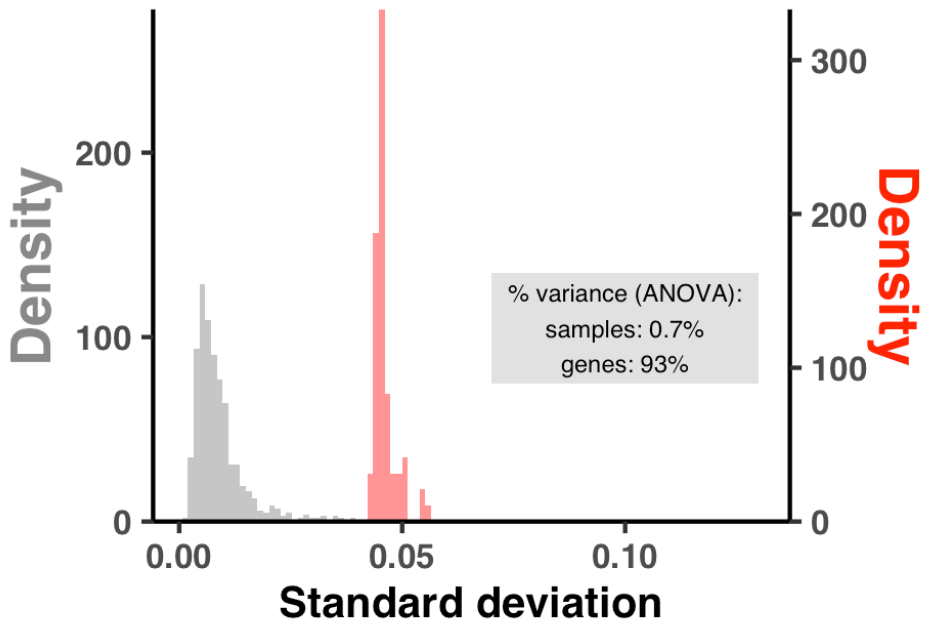

Figure S11. pN/pS$^{(gene)}$ varies more significantly between genes in a given sample than between samples for a given gene. The x-axis is the standard deviation of either a sample's pN/pS$^{(gene)}$ values observed over genes (orange), or a gene's pN/pS$^{(gene)}$ values observed over the 74 samples (gray). The gray box denotes the amount of variance explained by genes and samples in an ANOVA from the linear model pN/pS$^{(gene)}$ ~ gene + sample.

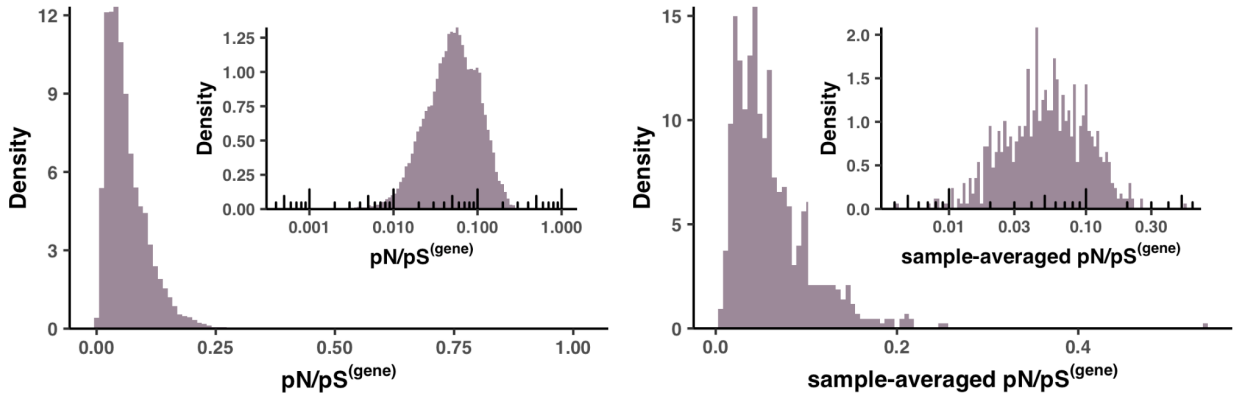

Figure S12. Distributions of pN/pS$^{(gene)}$. Left panel shows the distribution of pN/pS$^{(gene)}$, and the right panel shows the distribution of sample-averaged pN/pS$^{(gene)}$. Insets show the same distributions with a $\log_{10}$-transformed x-axis.

Supplementary Information

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 (*82*), 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 (*85*) 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 S1 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.

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 pre-existing 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 (*86*) and alpha carbon RMSD. Overall, the distributions of these metrics (Figures S4a, SS4b) illustrate the overarching similarity between AlphaFold and MODELLER structures. Since a score of 0.5 indicates that proteins likely belong to the same fold family (*87*), 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 S4c), and we found this discrepancy to be most pronounced when TM scores were low (<0.8) (Figure S4d). 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 S4e) 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 (*45*).

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.

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 ($pN^{(site)} = 0$ for ns-models, $pS^{(site)} = 0$ 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 S7).

We fit linear models of $\log_{10}(pN^{(site)})$ and $\log_{10}(pS^{(site)})$ 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 r_{RSA} = 0.353) (Figure 2c), whereas s-models exhibited correlations centered around 0 (average r_{RSA} = -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 $\log_{10}(pN^{(site)})$ and RSA (Figure S7). 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 $\log_{10}(pN^{(site)})$ and $\log_{10}(pS^{(site)})$ 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 $\log_{10}(pN^{(site)})$ and DTL (Figure S8). Interestingly, we found that $\log_{10}(pS^{(site)})$ on average negatively correlates with DTL (average Pearson coefficient -0.057). The overall positive correlation of DTL with $\log_{10}(pN^{(site)})$ 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 $\log_{10}(pN^{(site)})$ and $\log_{10}(pS^{(site)})$ with RSA and DTL together (Figure S14; 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 S13). 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 $R^2_{RSA\text{-}DTL}$ = 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 (*88*), 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 2e, Table S8). 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 $pN^{(group)}$ and $pS^{(group)}$, which are weighted averages of $pN^{(site)}$ and $pS^{(site)}$ 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. $pN^{(group)}$ values from each group collectively describe a 2D surface (Figure 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 $pN^{(site)}$ data observed within gene-sample pairs (Figures S7, S8),

the $pN^{(group)}$ surface is smooth and roughly linear (Figure 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 2e, bottom-left) statistically exhibited the lowest rates of ns-polymorphism, and sites with both high RSA and DTL (left panel of Figure 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 $pN^{(group)}$ due to structural constraint (top-left) while others exhibit low $pN^{(group)}$ 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 $pN^{(group)} \sim i + j$, 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 S15, 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 S16).

Site averaging yielded an unexpected relationship between s-polymorphism rates and RSA/DTL. $pS^{(group)}$ is not as strongly affected by RSA or DTL as $pN^{(group)}$, as indicated by the noisy contour lines of its surface (right panel Figure 2e). Even still, the linear model $pS^{(group)} \sim i + j$ 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 S15). We have observed this surprising finding through other means as well: in the sample-gene models, (1) the

mean Pearson correlation coefficient between $pS^{(site)}$ and RSA is -0.013 (Figure 2c), and (2) the mean Pearson correlation coefficient between $pS^{(site)}$ and DTL is -0.052 (Figure 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.

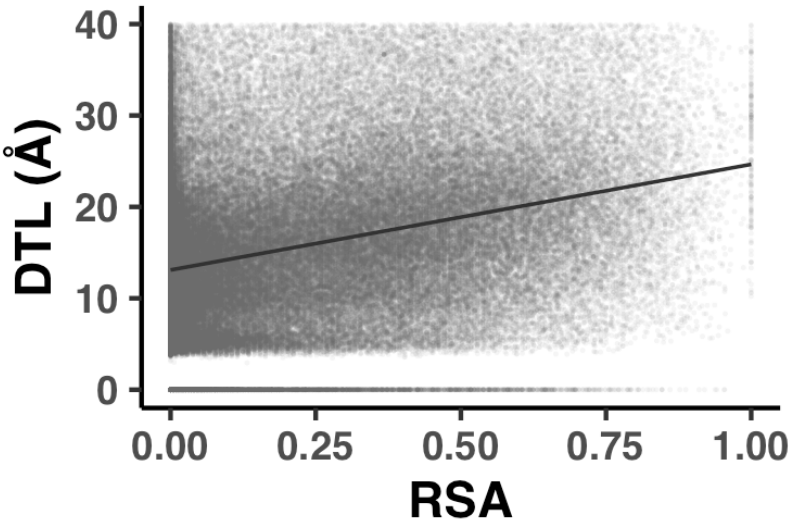

Figure S13. RSA and DTL are not problematically correlated. Scatter plot of RSA vs. DTL for the 143,181 sites belonging to genes with a predicted structure and at least one predicted ligand. The line of best fit is shown in black, The Pearson coefficient is 0.313 and the R^2 is 0.098.

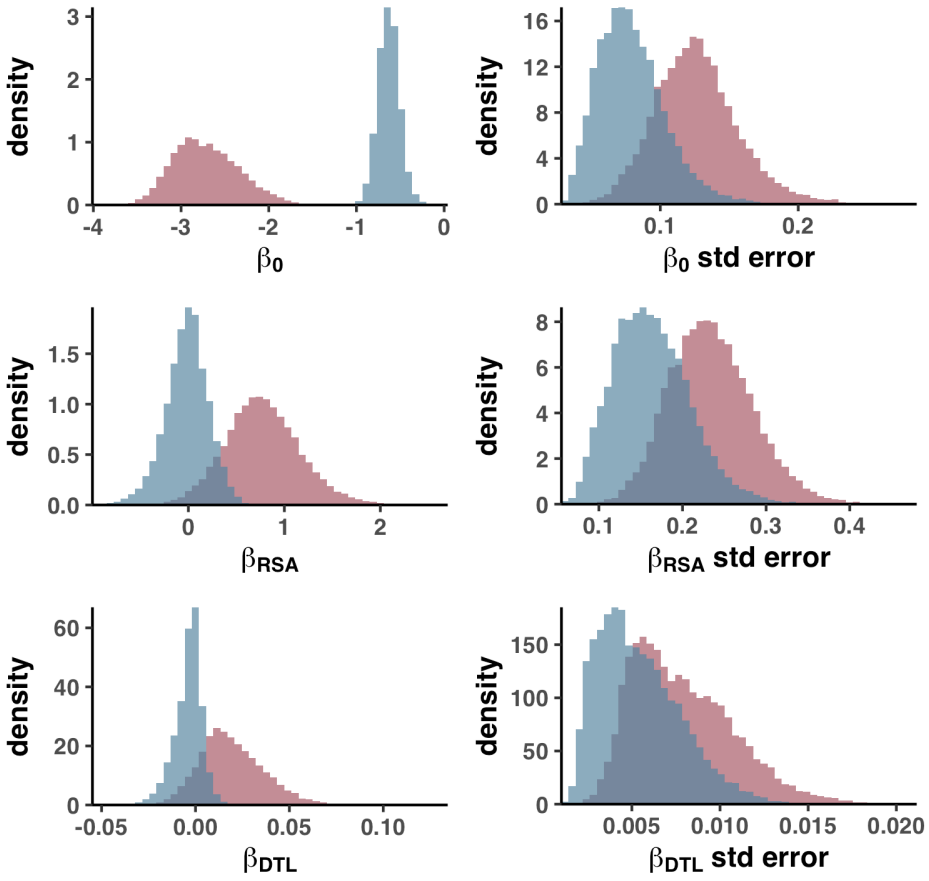

Figure S14. Parameter estimate and standard error distributions of the multidimensional linear regression models for $pN^{(site)}$ and $pS^{(site)}$. Red denotes parameter/error distributions for the 16,285 nonsynonymous models of the form $pN^{(site)} = \beta_0 + \beta_{RSA}RSA + \beta_{DTL}DTL$ and blue denotes parameter/error distributions for the 24,553 models of the form $pS^{(site)} = \beta_0 + \beta_{RSA}RSA + \beta_{DTL}DTL$.

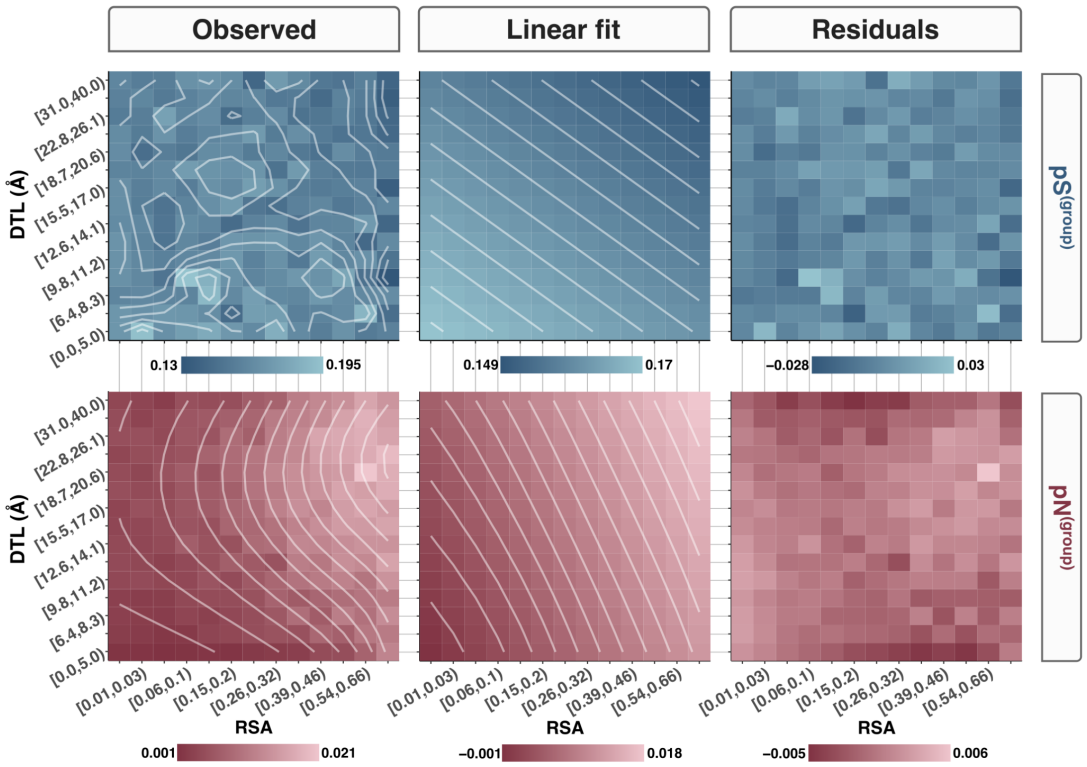

Figure S15. Observations, fits, and residuals of linear regressions for $pN^{(group)}$, $pS^{(group)}$, and $pN/pS^{(group)}$. The x-axis and y-axis for each heatmap are RSA and DTL groups, respectively. The first column shows the observed values (those seen in Figure 2e), the second column shows the 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.

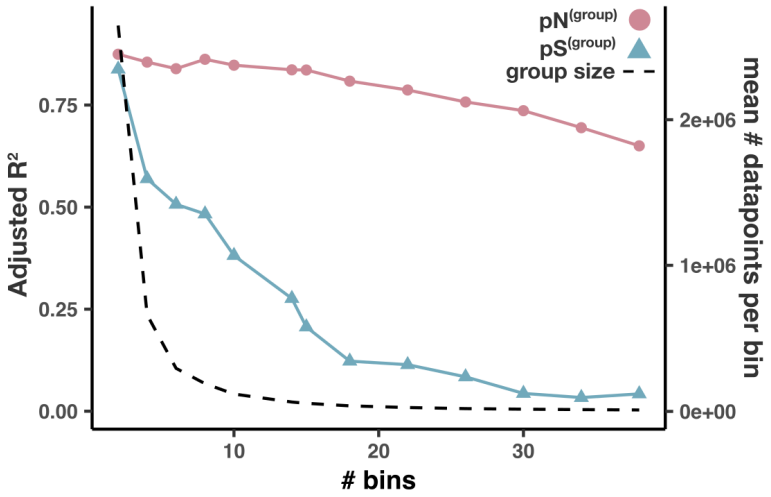

Figure S16. Model quality decreases for $pN^{(site)}$ and $pS^{(group)}$ as the number of RSA and DTL groups increases. The x-axis represents how many bins RSA and DTL are each split into. For example, the heatmaps in Figure 2e correspond to # bins = 15, since RSA and DTL are split into 15 bins, totaling 225 (=15x15) groups. The left y-axis corresponds to the adjusted R^2 value for the models $pN^{(group)}$ (red) and $pS^{(group)}$ (blue). The right y-axis corresponds to the average number of data points (# sites multiplied by # samples) found in a group (dashed black line).

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

To validate our $pN/pS^{(gene)}$ calculations, we ascribed a sample-averaged $pN/pS^{(gene)}$ value to each gene and compared the values to $dN/dS^{(gene)}$ (Table S11), 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 $dN/dS^{(gene)}$ 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 $pN/pS^{(gene)}$ and $dN/dS^{(gene)}$ were cross-contaminated due to HIMB83 recruiting significant proportions of reads from HIMB122-like populations. We found that log-transformed sample-averaged $pN/pS^{(gene)}$ highly correlated with log-transformed $dN/dS^{(gene)}$ (Pearson R^2 = 0.380), showing that the two metrics are commensurable. Nevertheless, differences were expected and observed. The ratio between sample-averaged $pN/pS^{(gene)}$ and $dN/dS^{(gene)}$ was on average 6.23 (Figure S10), matching expectations that slightly deleterious, nonsynonymous mutants commonly drift to observable frequencies, yet far less commonly drift to fixation.

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

Sample-averaged $pN/pS^{(gene)}$ values varied significantly between genes, varying from 0.004-0.539, with a mean of 0.063 (Figure S12, Table S9). What causes such variation in purifying selection strengths? Across diverse taxa (*89*), 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 (*90*). 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 S17a, Table S12).

Comparing sample-median TA values to sample-averaged $pN/pS^{(gene)}$ values yielded a strong, negative correlation (Figure S17b, Pearson r = -0.539, R^2 = 0.290) according to an inverse power-law relationship. The specific form of the linear model used was $\log_{10}(\text{median}_s(TA)+0.01) \sim \log_{10}(\text{mean}_s(pN/pS^{(gene)}))$, where median_s and mean_s 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 $\text{median}_s(TA)$. 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 (*90*). 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 $pN/pS^{(gene)}$ 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 $\log_{10}(TA+0.01)$ and $\log_{10}(pN/pS^{(gene)})$ (Figure S17c), yet only 11.5% of genes passed significance tests (one-sided Pearson, 25% Benjamini-Hochberg false discovery rate) (Figure S17d). 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 $pN/pS^{(gene)}$ 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 $pN/pS^{(gene)}$. These fluctuations therefore muddy signals that may exist between $pN/pS^{(gene)}$ and TA. Smoothing these fluctuations by averaging across samples thereby reveals the strong negative correlation observed (Figure S17b). In other words, TAs that are not averaged across environments are unreliable proxies for overall transcription level.

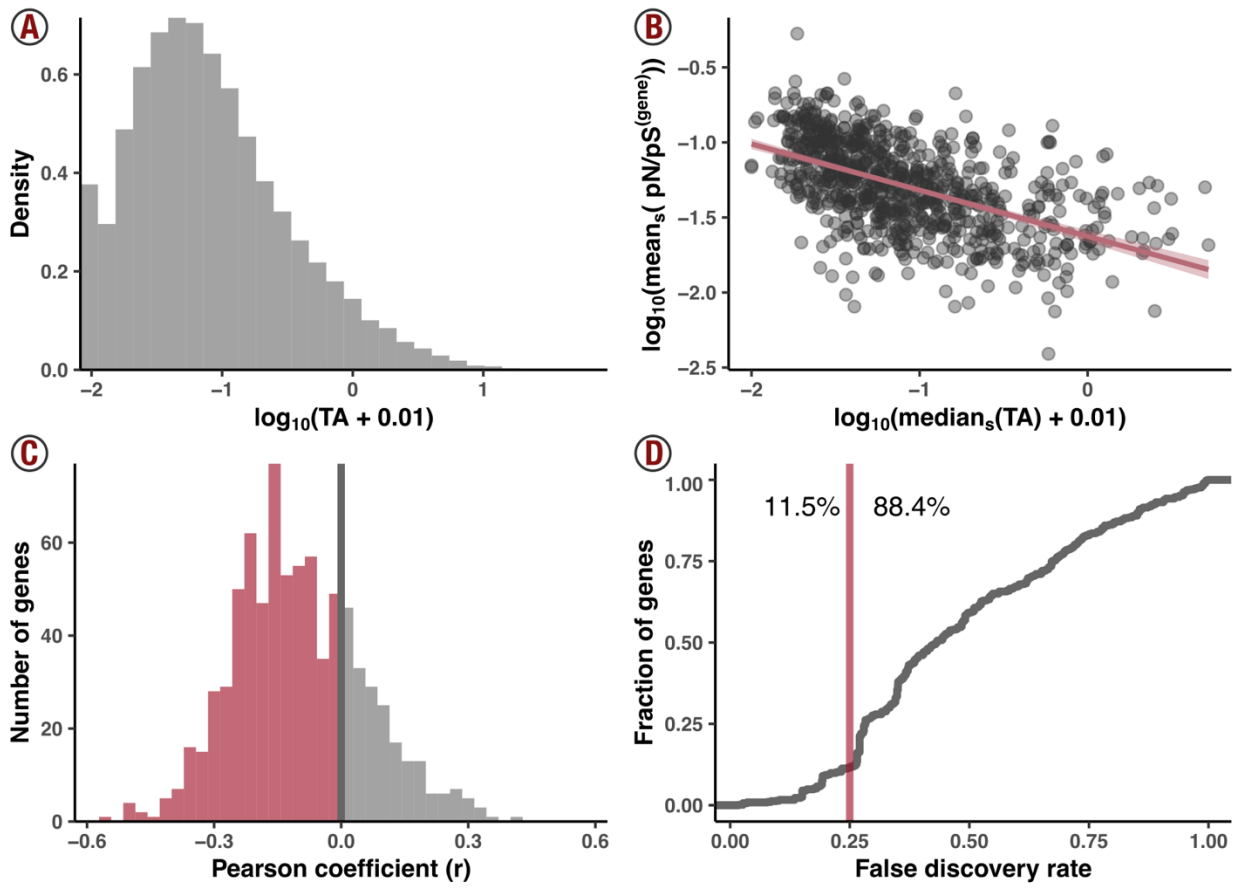

Figure S17. Associations of transcript abundance (TA) data with pN/pS$^{(gene)}$. (A) Log-transformed distribution of TA values across genes and samples. See Methods for details on TA calculation. 0.01 has been added to the log-transformation to avoid the exclusion of zeros. **(B) TA is a strong predictor of pN/pS$^{(gene)}$ 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$^{(gene)}$, 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$^{(gene)}$ 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$^{(gene)}$) 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$^{(gene)}$).

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 $pN/pS^{(core)}$ 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 $pN/pS^{(core)}$), ns-polymorphism throughout the genome distributed (a) in more solvent-exposed sites (Figure 4a) and (b) farther from predicted binding sites (Figure 4b). Our bootstrapping stability analysis (Figure S18, Table S13) showed that in 99.5% of gene resamplings, the mean RSA of ns-polymorphism negatively associated with $pN/pS^{(core)}$ (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 $pN/pS^{(core)}$. This latter finding indicates that the signal in Figure 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.

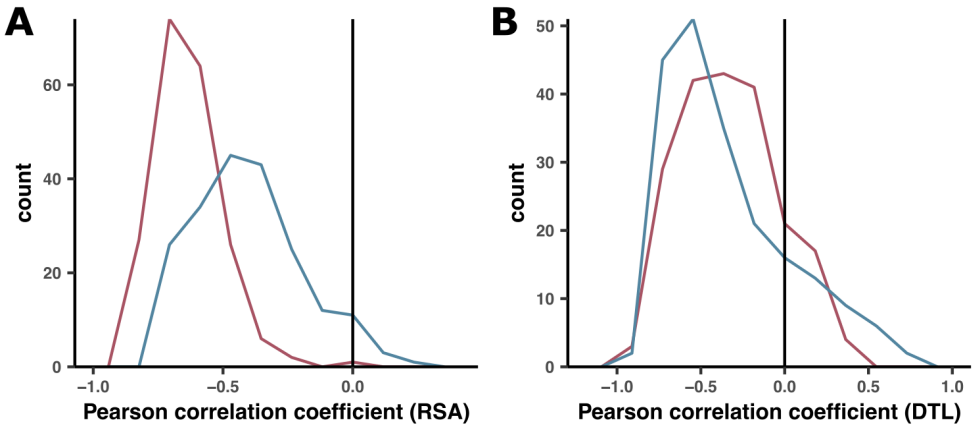

Figure S18. Robustness of negative associations between sample selection strength ($pN/pS^{(core)}$) and mean RSA/DTL of polymorphisms. We tested the robustness of results in Figure 4 by performing a bootstrapping stability analysis in which we created 200 bootstrapped estimates of the correlation coefficients, 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 $pN/pS^{(core)}$. These correspond to Figures 4a and 4c, respectively. **(B)** Histograms of the correlation coefficients between the mean DTL of s-polymorphism (blue) and ns-polymorphism (red) versus $pN/pS^{(core)}$. These correspond to Figures 4b and 4d, respectively.

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 S19, S20, S21, S22). 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 (*91, 92*) 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 S19). 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 S20). 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 S20). These same variables can also dynamically change the color and size of individual variants (Figure S21). 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 S21). 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 (Schrödinger, LLC) for further graphical refinement or statistical analyses (Figure S22).

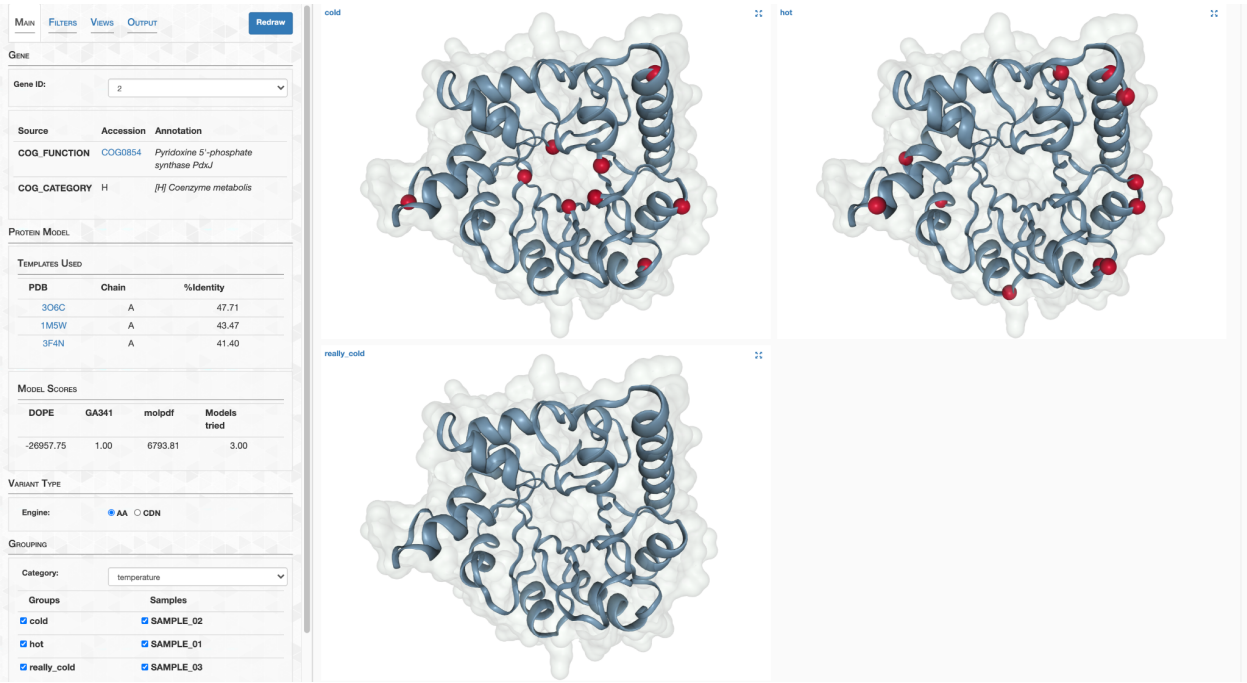

Figure S19. Screenshot of the interface with the "Main" tab active. The user has chosen to visualize Gene ID 2 from the left-hand side panel. Functional annotations from COG indicate this is a Pyridoxine 5'-phosphate synthase, and its structure was modeled using the PDB IDs 3O6C, 1M5W, and 3F4N templates. 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).

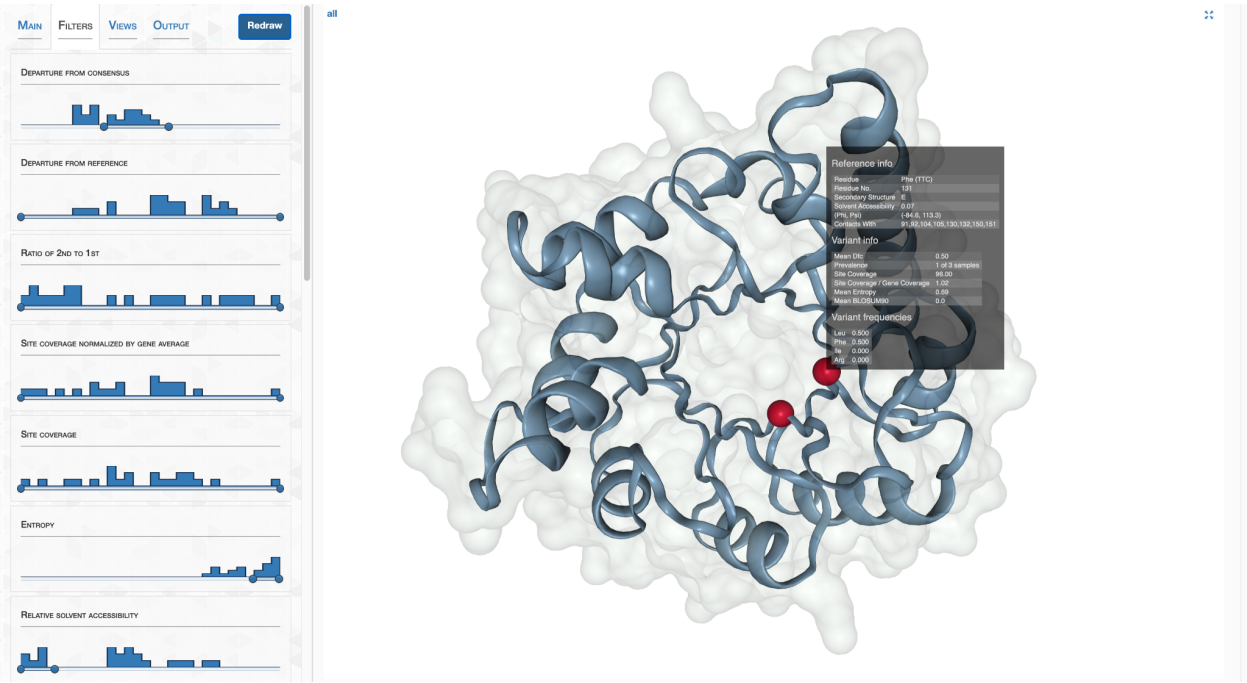

Figure S20. Screenshot of the interface with the "Filter" tab active. Variants can be filtered in the "Filters" tab, which shows a suite of filters, each represented as an interactive slider with endpoints that can be clicked and dragged by the user. Above each slider is a histogram detailing how the 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.

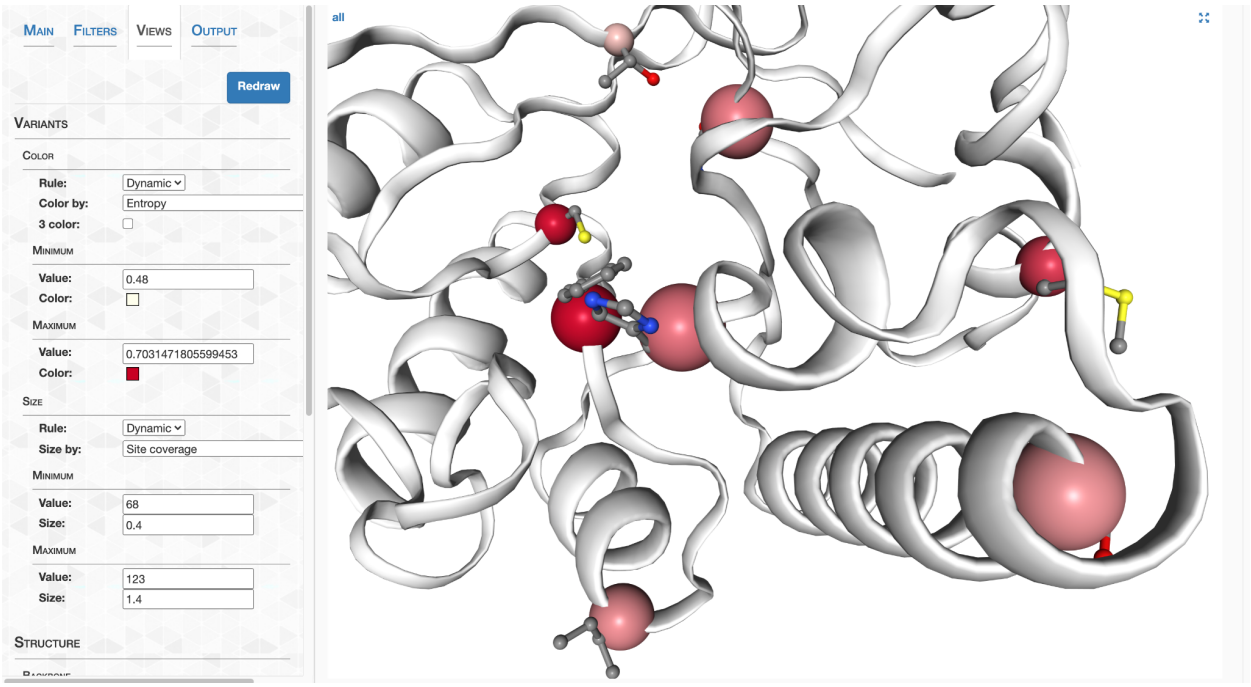

Figure S21. Screenshot of the interface with the "Views" tab active. In the “Views” tab, variants can be colored and sized according to variables. In this screenshot, the user has colored variants according to their entropy values on a linear gradient between white and red, and sized them according to their metagenomic coverage values.

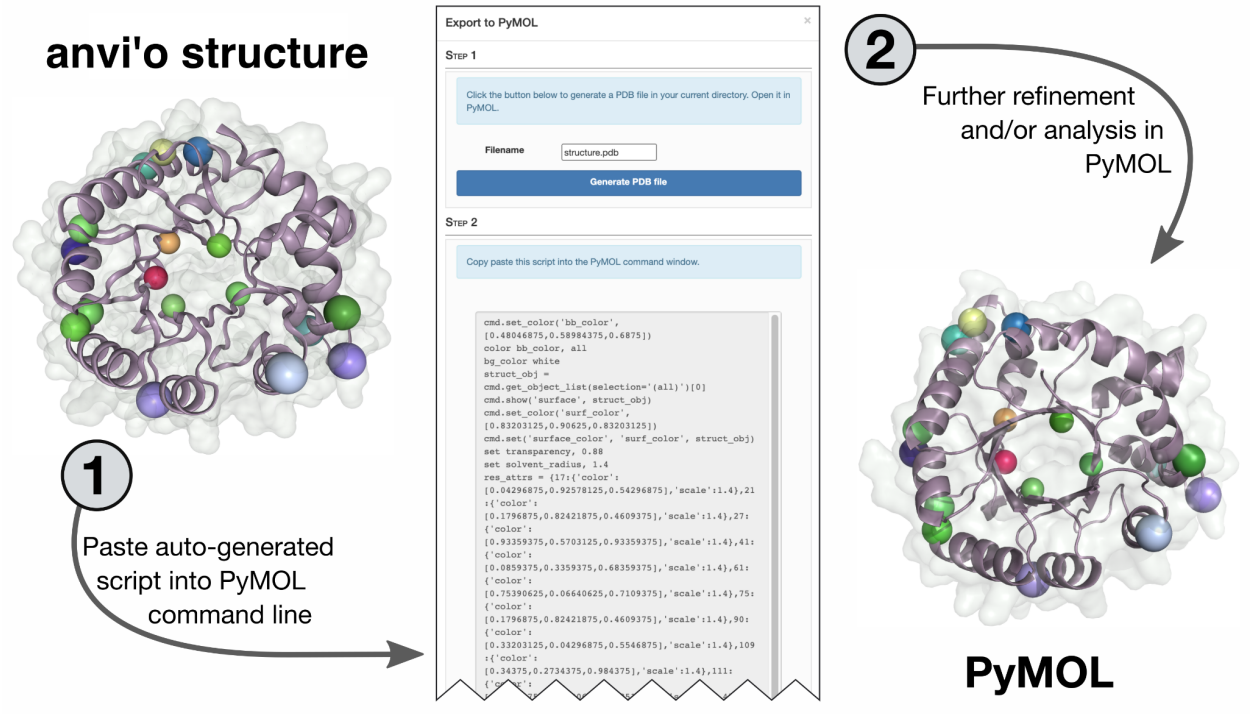

Figure S22. The interface can seamlessly migrate user sessions into PyMOL for visual refinement and more sophisticated analysis than is possible with `anvi-display-structure`. Under the “Output” tab, users can select “Generate in PyMOL” to auto-generate a script (middle) that when pasted into the PyMOL command line, reproduces the current interface view directly in PyMOL.

Supplementary Tables

Table S1. 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 S2. 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 S3. Mean per-site polymorphism rates ($pN_{(site)}$ and $pS_{(site)}$) 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 S4. 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 S5. 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 S6. Summary of models used for estimating the explanatory power of RSA and DTL on polymorphism rates (see Methods).

Table S7. Summary statistics for the polymorphism models of gene-sample pairs.

Table S8. Summary of per-group polymorphism data for **(A)** $pN_{(group)}$, **(B)** $pS_{(group)}$, **(C)** $pN/pS_{(group)}$, and **(D)** the size of each group.

Table S9. Summary of per-gene polymorphism data for **(A)** $pN/pS_{(gene)}$, **(B)** sample-averaged $pN/pS_{(gene)}$, **(C)** $pN_{(gene)}$, **(D)** $pS_{(gene)}$ and **(E)** the number of potential synonymous and nonsynonymous point mutations of each gene.

Table S10. Correlations of $pN/pS_{(gene)}$ for each 1a.3.V core gene with respect to the measured environmental parameters: nitrates, chlorophyll, temperature, salinity, phosphate, silicon, depth, and oxygen.

Table S11. Comparison between dN/dS between HIMB83 and HIMB122 homologs and sample-averaged $pN/pS_{(gene)}$ of 1a.3.V genes.

Table S12. Per sample and gene measures of transcript abundance (TA) and related quantities.

Table S13. Bootstrap estimates of Pearson correlation coefficients and p-values from Figure S18.

REFERENCES

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, Uncharted biosynthetic potential of the ocean microbiome. *bioRxiv* (2021).

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.