Dynamics of dN/dS within recombining bacterial populations

Zhiru Liu[1], Benjamin H. Good[1,2,3,†]

[1]*Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[2]*Department of Biology, Stanford University, Stanford, CA 94305, USA*
[3]*Chan Zuckerberg Biohub – San Francisco, San Francisco, CA 94158, USA*

[†] *Correspondence should be addressed to: B.H.G. (bhgood@stanford.edu)*

Abstract

The ratio of nonsynonymous to synonymous substitutions (dN/dS) encodes important information about the selection pressures acting on protein-coding genes. In bacterial populations, dN/dS often declines with the sequence divergence between strains, but the mechanisms responsible for this broad empirical trend are still debated. Existing models have primarily focused on de novo mutations, overlooking the older genetic variants that are continually introduced through horizontal gene transfer and recombination. Here we introduce a phenomenological model of dN/dS in recombining populations of bacteria, which allows us to disentangle the effects of recombination among pairs of closely related strains. We find that clonally inherited regions of the genome exhibit consistently higher dN/dS ratios, and that the accumulation of recombined segments can quantitatively explain the majority of the decline in dN/dS. We use these observations to re-examine models of purifying selection and adaptive reversion in human gut bacteria, and uncover evidence for widespread weak selection at a large fraction of protein coding sites. Our findings show that horizontal gene transfer can be an important factor in shaping genome-wide patterns of selective constraint, and raise new questions about the effectiveness of natural selection in complex bacterial populations.

Introduction

Patterns of DNA sequence conservation play a central role in modern evolutionary biology. In protein-coding genes, the degree of conservation is often quantified by comparing the rates of nonsynonymous and synonymous substitutions (dN/dS) (1). Under the assumption that synonymous mutations are effectively neutral, the dN/dS ratio provides a simple readout of the net direction of natural selection (2, 3): values less than one imply that many nonsynonymous variants have been eliminated by negative selection, while ratios greater than one imply that some of the observed substitutions were positively selected. This simple principle has widespread applications throughout evolutionary biology, from the identification of functionally important DNA sequences (4–6) to mapping the causative loci responsible for driving short-term evolutionary adaptation (7–10).

While originally developed for comparisons between species, dN/dS is also routinely applied to genomes from the same species (11, 12). In this case, the genetic differences between these

conspecific genomes represent polymorphisms that are circulating within the larger species population. An extensive literature has shown that the ratios of non-synonymous and synonymous variants at different allele frequencies can enable quantitative inferences of the underlying selection coefficients acting on new mutations (12–16), by partitioning the action of natural selection over different historical timescales. A similar approach is common in bacterial populations, where dN/dS can be computed for pairs of strains with different degrees of genetic relatedness. A widespread observation within many bacterial species is that dN/dS gradually decreases with increasing genetic divergence between strains, ranging from nearly neutral levels (dN/dS$\approx$1) among the most closely related strains to much lower levels (dN/dS$\approx$0.1) between typical pairs of conspecific isolates (8, 17–29). This robust trend suggests that dN/dS is time-dependent, with selection acting to reduce nonsynonymous variation as divergence time increases.

However, the mechanisms responsible for this time-dependent decay remain under debate. The classical explanation attributes the decline in dN/dS to the finite strength of purifying selection (22, 25, 28): weaker deleterious mutations take some time to be purged from the population, leading to an enrichment of non-synonymous variants among the most recently diverged strains. A more recent alternative proposes that the higher dN/dS values at short times arise from locally adaptive substitutions that undergo reversions on longer timescales, producing a long-term trend that is reminiscent of classical evolutionary constraint (30). While these two models rely on fundamentally different mechanisms, previous work suggests that they can both reproduce the observed dN/dS decay trends in multiple species of human gut bacteria (28–30), making them difficult to distinguish on the basis of these aggregate statistics alone.

Crucially, both of these existing explanations overlook a key process of bacterial evolution, where homologous recombination replaces a portion of the genome with a corresponding segment from another donor strain (31). Previous studies have shown that recombined regions in recently diverged genomes often exhibit lower dN/dS than clonal regions (24). This suggests that recombination could mimic the effects of natural selection by introducing older genetic variants that have already been filtered by selection. Given the widespread prevalence of homologous recombination within many bacterial species (32–35), this raises the possibility that the temporal trend in dN/dS could be strongly shaped by the accumulation of recombination events over time.

Here, we investigate this hypothesis by studying the impact of bacterial recombination on the dynamics of dN/dS. We show that a simple model in which recombination introduces divergent segments from unrelated donors can quantitatively recapitulate the observed decline in genome-wide dN/dS values within a large panel of human gut bacteria. After correcting for the confounding effects of recombination, we show how the residual dN/dS dynamics in clonally inherited regions can constrain existing models of purifying selection and adaptive reversion. These findings highlight the dominant role of recombination in shaping genome-wide patterns of selective constraint, and offer new insights into the dynamics of selection in natural bacterial populations.

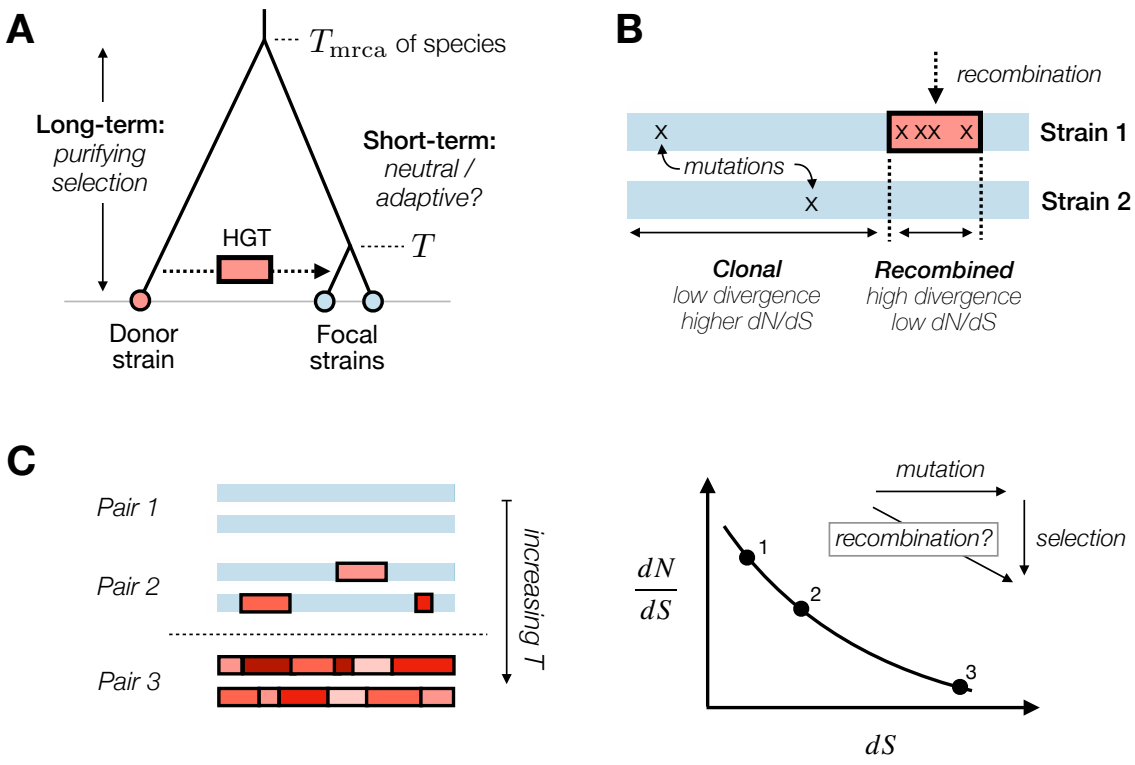

Figure 1: Modeling the impact of homologous recombination on dN/dS between conspecific strains of bacteria. (A) Pairs of recently diverged strains that shared a common ancestor T generations ago can acquire homologous DNA segments from more distantly related strains of the same species, which shared a common ancestor $\sim T_{\mathrm{mrca}}$ generations ago. (B) Recombined segments (red) introduce a high density of older genetic variants that have already been filtered by natural selection (low dN/dS), while clonally inherited regions (blue) exhibit lower genetic divergence and higher dN/dS from more recent de novo mutations. (C) The observed values of dN and dS in different pairs of strains (left) are influenced by their total proportion of clonal (blue) versus recombined (red) regions. The steady accumulation of recombination events produces a downward trend in dN/dS as strains diverge, potentially obscuring the underlying dynamics of natural selection.

Results

Quantifying the impact of homologous recombination on dN/dS dynamics

To illustrate how horizontal gene transfer and recombination can alter the temporal dynamics of dN/dS, we consider a simple conceptual model based on the genealogies of the underlying strains (33, 34, 36, 37; Fig. 1). We imagine a genome-wide comparison between two conspecific strains that descended from a single cell division event exactly T generations in the past (Fig. 1A). The genetic differences between these bacterial genomes will arise from two primary sources: (i) de novo mutations that are inherited through clonal descent, and (ii) pre-existing variants that are introduced via homologous recombination with other strains (Fig. 1B).

Under the assumption that synonymous mutations are effectively neutral, de novo mutations will generate single nucleotide differences (SNVs) at a characteristic rate of $d_S \approx 2\mu T$, where μ is the corresponding mutation rate in units of mutations per site per generation. In contrast, the synonymous divergence introduced by recombination will depend on the net rate of recombination per site (R), as well as the genetic divergence of the imported segments. Previous work has shown that these recombined DNA fragments often originate from unrelated donor strains (33, 34, 36, 37), which carry an average synonymous divergence of $\overline{d_S} = 2\mu T_{\mathrm{mrca}}$ (Fig. 1A). When $T \ll 1/R \ll T_{\mathrm{mrca}}$, the introduction of SNVs through recombination can substantially inflate d_S relative to the baseline expectation from de novo mutations alone. This hybrid divergence process continues until the majority of the clonal backbone has been overwritten by homologous recombination ($T \gtrsim 1/R$), after which d_S rapidly converges to the steady-state value $\overline{d_S}$.

This phenomenological model can also be extended to incorporate non-synonymous SNVs. In clonal regions of the genome, non-synonymous divergence accumulates at a time-dependent rate $d_N(T)$, whose detailed shape will depend on the underlying model of selection (described in more detail in the following section). Recombined segments, by contrast, will carry an average non-synonymous divergence of $\overline{d_N(T)}$, which can be estimated from the standing genetic diversity between unrelated pairs of strains (i.e. those with $T \gtrsim 1/R$). When $d_N(T)$ is a non-linear function of time, the value of dN/dS in the clonal regions of the genome will generally differ from the recombined segments. This reflects the fact that mutations on recombined segments have already been circulating in the population for a longer period of time, providing more opportunities for natural selection to act (Fig. 1A,B). As genomes accumulate more recombination events, the growing contribution from the recombined segments can alter the genome-wide average of dN/dS, similar to how they inflate the overall genetic divergence (Fig. 1C). Notably, this recombination process can generate a smoothly declining trend in the genome-wide dN/dS over time, even if the dN/dS ratios are largely constant within both the clonal and recombined regions.

To test whether this hypothesis could explain the dN/dS dynamics in empirical data, we leveraged a previously curated dataset of commensal human gut bacteria (28), containing 5,416 reconstructed strain genotypes from 43 prevalent species (SI Section 1). The genetic distances between these conspecific strains span nearly three orders of magnitude, from a typical value of $d_S \sim 1\%$ to just a handful of detected SNVs ($d_S \lesssim 10^{-5}$); the corresponding d_N/d_S levels vary over an order of magnitude as well (Fig. 2B). In a recent study (34), we used a hidden Markov model (HMM)

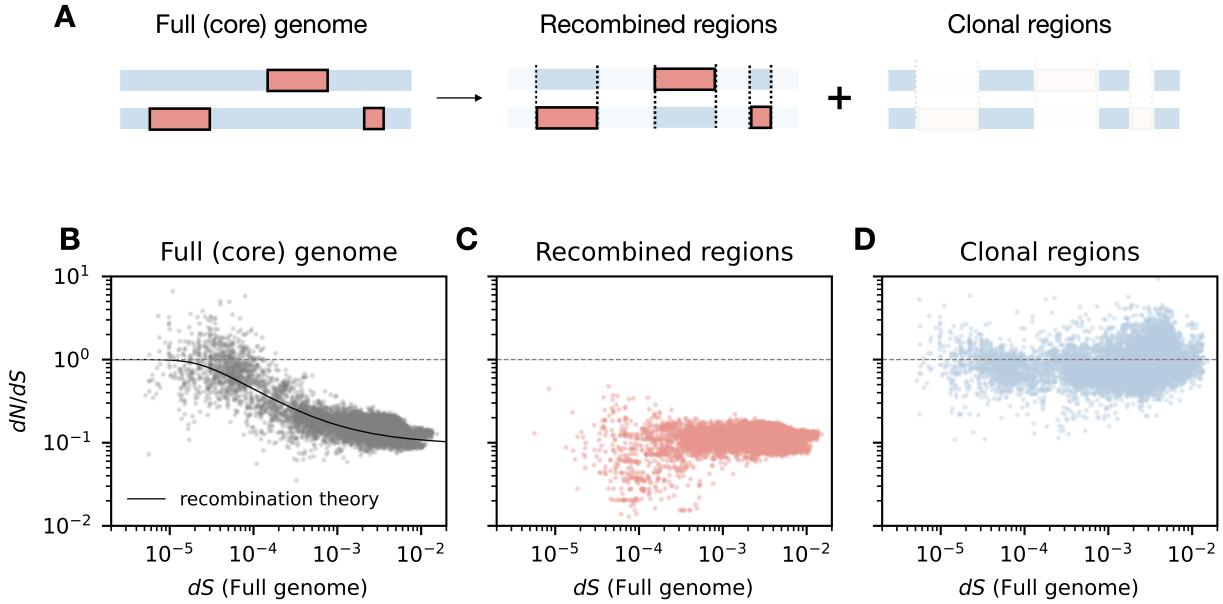

Figure 2: Temporal dynamics of dN/dS reflect the accumulation of homologous recombination events in human gut commensals. (A) Schematic illustration of how recombination events are isolated. Recombined regions (red blocks) and clonal regions (blue blocks) are identified based on the local density of synonymous SNVs. dN/dS values are then computed separately for each region type, as shown in panels C and D. (B) Genome-wide dN/dS dynamics in 29 human gut bacterial species reproduced from Ref. 28. Each point represents the genome-wide dN/dS value estimated from non-degenerate (1D) and fourfold-degenerate (4D) sites for an individual pair of closely related strains (SI Section 2). The solid line shows the prediction of the recombination model in SI Section 3, while the dashed line indicates the neutral expectation (dN/dS=1). (C-D) Analogous versions of panel B showing dN/dS values computed within the recombined (C) or clonal (D) of the genome separately.

approach to label individual recombination events and clonally inherited regions within the most closely related genomes in this dataset (Fig. 2A). Since our HMM method relies on the spatial distribution of synonymous variants, these labels should be independent of the corresponding d_N levels. The conceptual model in Fig. 1 then leads to a clear prediction for two different sources of data: if recombination is the primary driver of the dN/dS decay in Fig. 2B, then stratifying sites by their recombination status should reveal distinct dN/dS trajectories.

Consistent with this prediction, we found that the dN/dS dynamics in our dataset dramatically differ between clonal and recombined segments (Fig. 2C,D). Within recombined regions, dN/dS is consistently low ($d_N/d_S \sim 0.1$), closely matching the values observed between unrelated pairs of strains. In contrast, dN/dS remains close to one within the clonal regions, even as genome-wide divergence increases by several orders of magnitude.

To demonstrate how these local patterns give rise to the full genome-wide trend in Fig. 2B, we formulated a simple mathematical model of dN/dS accumulation that accounts for the growing contribution of recombined segments over time (Fig. S1; SI Section 3). Our phenomenological model simply averages the observed dN/dS values in the clonal and recombined regions, and weights them by the empirically determined fraction of the genome covered by recombination at each divergence level dS. As a result, the model has no additional free parameters that require fitting to the observed trend.

Despite its simplicity, we found that this mixture model quantitatively reproduces the observed dN/dS decay across the full range of divergence levels in Fig. 2B. In particular, it captures the characteristic transition from high to low dN/dS values at intermediate levels of synonymous divergence. In previous dN/dS models, the location of this transition was related to the underlying timescales of the natural selection (e.g., the selection coefficient in the purifying selection model of Ref. 28, or the reversion rate in the adaptive reversion model in Ref. 30). In our present model, this transition naturally emerges from the increasing genomic footprint of recombination, and occurs around the point when recombined segments begin to dominate the genome (Fig. 1C).

To explore whether these results extend beyond human gut bacteria, we also re-analyzed the dN/dS dynamics in 110 strains of *Staphylococcus aureus* (SI Section 7). *S. aureus* is a common bacterial pathogen with a distinct ecology and recombination dynamics compared to many gut commensals (38–42). Although clinically isolated strains of *S. aureus* are largely clonal in their recent evolutionary history, three large recombination events have been previously reported in sequence types ST34, ST239 and ST582 (42–44). We used these previously identified events to stratify the genome into clonal and recombined regions as above, and analyzed the corresponding dN/dS patterns in an analogous manner (Fig. S2). Once again, we found that dN/dS dynamics separate into distinct trends by genomic region, consistent with the predictions of our basic recombination model. These results suggest that recombination may be a general mechanism underlying the frequently observed temporal trend in dN/dS across diverse bacterial species.

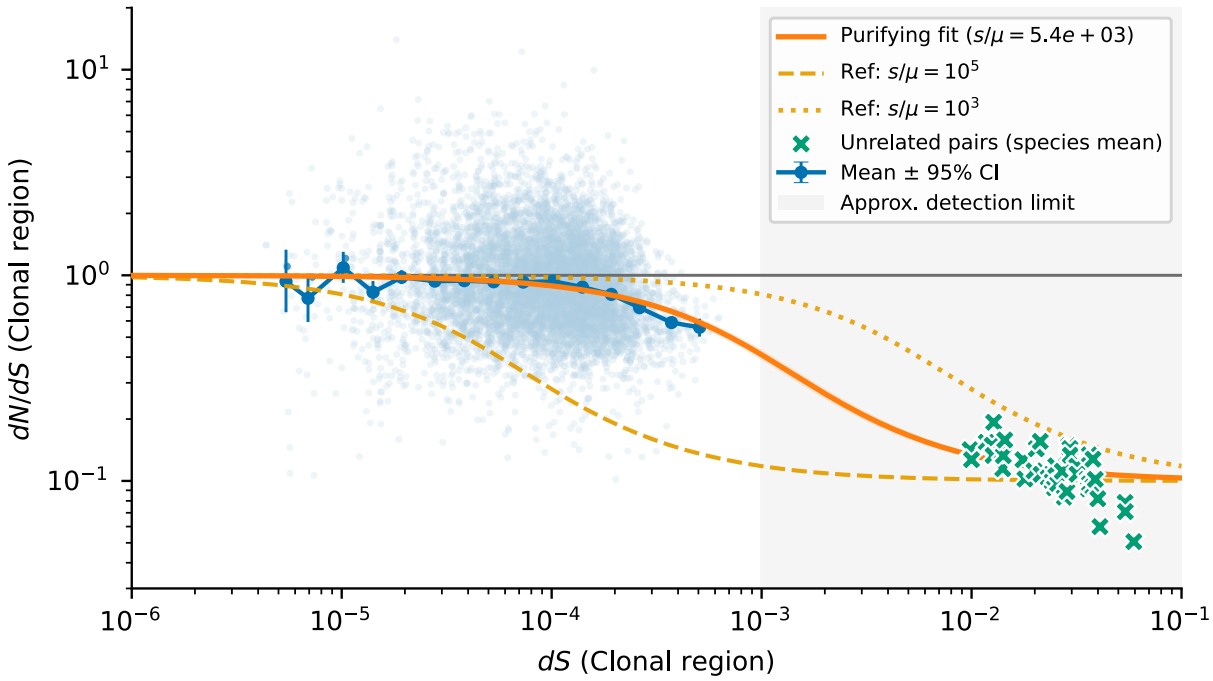

Figure 3: Clonal dN/dS dynamics in human gut bacteria after correcting for recombination. Light blue circles show the clonal d_N/d_S values for individual genome pairs (as in Fig. 2D) as a function of their clonal synonymous divergence (SI Section 2). Dark blue circles with error bars indicate the mean and 95% confidence intervals of pooled counts in logarithmically spaced bins of d_S (SI Section 2). Orange solid line shows the best-fit two-class purifying selection model (SI Section 4), with bootstrap confidence intervals narrower than the linewidth. For comparison, the dashed and dotted orange curves show reference predictions with $s/\mu = 10^5$ and 10^3, respectively, while the horizontal gray line marks the purely neutral expectation ($d_N/d_S = 1$). Green crosses denote the mean d_N/d_S and d_S values between fully recombined genome pairs for each individual species, which provide an estimate of the d_N/d_S levels when $T \approx T_{\mathrm{mrca}}$. The shaded vertical band indicates the approximate detection limit ($d_S > 10^{-3}$) beyond which most of the genome is overwritten by recombination and clonal tracts cannot be reliably identified (SI Section 1).

Estimating the strength of purifying selection from recombination-corrected dN/dS

While the mixture model in Fig. 2B can recapitulate the overall shape of the dN/dS curve, it still requires external estimates of the dN/dS levels in each region (Fig. 2C,D). These levels are set by the selective forces that act on nonsynonymous variants as they circulate through the broader population, which ultimately produce the low dN/dS values on a typical recombined segment. To probe these dynamics without the confounding influence of recombination, we can use the labels in Fig. 2 to construct a "corrected" dN/dS trajectory, using the local values of d_N and d_S in each genomic region.

The short-term dynamics of dN/dS can be assessed using the clonally inherited regions that are preserved between closely related genome pairs. Fig. 2D shows that the dN/dS levels on the clonal backbone remain close to one even as the nominal dS values span more than three orders of magnitude ($10^{-5} < d_S < 10^{-2}$; Fig. 2D). However, because these genome-wide dS levels are inflated by recombination, they overestimate the true divergence time between closely related strains.

To more accurately capture the temporal dynamics within the clonal regions, it is helpful to re-plot the clonal dN/dS values as a function of their *clonal* synonymous divergence $d_S^{(c)}$ (Fig. 3), which provides a better proxy for the elapsed divergence time between closely related pairs of genomes. The observed clonal divergences span a much narrower range ($10^{-5} < d_S^{(c)} < 10^{-3}$), which can help explain the relatively modest decay of dN/dS within these regions. However, the rate of the decay is still about $\sim$10-fold slower than the original data in Fig. 2, highlighting the importance of correcting for recombination. This global trend is also recapitulated at the level of individual species (Fig. S3), with most species showing only a modest decay from $d_N/d_S \approx 1$ as $d_S^{(c)}$ increases. A notable exception is *Alistipes putredinis*, which begins with elevated values ($d_N/d_S \approx 2$) at the shortest genetic distances and declines below one toward the end of the observable range. Apart from this example in *A. putredinis*, none of the other species showed consistently elevated dN/dS levels within the subset of sites that we are considering.

For more distantly related strains, recombination events eventually cover the entire genome (Fig. 1C), erasing clonal segments and precluding direct measurement of the clonal dN/dS curve over intermediate timescales. However, we can still estimate the long-term dN/dS levels by examining typical pairs of strains, whose genomes consist entirely of recombined segments with an average synonymous divergence $\overline{d_S} = 2\mu T_{\mathrm{mrca}} > 10^{-2}$. While the spread in this distribution is no longer informative of the differences in the underlying divergence times, the average values still provide a point estimate of the dN/dS levels when $T \approx T_{\mathrm{mrca}}$. Combining this estimate with the clonal regions above, we arrive at a corrected dN/dS curve with a gap at intermediate times (Fig. 3), corresponding to the transition from clonal to quasi-sexual evolution.

Having corrected for the confounding effects of recombination, we can use the residual dN/dS curves in Fig. 3 to re-examine the support for different models of selection. We begin by considering the standard purifying selection null model, which assumes that nonsynonymous variants are progressively removed over time (22, 28, 30). In this model, the clonal dN/dS curve follows the

approximate form,

$$\frac{d_N}{d_S} \approx \int ds\, \rho(s) \cdot \frac{1 - e^{-\frac{s\, d_S}{2\mu}}}{\frac{s\, d_S}{2\mu}}\,, \tag{1}$$

where $\rho(s)$ is the underlying distribution of fitness costs of nonsynonymous variants (SI Section 4). The key features of the data can be captured by a two-class version of this model, where a fraction of the sites are truly neutral ($s = 0$) and the rest have the same characteristic fitness cost s_1.

Fitting this model to the data in Fig. 3 yields a best-fit value of $s_1/\mu \approx 10^4$ and a corresponding neutral fraction of $\approx 10\%$ (SI Section 4). Similar values are obtained when *A. putredinis* is excluded (Fig. S4), indicating that the higher d_N/d_S values in this outlier species had little influence on the fitted parameters. These inferred fitness costs are about 10-fold smaller than previous estimates (28–30) obtained from the uncorrected data in Fig. 2 For typical bacterial mutation rates ($\mu \sim 10^{-10} - 10^{-9}$ per site per generation), the associated selection coefficients are remarkably weak ($s_1 \sim 10^{-6} - 10^{-5}$), and would be difficult to observe on normal experimental timescales. Nevertheless, Fig. 3 shows they are still sufficiently costly that they are eliminated by the time that T approaches $\overline{d_S}/2\mu \approx T_{\mathrm{mrca}}$.

These weak fitness costs are not an artifact of the two-class model in Fig. 3. Adding a third class of mutations only marginally improves the fit (Fig. S5), and the estimated target size of these mutations is comparatively small (<10%). A more general bound on the strength of selection can be obtained by leveraging the functional form of Eq. 1: the fact that dN/dS remains close to one for clonal divergences as large as $d_S^{(c)} \sim 10^{-4}$ implies that more than 90% of all nonsynonymous mutations must have fitness costs less than $s \sim 10^5 \cdot \mu \lesssim 10^{-4}$ (SI Section 4). At the same time, the fact that dN/dS eventually decays to ≈ 0.1 implies that most of these mutations are eventually purged, which requires that $s \gtrsim 2\mu/\overline{d_S} \approx 1/T_{\mathrm{mrca}}$. Taken together, these twin constraints imply that the purifying selection model in Eq. 1 can only explain the data if the fitness costs are weak but pervasive, with only a small minority of sites experiencing very strong ($s \gg 10^{-4}$) or nearly neutral ($s \ll 1/T_{\mathrm{mrca}}$) levels of selective constraint.

Mutation-type-specific dN/dS dynamics reveal stronger selection against nonsense variants

In the analysis above, all nonsynonymous mutations were aggregated together to compute the temporal dynamics of dN/dS. However, nonsense mutations, which introduce premature stop codons in the gene and often result in a loss of function, are generally expected to be subject to stronger selective pressures than missense mutations. If this is the case, then the model in Eq. 1 predicts that their clonal dN/dS trajectories should diverge from their missense mutation counterparts.

We tested this hypothesis by computing separate dN/dS curves for nonsense and missense mutations, respectively, and examined how each curve varies as a function of the clonal synonymous divergence $d_S^{(c)}$ (Fig. 4). The dN/dS decay of missense mutations closely resembled the aggregate results above, as expected given that they constitute the vast majority of nonsynonymous sites.

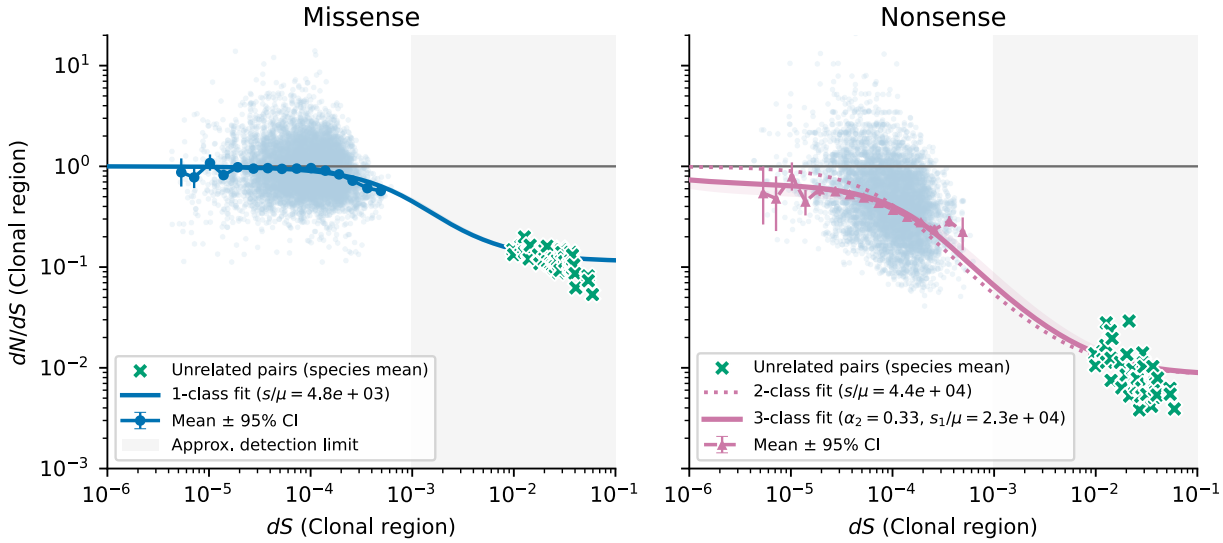

Figure 4: Analogous versions of Fig. 3 plotted for missense (left) and nonsense (right) mutations separately. In the left panel, the solid line denote the best-fit two-class purifying selection model (SI Section 4), with a 95% confidence interval smaller than the linewidth. In the right panel, the dashed line denotes the best-fit two-class model, while the solid line shows a three-class model with an additional class of strongly deleterious mutations with $s_2/\mu \approx 10^7$. These data suggest that nonsense mutations experience systematically higher fitness costs than missense mutations, but the average fitness cost of both mutation types remains comparatively weak.

In contrast, nonsense mutations showed a noticeably steeper decline in dN/dS with increasing divergence $d_S^{(c)}$, reaching a long-term value of ~ 0.01. This indicates that nonsense mutations are under stronger purifying selection, with a long-term neutral fraction of only ~ 0.01.

To quantify the underlying fitness costs, we fit these data to the same two- and three-class models that we used for the aggregate data above. The missense fits recovered nearly the same s/μ estimate as the aggregate data in Fig. 3. By contrast, we found that the nonsense mutations were better captured by the three-class model, which can account for $d_N/d_S < 1$ at short times (Fig. 4; SI Section 4). The inferred parameters suggest that roughly 30% of nonsense sites fall into a strongly deleterious class with $s_2/\mu \gg 10^5$, which are effectively purged before the earliest observable divergence times in Fig. 4 ($d_S \ll 10^{-5}$). Given that the total target size for nonsense mutations is about 20-fold lower than missense mutations, these strongly deleterious variants correspond to only a few percent of all nonsynonymous sites — consistent with our earlier estimates from the aggregate d_N/d_S dynamics in Fig. 3. Interestingly, for the remaining 60% of nonsense sites, the inferred selection coefficients are only about four-fold stronger than for missense sites, suggesting that many nonsense mutations are only weakly deleterious as well. These results demonstrate that stratifying dN/dS curves by mutation type can help resolve the heterogeneity in selection pressures across the genome: while most nonsynonymous mutations experience weak purifying selection, a subset of nonsense mutations are subject to strong constraint and are rapidly eliminated from the population.

Comparisons to alternative models of adaptive reversion

While the purifying selection model captures much of the observed dN/dS dynamics within the clonal regions, it is not the only possible explanation. One intruiging alternative was recently proposed in Ref. 30, in which the nonsynonymous differences could instead arise from locally adaptive substitutions that revert on longer timescales when the local environment changes. Previous work has shown that this adaptive reversion model can produce clonal dN/dS trajectories with the same functional form as Eq. 1, despite arising from a different set of biological assumptions and parameter meanings (30; see SI Section 5 for a full derivation). To understand the implications of this alternative scenario, we sought to determine how the new recombination-corrected measurements in Figs. 3 and 4 further constrain the underlying parameters of the model.

For example, the speed of the decay in the adaptive reversion model is controlled by the reversion timescale τ_{flip}, which is dominated by the time that it takes for the environment to shift to a state where the reverse mutation is preferred (30; SI Section 5). Using this mapping, the speed of the decay in Fig. 3 yields an associated reversion time of $\mu \cdot \tau_{\text{flip}} \equiv (s/2\mu)^{-1} \approx 4 \times 10^{-4}$, which translates to $\tau_{\text{flip}} \sim 300 - 3000$ years given current estimates of the mutation clock ($\mu \sim 10^{-7} - 10^{-6}$ per site per year; 8, 45). This revised timescale is significantly longer than a human lifetime, which places important constraints on the environmental processes that could be driving the shifts in the adaptive landscape.

A second key feature of the adaptive reversion model in Ref. 30 is that it predicts that the vast majority ($\sim$90%) of the clonal nonsynonymous SNVs in Fig. 3 must be comprised of reversible

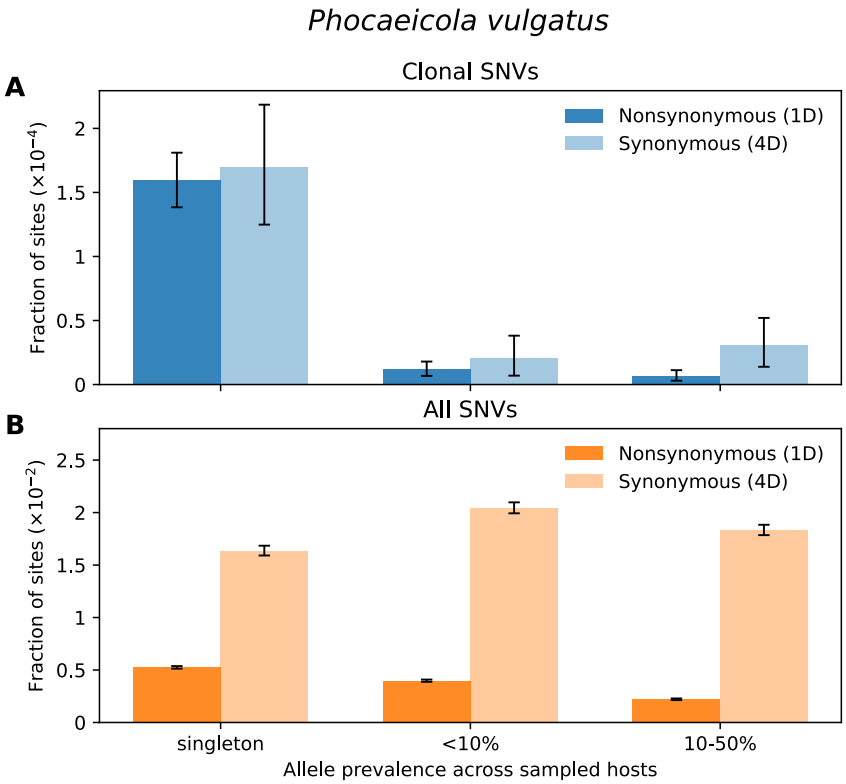

Figure 5: Measuring parallel evolution of clonal SNVs between closely related strains of *Phocaeicola vulgatus*. (A) Bar plots showing clonal SNVs in *P. vulgatus* stratified by the prevalence of the minor allele across other hosts in our cohort. Each bar shows the fraction of non-degenerate (1D; dark blue) or fourfold degenerate (4D; light blue) sites that harbor a SNV identified between fully clonal strain pairs (i.e., those with no detected recombination events). The singleton class refers to SNVs observed in only one of the 120 samples within the major clade of *P. vulgatus* (SI Section 6); other SNVs are grouped into two broad frequency ranges (<10% and >10%). These data show that most clonal SNVs are private to a single strain pair. (B) The same analysis applied to all SNVs in the dataset, most of which reside on recombined segments; the larger fraction of common alleles highlights the importance of stratifying SNVs by their recombination status. In both panels, error bars denote 95% confidence intervals assuming a Poisson distribution of observed SNV counts per bin. Analogous results for several additional species are shown in Fig. S8.

adaptive mutations. This large fraction of adaptive variants introduces an additional constraint linking the total number of adaptive sites and the degree of parallel evolution in other strains (SI Section 5). For example, if local adaptation was driven by a modest number of reversible adaptive sites, one would expect to see the same mutations independently arise in multiple pairs of strains. We quantified this signature of parallel evolution by identifying all purely clonal strain pairs (i.e. those with no detected recombination events) and asking how often the observed SNV differences between them were also found in other strains within our larger cohort (SI Section 6). Fig. 5A shows an example of this calculation for *Phocaeicola vulgatus*, one of the most prevalent gut commensals in our dataset. The data show that nearly 90% of all clonal nonsynonymous SNVs were private to a single clonal lineage and were never observed elsewhere among the 120 sampled hosts. Similar patterns are observed across the other species in our dataset (Fig. S8), with clonal SNVs rarely recurring in multiple strains (and at comparable rates between synonymous and nonsynonymous mutations). Stratifying SNVs by their recombination status is crucial for this calculation, since much higher rates of "parallelism" are observed when recombined sites are included (Figs. 5B and S8).

By extending the calculations in Ref. 30, we show that these low rates of parallel evolution can explain the dN/dS curves in Fig. 3 only if more than 1% of all nonsynonymous sites in the genome – corresponding to tens of thousands of sites – are targets of reversible adaptive evolution (SI Section 5). These numbers could be consistent with a smaller number of effective loci if mutations in the same locus interact with each other epistatically (e.g. loss-of-function mutations) (30). However, a similar calculation shows that at least $\sim$500 adaptive loci, each containing tens to hundreds of sites, are required to match the observed dN/dS curves in Fig. 3 (SI Section 5). This new estimate is an order of magnitude larger than the $\sim$50 adaptive loci previously estimated in Ref. 30, and reflects the $\sim$10-fold longer reversion timescale implied by the recombination-corrected data.

Finally, while the purifying selection model mechanically predicts that $d_N/d_S \to 1$ at short times, the adaptive reversion model can only match this feature of the data if the fraction of adaptive sites and their corresponding substitution rates are precisely tuned so that their product is of order μ (SI Section 5). Moreover, this condition must separately hold within each species in our dataset, as well as for different classes of sites (e.g. the missense and nonsense mutations in Fig. 4), where both the decay rates and fractions of adaptive sites can differ. The mechanisms that could give rise to such fine-tuning currently remain unclear.

Interestingly, the parallelism analysis in Fig. 5 also reveals a small subset of clonal mutations that are observed at higher prevalence within the broader population. Some of these mutations likely correspond to shorter recombination events that were not detected by our HMM algorithm. However, we also note that the dN/dS ratios of these recurrent clonal SNVs are somewhat elevated relative to the population average (Fig. 5B), which captures the long-term expectation for recombined sites. Although the small number of prevalent clonal SNVs limits statistical power, this relative enrichment of nonsynonymous variants raises the possibility that a small minority of clonal mutations may still correspond to adaptive reversions, consistent with prior direct observations (28, 46, 47). Nevertheless, the small overall fraction of these recurrent clonal SNVs in Figs. 5 and S8 suggests they cannot be driving the large-scale temporal trends in Fig. 3.

Discussion

The temporal dynamics of dN/dS carry important information about the selection pressures operating in natural bacterial populations. Here we have shown that in a broad range of human gut bacteria, the dN/dS ratios between closely related strains are dominated by the steady accumulation of recombination events, rather than the gradual unfolding of selective effects over time. By explicitly masking these recombination events, we demonstrated that the residual dN/dS levels in the clonally inherited portions of the genome remain informative, and can be used to constrain different models of selection. We found that a simple purifying selection null model can quantitatively reproduce the recombination-corrected dN/dS curves (Figs. 3 and 4), with estimated fitness costs that are about $\sim$10-fold lower than previous analyses of the non-recombination-corrected data (28–30). The implied fitness costs in this model are remarkably weak ($s \ll 10^{-4}$), and would be difficult to observe on the timescales of a laboratory experiment or a typical human generation time. Nevertheless, these nonsynonymous variants are still reliably purged on the longer evolutionary timescales separating typical pairs of strains from different hosts ($s \gtrsim 1/T_{\mathrm{mrca}}$).

How natural selection is able to optimize such weak fitness costs is still a major open question — particularly given the pervasive hitchhiking and selective bottlenecks that are observed in the gut microbiome on shorter evolutionary timescales (8, 28, 48–52). However, it is worth emphasizing that the selection coefficients in Eq. 1 are really effective fitness costs (53), which average over multiple rounds of host colonization, growth, and dispersal. In metapopulation contexts like the microbiome, it is currently unclear how much of this selection is driven by competition between individual cells, competition between co-colonizing lineages, or larger-scale ecological processes (e.g. multi-level selection between hosts; 54). Understanding how the long-term dynamics in Eq. 1 emerge from these local evolutionary processes is an important direction for future research — one that will likely require new advances in modeling selection and genetic linkage in metapopulation contexts.

Our recombination-corrected data also shed new light on the proposal that the declining dN/dS levels could be driven by a small number of adaptive loci that undergo frequent local fixations and reversions. Although this adaptive reversion model can be modified to accommodate our new data, it requires a much larger number of adaptive loci (>500) as well as the fine-tuning of multiple small parameters to match the nearly neutral dN/dS levels observed at short times. These arguments suggest that adaptive reversions may not fully explain the global dN/dS patterns in Fig. 3. However, their impact on microbiome evolution is still worth exploring. Our observation that some clonal SNVs are frequently observed in other hosts (Figs. 5 and S8) could hint at a smaller number of genuine reversion events. If so, these variants could serve as useful entry points for identifying the functional targets of adaptive reversions. It is also possible that recombination itself could be an important source of adaptive mutations (55) (and subsequent reversions), even if the recombined segments are not included in the clonal dN/dS curves in Fig. 3 (SI Section 5). Explicitly accounting for such processes remains an important avenue for future work.

More broadly, our findings highlight a fundamental limitation in using dN/dS dynamics to study selection in natural bacterial populations. Because recombination rapidly erases clonal segments

in many species of bacteria, the timescale over which clonal SNVs can be observed is sharply constrained. This leaves the critical regime, where dN/dS transitions from nearly neutral levels to its long-term equilibrium, as effectively unobservable with our present approach. To probe selection across these intermediate timescales, alternative population genetic measures such as the site frequency spectrum (56) or the decay of linkage disequilibrium (57–59) may provide complementary insights. Understanding how these population genetic observables map on to real-time evolutionary dynamics from longitudinal studies may ultimately help clarify how natural selection operates in this complex, multi-scale ecosystem.

References

[1] Ziheng Yang and Joseph P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503, December 2000. ISSN 0169-5347. doi: 10.1016/S0169-5347(00)01994-7.

[2] Motoo Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, May 1977. ISSN 1476-4687. doi: 10.1038/267275a0.

[3] Takashi Miyata and Teruo Yasunaga. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1):23–36, March 1980. ISSN 1432-1432. doi: 10.1007/BF01732067.

[4] Julian Echave, Stephanie J Spielman, and Claus O Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109–121, 2016.

[5] Lukas FK Kuderna, Jacob C Ulirsch, Sabrina Rashid, Mohamed Ameen, Laksshman Sundaram, Glenn Hickey, Anthony J Cox, Hong Gao, Arvind Kumar, Francois Aguet, et al. Identification of constrained sequence elements across 239 primate genomes. *Nature*, 625(7996): 735–742, 2024.

[6] Siwei Chen, Laurent C Francioli, Julia K Goodrich, Ryan L Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A Watts, Christopher Vittal, Laura D Gauthier, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625 (7993):92–100, 2024.

[7] Xavier Didelot, A Sarah Walker, Tim E Peto, Derrick W Crook, and Daniel J Wilson. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*, 14(3):150–162, 2016.

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

[9] Daniel J Wilson. Genomegamap: within-species genome-wide dn/ds estimation from over 10,000 genomes. *Molecular biology and evolution*, 37(8):2450–2460, 2020.

[10] Sean Whalen and Katherine S Pollard. Enhancer function and evolutionary roles of human accelerated regions. *Annual review of genetics*, 56:423–439, 2022.

[11] Sergey Kryazhimskiy and Joshua B. Plotkin. The Population Genetics of dN/dS. *PLoS Genetics*, 4(12):e1000304, December 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000304.

[12] David S Lawrie and Dmitri A Petrov. Comparative population genomics: power and principles for the inference of functionality. *Trends in Genetics*, 30(4):133–139, 2014.

[13] Adam Eyre-Walker, Megan Woolfit, and Ted Phelps. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2):891–900, 2006.

[14] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, 4(5):e1000083, 2008.

[15] Aaron P Ragsdale, Claudia Moreau, and Simon Gravel. Genomic inference using diffusion models and the allele frequency spectrum. *Current Opinion in Genetics & Development*, 53: 140–147, 2018.

[16] Bernard Y Kim, Christian D Huber, and Kirk E Lohmueller. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, 206 (1):345–361, 2017.

[17] I. King Jordan, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin. Microevolutionary Genomics of Bacteria. *Theoretical Population Biology*, 61(4):435–447, June 2002. ISSN 0040-5809. doi: 10.1006/tpbi.2002.1588.

[18] Lucy Baker, Tim Brown, Martin C. Maiden, and Francis Drobniewski. Silent Nucleotide Polymorphisms and a Phylogeny for Mycobacterium tuberculosis. *Emerging Infectious Diseases*, 10(9):1568–1577, September 2004. ISSN 1080-6040. doi: 10.3201/eid1009.040046.

[19] Timothy D. Read, Scott N. Peterson, Nicolas Tourasse, Les W. Baillie, Ian T. Paulsen, Karen E. Nelson, Hervé Tettelin, Derrick E. Fouts, Jonathan A. Eisen, Steven R. Gill, Erik K. Holtzapple, Ole Andreas Økstad, Erlendur Helgason, Jennifer Rilstone, Martin Wu, James F. Kolonay, Maureen J. Beanan, Robert J. Dodson, Lauren M. Brinkac, Michelle Gwinn, Robert T. DeBoy, Ramana Madpu, Sean C. Daugherty, A. Scott Durkin, Daniel H. Haft, William C. Nelson, Jeremy D. Peterson, Mihai Pop, Hoda M. Khouri, Diana Radune, Jonathan L. Benton, Yasmin Mahamoud, Lingxia Jiang, Ioana R. Hance, Janice F. Weidman, Kristi J. Berry, Roger D. Plaut, Alex M. Wolf, Kisha L. Watkins, William C. Nierman, Alyson Hazen, Robin Cline, Caroline Redmond, Joanne E. Thwaite, Owen White, Steven L. Salzberg, Brendan Thomason, Arthur M. Friedlander, Theresa M. Koehler, Philip C. Hanna, Anne-Brit Kolstø, and Claire M. Fraser. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature*, 423(6935):81–86, May 2003. ISSN 1476-4687. doi: 10.1038/nature01586.

[20] Michaela M Gutacker, James C Smoot, Cristi A Lux Migliaccio, Stacy M Ricklefs, Su Hua, Debby V Cousins, Edward A Graviss, Elena Shashkina, Barry N Kreiswirth, and James M Musser. Genome-Wide Analysis of Synonymous Single Nucleotide Polymorphisms in Mycobacterium tuberculosis Complex Organisms: Resolution of Genetic Relationships Among Closely Related Microbial Strains. *Genetics*, 162(4):1533–1543, December 2002. ISSN 1943-2631. doi: 10.1093/genetics/162.4.1533.

[21] Edward J. Feil, Jessica E. Cooper, Hajo Grundmann, D. Ashley Robinson, Mark C. Enright, Tony Berendt, Sharon J. Peacock, John Maynard Smith, Michael Murphy, Brian G. Spratt, Catrin E. Moore, and Nicholas P. J. Day. How Clonal Is Staphylococcus aureus? *Journal of Bacteriology*, 185(11):3307–3316, June 2003. doi: 10.1128/jb.185.11.3307-3316.2003.

[22] Eduardo P. C. Rocha, John Maynard Smith, Laurence D. Hurst, Matthew T. G. Holden, Jessica E. Cooper, Noel H. Smith, and Edward J. Feil. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, 239(2):226–235, March 2006. ISSN 0022-5193. doi: 10.1016/j.jtbi.2005.08.037.

[23] K. J. Balbi, E. P.C. Rocha, and E. J. Feil. The Temporal Dynamics of Slightly Deleterious Mutations in Escherichia coli and Shigella spp. *Molecular Biology and Evolution*, 26(2):345–355, February 2009. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msn252.

[24] Santiago Castillo-Ramírez, Simon R. Harris, Matthew T. G. Holden, Miao He, Julian Parkhill, Stephen D. Bentley, and Edward J. Feil. The Impact of Recombination on dN/dS within Recently Emerged Bacterial Clones. *PLoS Pathogens*, 7(7):e1002129, July 2011. ISSN 1553-7374. doi: 10.1371/journal.ppat.1002129.

[25] Christian Rödelsperger, Richard A Neher, Andreas M Weller, Gabi Eberhardt, Hanh Witte, Werner E Mayer, Christoph Dieterich, and Ralf J Sommer. Characterization of Genetic Diversity in the Nematode Pristionchus pacificus from Population-Scale Resequencing Data. *Genetics*, 196(4):1153–1165, April 2014. ISSN 1943-2631. doi: 10.1534/genetics.113.159855.

[26] Jessica K Calland, Ben Pascoe, Sion C Bayliss, Evangelos Mourkas, Elvire Berthenet, Harry A Thorpe, Matthew D Hitchings, Edward J Feil, Jukka Corander, Martin J Blaser, et al. Quantifying bacterial evolution in the wild: a birthday problem for campylobacter lineages. *PLoS genetics*, 17(9):e1009829, 2021.

[27] Sara L Loo, Anna Ong, Wunna Kyaw, Loïc M Thibaut, Ruiting Lan, and Mark M Tanaka. Nonsynonymous polymorphism counts in bacterial genomes: a comparative examination. *Applied and environmental microbiology*, 87(1):e02002–20, 2020.

[28] Nandita R. Garud, Benjamin H. Good, Oskar Hallatschek, and Katherine S. Pollard. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biology*, 17(1):e3000102, January 2019. ISSN 1545-7885. doi: 10.1371/journal.pbio.3000102.

[29] William R Shoemaker, Daisy Chen, and Nandita R Garud. Comparative population genetics in the human gut microbiome. *Genome Biology and Evolution*, 14(1):evab116, 2021.

[30] Paul A Torrillo and Tami D Lieberman. Reversions mask the contribution of adaptive evolution in microbiomes. *eLife*, 13:e93146, September 2024. ISSN 2050-084X. doi: 10.7554/eLife.93146.

[31] William P. Hanage. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harbor Perspectives in Biology*, 8(7):a018069, July 2016. ISSN , 1943-0264. doi: 10.1101/cshperspect.a018069.

[32] Michiel Vos and Xavier Didelot. A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, 3(2):199–208, February 2009. ISSN 1751-7370. doi: 10.1038/ismej.2008.93.

[33] Thomas Sakoparnig, Chris Field, and Erik van Nimwegen. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife*, 10:e65366, January 2021. ISSN 2050-084X. doi: 10.7554/eLife.65366.

[34] Zhiru Liu and Benjamin H. Good. Dynamics of bacterial recombination in the human gut microbiome. *PLOS Biology*, 22(2):e3002472, February 2024. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002472.

[35] Ellis L Torrance, Corey Burton, Awa Diop, and Louis-Marie Bobay. Evolution of homologous recombination rates across bacteria. *Proceedings of the National Academy of Sciences*, 121 (18):e2316302121, 2024.

[36] William P Hanage, Christophe Fraser, and Brian G Spratt. The impact of homologous recombination on the generation of diversity in bacteria. *Journal of theoretical biology*, 239(2): 210–219, 2006.

[37] Purushottam D Dixit, Tin Yau Pang, F William Studier, and Sergei Maslov. Recombinant transfer in the basic genome of escherichia coli. *Proceedings of the National Academy of Sciences*, 112(29):9070–9075, 2015.

[38] Jodi A. Lindsay. Genomic variation and evolution of *Staphylococcus aureus*. *International Journal of Medical Microbiology*, 300(2):98–103, February 2010. ISSN 1438-4221. doi: 10.1016/j.ijmm.2009.08.013.

[39] Andreas F. Haag, J. Ross Fitzgerald, and José R. Penadés. Staphylococcus aureus in Animals. *Microbiology Spectrum*, 7(3):10.1128/microbiolspec.gpp3–0060–2019, May 2019. doi: 10.1128/microbiolspec.gpp3-0060-2019.

[40] Marta Matuszewska, Gemma G. R. Murray, Ewan M. Harrison, Mark A. Holmes, and Lucy A. Weinert. The Evolutionary Genomics of Host Specificity in Staphylococcus aureus. *Trends in Microbiology*, 28(6):465–477, June 2020. ISSN 0966-842X, 1878-4380. doi: 10.1016/j.tim.2019.12.007.

[41] Santiago Castillo-Ramírez, Jukka Corander, Pekka Marttinen, Mona Aldeljawi, William P. Hanage, Henrik Westh, Kit Boye, Zeynep Gulay, Stephen D. Bentley, Julian Parkhill, Matthew T. Holden, and Edward J. Feil. Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. *Genome Biology*, 13 (12):R126, December 2012. ISSN 1474-760X. doi: 10.1186/gb-2012-13-12-r126.

[42] Xavier Didelot and Daniel J. Wilson. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology*, 11(2):e1004041, February 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004041.

[43] D. Ashley Robinson and Mark C. Enright. Evolution of Staphylococcus aureus by Large Chromosomal Replacements. *Journal of Bacteriology*, 186(4):1060–1064, February 2004. doi: 10.1128/JB.186.4.1060-1064.2004.

[44] R. R. Miller, J. R. Price, E. M. Batty, X. Didelot, D. Wyllie, T. Golubchik, D. W. Crook, J. Paul, T. E. A. Peto, D. J. Wilson, M. Cule, C. L. C. Ip, N. P. J. Day, C. E. Moore, R. Bowden, and M. J. Llewelyn. Healthcare-associated outbreak of meticillin-resistant Staphylococcus aureus bacteraemia: Role of a cryptic variant of an epidemic clone. *Journal of Hospital Infection*, 86 (2):83–89, February 2014. ISSN 0195-6701, 1532-2939. doi: 10.1016/j.jhin.2013.11.007.

[45] Matthew M Carter, Zhiru Liu, Matthew R Olm, Melanie Martin, Daniel D Sprockett, Benjamin C Trumble, Hillard Kaplan, Jonathan Stieglitz, Daniel Eid Rodriguez, David A Relman, et al. Prehistoric global migration of vanishing gut microbes with humans. *bioRxiv*, pages 2025–08, 2025.

[46] Ana Sousa, Ricardo S Ramiro, João Barroso-Batista, Daniela Güleresi, Marta Lourenço, and Isabel Gordo. Recurrent reverse evolution maintains polymorphism after strong bottlenecks in commensal gut bacteria. *Molecular biology and evolution*, 34(11):2879–2892, 2017.

[47] Alexandra J Poret, Matthew Schaefers, Christina Merakou, Kathryn E Mansour, Connor D Ahern, Georgia K Lagoudas, Alyssa Haynes, Ashley R Cross, Joanna B Goldberg, Roy Kishony, et al. De novo mutations mediate phenotypic switching in an opportunistic human lung pathogen. *Nature Communications*, 16(1):6799, 2025.

[48] João Barroso-Batista, Ana Sousa, Marta Lourenço, Marie-Louise Bergman, Daniel Sobral, Jocelyne Demengeot, Karina B Xavier, and Isabel Gordo. The first steps of adaptation of escherichia coli to the gut are dominated by soft sweeps. *PLoS genetics*, 10(3):e1004182, 2014.

[49] Morteza Roodgar, Benjamin H Good, Nandita R Garud, Stephen Martis, Mohan Avula, Wenyu Zhou, Samuel M Lancaster, Hayan Lee, Afshin Babveyh, Sophia Nesamoney, et al. Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. *Genome research*, 31(8):1433–1446, 2021.

[50] Daisy W Chen and Nandita R Garud. Rapid evolution and strain turnover in the infant gut microbiome. *Genome Research*, 32(6):1124–1136, 2022.

[51] Katherine S Xue, Sophie Jean Walton, Doran A Goldman, Maike L Morrison, Adrian J Verster, Autumn B Parrott, Feiqiao Brian Yu, Norma F Neff, Noah A Rosenberg, Benjamin D Ross, et al. Prolonged delays in human microbiota transmission after a controlled antibiotic perturbation. *bioRxiv*, 2023.

[52] Eitan Yaffe, Les Dethlefsen, Arati V Patankar, Chen Gui, Susan Holmes, and David A Relman. Brief antibiotic use drives human gut bacteria towards low-cost resistance. *Nature*, pages 1–10, 2025.

[53] Benjamin H Good and Oskar Hallatschek. Effective models and the search for quantitative principles in microbial evolution. *Current opinion in microbiology*, 45:203–212, 2018.

[54] Bob Week, Shelbi L. Russell, Hinrich Schulenburg, Brendan J. M. Bohannan, and Marjolein Bruijning. Applying evolutionary theory to understand host–microbiome evolution. *Nature*

Ecology & Evolution, September 2025. ISSN 2397-334X. doi: 10.1038/s41559-025-02846-w. URL https://doi.org/10.1038/s41559-025-02846-w.

[55] Brian J Arnold, I-Ting Huang, and William P Hanage. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4):206–218, 2022.

[56] Jonathan C Mah, Kirk E Lohmueller, and Nandita R Garud. Inference of the demographic histories and selective effects of human gut commensal microbiota over the course of human history. *Molecular Biology and Evolution*, 42(2):msaf010, 2025.

[57] Benjamin H Good. Linkage disequilibrium between rare mutations. *Genetics*, 220(4):iyac004, 2022.

[58] Anastasia S Lyulina, Zhiru Liu, and Benjamin H Good. Linkage equilibrium between rare mutations. *Genetics*, 228(3):iyae145, 2024.

[59] Richard Wolff and Nandita R Garud. Pervasive selective sweeps across human gut microbiomes. *BioRxiv*, pages 2023–12, 2024.

[60] Stephen Nayfach, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S. Pollard. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research*, October 2016. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.201863.115.

[61] Marina García-López, Jan P. Meier-Kolthoff, Brian J. Tindall, Sabine Gronow, Tanja Woyke, Nikos C. Kyrpides, Richard L. Hahnke, and Markus Göker. Analysis of 1,000 Type-Strain Genomes Improves Taxonomic Classification of Bacteroidetes. *Frontiers in Microbiology*, 10, 2019. ISSN 1664-302X.

[62] Vaishali Katju and Ulfar Bergthorsson. Old Trade, New Tricks: Insights into the Spontaneous Mutation Process from the Partnering of Classical Mutation Accumulation Experiments with High-Throughput Genomic Approaches. *Genome Biology and Evolution*, 11(1):136–165, January 2019. ISSN 1759-6653. doi: 10.1093/gbe/evy252.

[63] Xiaofang Jiang, A Brantley Hall, Timothy D Arthur, Damian R Plichta, Christian T Covington, Mathilde Poyet, Jessica Crothers, Peter L Moses, Andrew C Tolonen, Hera Vlamakis, et al. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science*, 363(6423):181–187, 2019.

[64] Tami D Lieberman, Jean-Baptiste Michel, Mythili Aingaran, Gail Potter-Bynoe, Damien Roux, Michael R Davis, David Skurnik, Nicholas Leiby, John J LiPuma, Joanna B Goldberg, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature genetics*, 43(12):1275–1280, 2011.

[65] Jianglin Zhou, Hongguang Ren, Mingda Hu, Jing Zhou, Beiping Li, Na Kong, Qi Zhang, Yuan Jin, Long Liang, and Junjie Yue. Characterization of burkholderia cepacia complex core genome and the underlying recombination and positive selection. *Frontiers in Genetics*, 11: 506, 2020.

[66] Mohammad Abdul Bakir, Mitsuo Sakamoto, Maki Kitahara, Mitsuharu Matsumoto, and Yoshimi Benno. Bacteroides dorei sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, 56(7):1639–1643, 2006. ISSN 1466-5034. doi: 10.1099/ijs.0.64257-0.

[67] Richard G. Everitt, Xavier Didelot, Elizabeth M. Batty, Ruth R. Miller, Kyle Knox, Bernadette C. Young, Rory Bowden, Adam Auton, Antonina Votintseva, Hanna Larner-Svensson, Jane Charlesworth, Tanya Golubchik, Camilla L. C. Ip, Heather Godwin, Rowena Fung, Tim E. A. Peto, A. Sarah Walker, Derrick W. Crook, and Daniel J. Wilson. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nature Communications*, 5 (1):3956, May 2014. ISSN 2041-1723. doi: 10.1038/ncomms4956.

Acknowledgments

We thank T. Lieberman and P. Torrillo for clarifications on the adaptive reversion model, and J. Ferrare, S. Walton, O. Ghosh, and J. McEnany for comments and feedback on the manuscript. This work was supported in part by a Stanford Bio-X Bowes Fellowship (to Z.L.), NIH NIGMS Grant No. R35GM146949 (to B.H.G.), NSF grant PHY-2309135, the Gordon and Betty Moore Foundation grant no. 2919.02, and the Chan Zuckerberg Initiative DAF grant to the Kavli Institute for Theoretical Physics (KITP). B.H.G. is a Chan Zuckerberg Biohub – San Francisco Investigator.

Data and Code Availability

All code used in this study, along with instructions for accessing the relevant datasets, are available at Github (`https://github.com/zhiru-liu/dNdS_dynamics`).

Author contributions: Conceptualization: Z.L. and B.H.G.; theory and methods development: Z.L. and B.H.G..; analysis: Z.L. and B.H.G.; writing: Z.L. and B.H.G.

Competing interests: The authors declare no competing interests.

Supplementary Methods

1 Metagenomic analysis of human gut commensals

The dN/dS measurements in the main text were obtained from collection of shotgun metagenomic data that have been analyzed in several previous works (28–30, 34). This dataset comprises of 932 fecal metagenomes from 693 individuals from North America, Europe, and China. All samples were processed using the same reference-based pipeline described in Ref. 28. In brief, we used the MIDAS software package (60) to align the raw sequencing reads to a panel of reference genomes representing common gut bacterial species, and identified single nucleotide variants (SNVs) based on the raw read pileups in each sample. In this work, we restricted our attention to SNVs in core genes (those that were present in more than 90% of samples in which the species was reliably detected). We used the quasi-phasing approach in Ref. 28 to identify intra-host populations in which the genotype of the dominant lineage could be inferred with a high-degree of confidence. Applying this method to the entire dataset yielded a collection of 5,416 strain genotypes from 43 species. Filtered genotype and coverage matrices for these quasi-phased genomes are provided in the Supplementary Data.

To identify synonymous and nonsynonymous variants, we focused on the subset of SNVs that were biallelic across the entire collection of quasi-phased genomes, ensuring that each site had a single putative ancestral allele and a single derived allele. Under the assumption that the major allele is ancestral, we used the reference genome sequence to determine site degeneracy (e.g., non-degenerate or fourfold) and to classify each mutation as synonymous, missense, or nonsense. These annotations were then used to tally the number of observed differences and opportunities in each mutation class.

Note on species names: The analyses in this study were performed using version 1.2 of the MIDAS reference genome database (60). Subsequent taxonomic updates have divided the *Bacteroides* genus into *Bacteroides* and *Phocaeicola* (61), leading to changes in species names relative to the original MIDAS database. For consistency, we retain the MIDAS species names in supplementary figures, while adopting updated nomenclature in the main text where appropriate. The corresponding name translations are listed below for reference:

MIDAS DB v1.2	NCBI Taxonomy Browser
Bacteroides coprocola	*Phocaeicola coprocola*
Bacteroides massiliensis	*Phocaeicola massiliensis*
Bacteroides plebeius	*Phocaeicola plebeius*
Bacteroides vulgatus	*Phocaeicola vulgatus*
Eubacterium eligens	*Lachnospira eligens*

2 Estimating dN/dS ratios from pairwise comparisons

To estimate the synonymous and nonsynonymous substitution rates between a given pair of genomes, we used the observed sequence divergence at fourfold-degenerate (4D; synonymous) and non-degenerate (1D; nonsynonymous) sites. These divergences were estimated from the

ratio of the total number of SNV differences (k_i) and the total number of covered sites (L_i) for each class of mutations. Since our goal is to examine how d_N/d_S varies as a function of d_S, it is important to avoid the mechanical correlation that arises when the sampling noise in d_S is shared between both the input and output variables. Following the Poisson thinning approach in Garud et al. (28), we randomly partitioned the synonymous sites into two disjoint sets of sizes L_S^A and L_S^B, with corresponding difference counts k_S^A and k_S^B. This yields a pair of synonymous divergence estimates for each a pair of strains,

$$d_S^A = \frac{k_S^A}{L_S^A}, \qquad d_S^B = \frac{k_S^B}{L_S^B}\,, \tag{S1}$$

which should be statistically independent of each other after conditioning on the underlying divergence time. We therefore computed d_N/d_S using d_S^B in the denominator,

$$d_N/d_S \approx \frac{k_N}{L_N} \Big/ \frac{k_S^B}{L_S^B}\,, \tag{S2}$$

and plotted this ratio as a function of the independent variable d_S^A.

When comparing closely related genome pairs from the gut commensal dataset, we further stratified sites by recombination status to compare d_N/d_S between clonal and recombined regions. We used the recombination tracts that were previously inferred for these same pairs of strains in Ref. 34, using a pairwise hidden Markov model that leverages the spatial profile of synonymous sequence divergence across the genome. Closely related genome pairs were defined empirically as those sharing more than 50% identical genome blocks, where each block corresponds to 1,000 fourfold-degenerate sites. For each such pair, we used the inferred tract boundaries to assign each site to either the clonal or recombined class, and then computed k_S, L_S, k_N, and L_N separately for each class before applying the Poisson thinning procedure described above.

While the above procedure provides d_N/d_S for individual genome pairs, we also sought to compute the average d_N/d_S value for all pairs of strains with a similar estimated divergence time. To do so, we first performed our Poisson thinning procedure on the clonal synonymous sites of each genome pair and used d_S^A to assign pairs into logarithmically spaced divergence bins. Within each bin, we pooled all synonymous and nonsynonymous sites across pairs and computed a "ratio-of-means" estimate of the average d_N/d_S within the bin:

$$d_N/d_S \approx \frac{\sum_{\text{pairs}} k_N}{\sum_{\text{pairs}} L_N} \Big/ \frac{\sum_{\text{pairs}} k_S^B}{\sum_{\text{pairs}} L_S^B}\,. \tag{S3}$$

Confidence intervals were obtained from 600 bootstrap replicates, resampling both genome pairs and Poisson partitions. The same procedure was used to estimate the long-term d_N/d_S for each species by analyzing a sample of 20 unrelated, fully recombined genome pairs.

Finally, to compare missense and nonsense substitution rates within clonal regions, we annotated the mutation class of each 1D SNV difference and calculated the total number of missense ($L_N^{(m)}$) and nonsense ($L_N^{(n)}$) opportunities, where $L_N^{(m)} + L_N^{(n)} = 3L_N$. We then repeated the Poisson thinning and bootstrap procedures separately for each mutation class. In all cases, d_N/d_S was

calculated using the synonymous substitution rate per synonymous *opportunity*, while the x-axis in the relevant figures (e.g., Fig. 4) shows the synonymous substitution rate per synonymous *site*.

Note on transition/transversion composition. Partitioning 1D differences into missense and nonsense classes slightly changes the neutral per–opportunity mutation rate because the underlying transition:transversion (*Ti:Tv*) mix differs across these opportunity sets. In the standard genetic code, the Ti:Tv composition of missense opportunities is nearly identical to that of 4D sites, and therefore the expected neutral mutation rate per opportunity is effectively the same. Nonsense opportunities, however, are modestly enriched for transversions, with a Ti fraction of roughly 0.27 based on the genetic code. Assuming a typical bacterial rate bias of $\mu_{\mathrm{Ti}}/\mu_{\mathrm{Tv}} \approx 4$ (62), this composition would yield a neutral mutation rate per nonsense opportunity that is about 10% lower than at 4D sites. Thus, while this mechanism can produce a nonsense-only d_N/d_S modestly less than one even under neutrality, values much lower than this expectation are indicative of the effect of selection.

3 Modeling the dynamics of genome-wide dN/dS with recombination

The phenomenological model in Fig. 2 assumes that each pair of bacterial genomes contains a mixture of clonally inherited and recently recombined genomic regions. We assume that in the clonal regions, sequence divergence arises solely through mutation, so the expected clonal synonymous divergence is given by

$$d_S^{(c)}(T) = 2\mu T\,, \tag{S4}$$

where μ is the per-site mutation rate and T is the divergence time between the two strains. In the simplest version of the model, we assume that the nonsynonymous clonal divergence is equal to $d_N^{(c)}(T) \approx d_S^{(c)}(T)$, though we also consider more complex generalizations in Sections 4 and 5 below.

In contrast, the recombined regions of the alignment are assumed to originate from unrelated donors in the population. As a result, the sequence divergence within the recombined segments reflects the long-term evolutionary distance between unrelated strains. In the simplest version of the model, we assume that these recombined divergences are drawn from a single distribution with means $\overline{d_S}$ and $\overline{d_N}$, respectively, which can be directly measured from the larger dataset using the typical divergence between unrelated pairs of genomes.

The expected fraction of the genome that is covered by recombination events (f_r) depends on divergence time T between each pair of strains. In our previous work (34), we showed that recombination dynamics in gut bacteria can be both heterogeneous across strains and nonlinear in time. To account for this complexity, we adopted an empirical approach to estimate the functional form of $f_r(T)$: we used our previously inferred values of f_r and $d_S^{(c)}$ for each pair of closely related strains and fit these data to a logistic function of the form:

$$\bar{f}_r(d_S^{(c)}) = \frac{1}{1 + e^{-k\left(\log(d_S^{(c)}) - \log(dS^*)\right)}}\,, \tag{S5}$$

where k controls the steepness of the transition and dS^* sets the characteristic divergence level at which recombination begins to dominate (Fig. S1). Using this functional form, the total expected divergence across the genome is a weighted average of clonal and recombined contributions:

$$dS(T) = d_S^{(c)}(T) \cdot \left[1 - \bar{f}_r\big(d_S^{(c)}(T)\big)\right] + \overline{d_S} \cdot \bar{f}_r\big(d_S^{(c)}(T)\big)\,, \tag{S6a}$$

$$dN(T) = d_N^{(c)}(T) \cdot \left[1 - \bar{f}_r\big(d_S^{(c)}(T)\big)\right] + \overline{d_N} \cdot \bar{f}_r\big(d_S^{(c)}(T)\big)\,. \tag{S6b}$$

The ratio of these two expressions defines the expected genome-wide d_N/d_S as a function of divergence time. This deterministic prediction is shown in Fig. 2B.

In contrast to previous models (28, 30), which have free parameters (e.g. mutation rates, selection coefficients, and/or reversion rates) that must be estimated from the observed d_N/d_S curve, the phenomenological model in Eq. S6 only depends on quantities that can be independently measured from other features of the data. Once $\overline{d_N}$, $\overline{d_S}$, and $\bar{f}_r(d_S^{(c)})$ are specified, there are no additional free parameters for the theoretical curve in Fig. 2.

However, while this phenomenological recombination model provides an explanation for the observed decay of genome-wide d_N/d_S, it does not explicitly model the dynamics of d_N/d_S within the clonal regions, or the values of $\overline{d_N}$ and $\overline{d_S}$ on the recombined segments. In the following two sections, we consider two classes of models that generate these predictions from different underlying mechanisms.

4 Purifying selection model

Multiple formulations of the purifying selection model describe how d_N/d_S changes as a function of synonymous divergence (d_S) at a locus evolving without recombination (22, 28, 30). These models converge on the common functional form in Eq. 1, which captures the impact of selection against nonsynonymous mutations with a distribution of deleterious fitness effects (DFE) $\rho(s)$. The shape of this curve is governed by the ratio s/μ, which sets the timescale over which a mutation of effect s is purged. For an arbitrary distribution $\rho(s)$, sites with different selection strengths are purged at different ranges of d_S, and the d_N/d_S curve reflects their relative contributions. Since the actual DFE is unknown, we used two different strategies to constrain this parameter from the observed data.

Estimating parameters for multi-class DFEs. Following previous work, we first considered discretized versions of the DFE, where nonsynonymous mutations fall into one of K discrete fitness classes with characteristic costs $\{s_i\}$ and relative target sizes $\{\alpha_i\}$. In the simplest version of this model with $K = 2$ classes, a fraction α_0 of nonsynonymous sites are truly neutral ($s_0 = 0$), and the remaining fraction $\alpha_1 = 1 - \alpha_0$ share the same characteristic fitness cost s_1. In this case, the d_N/d_S curve reduces to

$$\frac{d_N}{d_S} \;=\; \alpha_0 + (1-\alpha_0) \cdot \frac{1 - e^{-\frac{s_1 d_S}{2\mu}}}{\frac{s_1 d_S}{2\mu}}\,. \tag{S7}$$

which depends on the two compound parameters α_0 and s_1/μ. At long times ($d_S \gg \mu/s_1$), d_N/d_S asymptotically approaches α_0, which can be interpreted as the equilibrium fraction of neutral sites.

At short timescales ($d_S \ll \mu/s_1$), the second term approaches $1-\alpha_0$, and thus $d_N/d_S \to 1$ regardless of s_1, reflecting the fact that deleterious mutations have not yet been purged.

We also considered three-class versions of the model,

$$\frac{d_N}{d_S} \;=\; \alpha_0 + \alpha_1 \cdot \frac{1-e^{-\frac{s_1\, d_S}{2\mu}}}{\frac{s_1\, d_S}{2\mu}} + (1-\alpha_0-\alpha_1)\cdot \frac{1-e^{-\frac{s_2\, d_S}{2\mu}}}{\frac{s_2\, d_S}{2\mu}}\,. \tag{S8}$$

which include an additional fitness class with $s_2 \geq s_1$. This three-class formulation can capture cases like Fig. 4B where the short-time d_N/d_S is significantly below 1, indicating that a substantial fraction of nonsynonymous mutations are strongly deleterious.

We fit Eqns. S7 and S8 to the binned d_N/d_S data using weighted nonlinear least squares. As described in SI Section 2, we used the Poisson–thinning procedure to partition the synonymous sites into two disjoint sets, "A" and "B", with d_S^A used for bin assignment and d_S^B for the d_N/d_S calculation. For each divergence bin i, let $K_{N,i}$ and $K_{S,i}^B$ denote the total numbers of nonsynonymous and synonymous differences, respectively, obtained by pooling counts across all genome pairs in the bin for the "B" set of sites, and $L_{N,i}$ and $L_{S,i}^B$ the corresponding pooled site counts. The binned d_N/d_S estimate is then

$$\hat{R}_i = \frac{K_{N,i}/L_{N,i}}{K_{S,i}^B/L_{S,i}^B}\,. \tag{S9}$$

To account for varying statistical uncertainty across bins, we used inverse–variance weights,

$$w_i^{-1} = \frac{1}{K_{N,i}} + \frac{1}{K_{S,i}^B}\,, \tag{S10}$$

on a log-transformed version of $\hat{R}_i$, which down-weights bins with few observed differences. We excluded divergence bins containing fewer than three genome pairs to ensure that there was sufficient data for fitting. The resulting objective function is then defined by

$$\mathrm{wMSE}(\theta) = \frac{\sum_i w_i \left[\log \hat{R}_i - \log R_{\mathrm{model}}(d_{S,i}^A;\theta)\right]^2}{\sum_i w_i}\,, \tag{S11}$$

where $R_{\mathrm{model}}(d_S^A;\theta)$ is the predicted d_N/d_S from the chosen model (two-, or three-class) with parameters θ. For all model variants, we fixed the neutral-site fraction α_0 to the long-term d_N/d_S value for each species, estimated from unrelated genome pairs (SI Section 1).

We quantified the uncertainty in our estimates by bootstrap resampling the genome pairs, repeating the Poisson partitioning of the synonymous sites, and re-fitting the model to each bootstrap replicate. The 2.5%, 50%, and 97.5% quantiles of the bootstrap parameter distribution were used to define the corresponding confidence interval.

General constraints on the DFE from the asymptotic behavior of dN/dS. For an arbitrary DFE, the functional form in Eq. 1 still imposes strict constraints on the fraction of strongly deleterious mutations from the short-time behavior of d_N/d_S. A robust empirical observation across most

of the species in our dataset is that the clonal d_N/d_S curve remains close to one for $d_S^{(c)}$ as large as $d_S^{(c)} \approx 10^{-4}$. Intuitively, this implies that strongly deleterious sites cannot be too abundant: if many mutations were purged almost immediately, d_N/d_S would already have dropped well below one at these short times.

We can formalize this argument by partitioning the integral in Eq. 1 at a user-specified threshold s^*, and then bounding the integrand in each term by its value at the lower limit of integration:

$$\begin{aligned}\frac{d_N}{d_S} &= \int_0^{s^*} ds\, \rho(s) \cdot \frac{1-e^{-\frac{s\, d_S}{2\mu}}}{\frac{s\, d_S}{2\mu}} + \int_{s^*}^{\infty} ds\, \rho(s) \cdot \frac{1-e^{-\frac{s\, d_S}{2\mu}}}{\frac{s\, d_S}{2\mu}}\,, \\ &\leq 1 \cdot \int_0^{s^*} \rho(s)\, ds + \frac{1-e^{-\frac{s^*\, d_S}{2\mu}}}{\frac{s^*\, d_S}{2\mu}} \cdot \int_{s^*}^{\infty} \rho(s)\, ds\,. \end{aligned} \tag{S12}$$

Rearranging this expression yields an upper bound on the fraction of mutations that are stronger than s^*:

$$\int_{s^*}^{\infty} \rho(s)\, ds \leq (1 - d_N/d_S) \left[1 - \frac{1-e^{-\frac{s^*\, d_S}{2\mu}}}{\frac{s^*\, d_S}{2\mu}}\right]^{-1}. \tag{S13}$$

As a concrete example, Fig. 3 shows that $d_N/d_S \approx 0.9$ for $d_S \sim 10^{-4}$. Eq. S13 then implies that more than 90% of nonsynonymous mutations must have $s/\mu < 10^5$ and more than half must have $s/\mu < 10^4$, independent of the detailed form of the DFE.

An analogous argument for the large d_S limit of the dN/dS curve yields a lower bound on the fraction of mutations that are more costly than a given s^* value:

$$\int_{s^*}^{\infty} \rho(s)\, ds \geq 1 - d_N/d_S \left[\frac{1-e^{-\frac{s^*\, d_S}{2\mu}}}{\frac{s^*\, d_S}{2\mu}}\right]^{-1}. \tag{S14}$$

Applying this bound to the right-hand side of the d_N/d_S curve in Fig. 3 shows that more than 80% of all nonsynonymous mutations must have $s/\mu > 10^2$, and more than half must have $s/\mu > 10^3$. Together, these two conditions provide general constraint linking the early- and late-time behavior of the d_N/d_S curve to the underlying scale of the DFE.

5 Adaptive reversion model

We also compared the data in Figs. 3 and 4 to a generalized version of the adaptive reversion model in Ref. 30. This metapopulation model assumes that a subset of nonsynonymous sites undergo recurrent locally adaptive substitutions, which are later reversed when environmental conditions change.

Derivation of model. Following Ref. 30, we assume that these locally adaptive sites exhibit a form of modular epistasis. Each site is part of a larger module (or "locus" in Ref. 30) with two states

that we denote as “mutant” (M) and “wildtype” (WT). We assume that mutations at any of the sites in the module will switch it from the wildtype to the mutant state, but compensatory mutations (M→WT) can only occur through a reversion of the original mutation. We also assume that there is strong intra-module epistasis, such that two or more mutations in the same module have the same phenotypic effect as a single mutation.

Substitutions and reversions within a single person’s microbiome are modeled as stochastic transitions with total rates $r_{\rm on}$ and $r_{\rm off}$, respectively. These two effective parameters encapsulate both the rate of environmental switching and the time that it takes for the population to reach the new locally adapted state. In many cases of interest, this latter timescale may be negligible, and $r_{\rm on}$ and $r_{\rm off}$ can be interpreted as switching rate between environmental states where either the mutant or wildtype version of the module is preferred. Ref. 30 focused on a symmetric version of the model where $r_{\rm on} \approx r_{\rm off} \approx 1/\tau_{\rm flip}$; for completeness, we consider general values of $r_{\rm on}$ and $r_{\rm off}$ to allow for cases where the mutant state is preferred in only a small fraction of microbiomes ($r_{\rm on} \ll r_{\rm off}$).

Under these assumptions, the probability that a given module is in its mutant state at time t follows a simple Markov process,

$$P_{\rm M}(t) = \overline{P}_{\rm M} + (P_{\rm M}(0) - \overline{P}_{\rm M})e^{-(r_{\rm on}+r_{\rm off})t}, \tag{S15}$$

where $\overline{P}_{\rm M} \equiv \frac{r_{\rm on}}{r_{\rm on}+r_{\rm off}}$ is the steady-state probability of finding the module in the mutant state. However, since dN/dS is a site-based metric, we must also understand the probability that a given *site* within the module will be found in its mutant or wildtype state. Under similar assumptions as above, we can model the dynamics at a given focal site as a three-state Markov process, corresponding to cases where (i) neither the module or the focal site are mutated (WT,WT); (ii) both the focal site and module are mutated (M,M); and (iii) cases where the module is mutated at another site within the module, while the focal site retains its wildtype allele (WT,M). The probabilities of finding the site in each of the three states evolve as

$$\begin{aligned}
\partial_t P_{\rm WT,WT} &= r_{\rm off}(1 - P_{\rm WT,WT}) - r_{\rm on}P_{\rm WT,WT}\,, \\
\partial_t P_{\rm M,M} &= \epsilon r_{\rm on}P_{\rm WT,WT} - r_{\rm off}P_{\rm M,M}\,, \\
\partial_t P_{\rm WT,M} &= (1-\epsilon) r_{\rm on}P_{\rm WT,WT} - r_{\rm off}P_{\rm WT,M}\,,
\end{aligned} \tag{S16}$$

where ϵ is an additional parameter that describes the relative probability that a locally adapted module is driven by a mutation at the focal site. The solution to Eq. S16 is given by

$$\begin{aligned}
P_{\rm M,M}(t) = \overline{P}_{\rm M,M} &+ (P_{\rm M,M}(0) - \overline{P}_{\rm M,M})e^{-r_{\rm off}t}\left[(1-\epsilon) + \epsilon e^{-r_{\rm on}t}\right] \\
&- (P_{\rm WT,M}(0) - \overline{P}_{\rm WT,M})\epsilon e^{-r_{\rm off}t}\left(1 - e^{-r_{\rm on}t}\right)\,,
\end{aligned} \tag{S17}$$

where

$$\overline{P}_{\rm M,M} = \frac{\epsilon r_{\rm on}}{r_{\rm on}+r_{\rm off}}\,, \qquad \overline{P}_{\rm WT,M} = \frac{(1-\epsilon)r_{\rm on}}{r_{\rm on}+r_{\rm off}}\,, \qquad \overline{P}_{\rm WT,WT} = \frac{r_{\rm off}}{r_{\rm on}+r_{\rm off}}\,, \tag{S18}$$

are the steady-state probabilities of finding the site in a given state. This expression reduces to Eq. S15 when each site comprises its own module ($\epsilon = 1$), while allowing for more complex dynamics when $\epsilon < 1$.

To calculate the expected genetic divergence between two independently evolving microbiomes, we assume that fixation within each microbiome is fast relative to the divergence time and that the shared ancestral allele is drawn from the steady-state multinomial distribution in Eq. S18. The expected probability of observing a difference at a given site is given by

$$P_{\text{diff}}(t) = \langle 2P_{\text{M,M}}(t)(1 - P_{\text{M,M}}(t))\rangle_{\vec{P}(0)}\,, \tag{S19}$$

$$\begin{aligned} &= 2\overline{P}_{\text{M,M}}(1-\overline{P}_{\text{M,M}})\left[1 - e^{-2r_{\text{off}}t}\left(1-\epsilon+\epsilon e^{-r_{\text{on}}t}\right)^2\right] \\ &\quad - 2(1-\epsilon)\overline{P}_{\text{M,M}}\left(\epsilon - \overline{P}_{\text{M,M}}(1-\epsilon)\right)e^{-2r_{\text{off}}t}\left(1-e^{-r_{\text{on}}t}\right)^2 \\ &\quad - 4\overline{P}^2_{\text{M,M}}(1-\epsilon)e^{-2r_{\text{off}}t}\left(1-\epsilon+\epsilon e^{-r_{\text{on}}t}\right)\left(1-e^{-r_{\text{on}}t}\right)\,. \end{aligned} \tag{S20}$$

In our parameter regime of interest where each site makes a small contribution to its module ($\epsilon \ll 1$), this further reduces to

$$P_{\text{diff}}(t) = 2\overline{P}_{\text{M,M}}(1 - e^{-2r_{\text{off}}t})\,. \tag{S21}$$

Summing over all nonsynonymous sites within a given focal set $\mathcal{L}_{\text{focal}}$ yields a corresponding prediction for the d_N/d_S ratio,

$$\frac{d_N}{d_S} = \int dr_{\text{on}}\, dr_{\text{off}}\, d\epsilon\, \rho(r_{\text{on}}, r_{\text{off}}, \epsilon)\cdot 2\overline{P}_{\text{M,M}}\cdot\frac{r_{\text{off}}}{\mu}\cdot\frac{1-e^{-\frac{r_{\text{off}}d_S}{\mu}}}{\frac{r_{\text{off}}d_S}{\mu}}\,, \tag{S22}$$

where $\rho(r_{\text{on}}, r_{\text{off}}, \epsilon)$ is the aggregate distribution of r_{on}, r_{off}, and ϵ across all of the sites within the focal set. This generalizes the derivation in Ref. 30 to allow for variable on and off rates, and applies to arbitrary subsets of nonsynonymous sites, in addition to the total genome-wide value. Notably, the functional form of Eq. S22 is identical to the purifying selection model in Eq. 1, even though the two expressions emerge from distinct evolutionary mechanisms.

Three-class model. To understand the constraints that this equivalence places on the underlying parameters of Eq. S22, we first considered a simplified version of the reversion model, similar to the one proposed in Ref. 30, where the nonsynonymous sites are divided into three discrete categories: (i) a set of effectively neutral sites with $r_{\text{on}} \approx r_{\text{off}} \approx \mu$, (ii) a set of reversable adaptive loci with characteristic rates r_{on} and r_{off}, and (iii) a set of strong, permanently deleterious sites with $r_{\text{on}} \approx 0$ and $r_{\text{off}} \approx \infty$. Denoting the fraction of sites in each class by α_0, α_r, and $\alpha_\infty = 1 - \alpha_0 - \alpha_r$, respectively, the d_N/d_S curve in Eq. S22 reduces to

$$\begin{aligned} \frac{d_N}{d_S} &= \alpha_0 + 2\alpha_r\langle\overline{P}_{\text{M,M}}\rangle\cdot\left(\frac{r_{\text{off}}}{\mu}\right)\cdot\frac{1-e^{-r_{\text{off}}d_S/\mu}}{\frac{r_{\text{off}}d_S}{\mu}}\,, \\ &= \alpha_0 + 2\alpha_r\frac{\langle\epsilon\rangle r_{\text{on}}}{\mu}\cdot\left(\frac{r_{\text{off}}}{r_{\text{on}}+r_{\text{off}}}\right)\cdot\frac{1-e^{-r_{\text{off}}d_S/\mu}}{\frac{r_{\text{off}}d_S}{\mu}}\,, \end{aligned} \tag{S23}$$

where

$$\langle\epsilon\rangle = \int dr_{\text{on}}\, dr_{\text{off}}\, d\epsilon\, \rho(r_{\text{on}}, r_{\text{off}}, \epsilon)\cdot\epsilon \tag{S24}$$

is the average value of ϵ across all of the sites within class (ii).

To more directly compare this expression with the one provided in Ref. 30, it is necessary to invoke some additional assumptions about the functional form of ϵ. We will assume that the probability that a given site will be the driver mutation of its module is proportional to its relative target size — a term we use in a generalized sense to refer to both its underlying mutation rate and any differences in its within-host fixation probability. Using this notation, we can express the value of ϵ for a given site i as

$$\epsilon_i = \frac{\omega_i}{\omega^{\text{tot}}_{m(i)}}\,, \tag{S25}$$

where ω_i denotes the relative target size for mutations at site i, $m(i)$ denotes the module that site i belongs to, and ω^{tot}_m denotes the total target size for adaptive mutations in module m. This last term includes the contributions from (i) other nonsynonymous sites in the focal set, (ii) any nonsynonymous sites outside the focal set (if $\mathcal{L}_{\text{focal}}$ is a subset of the genome), and (iii) any non-coding or non-point mutations (e.g. invertible promoters; 63) that can contribute to the overall adaptation of the module but are not included in dN/dS.

Summing over all of the nonsynonymous sites in class (ii), the average value of $\langle\epsilon\rangle$ can then be expressed as

$$\langle\epsilon\rangle = \frac{1}{\alpha_r|\mathcal{L}_{\text{focal}}|}\sum_{i=1}^{\alpha_r|\mathcal{L}_{\text{focal}}|}\epsilon_i = \frac{n_{\text{modules}}}{\alpha_r|\mathcal{L}_{\text{focal}}|}\left\langle\frac{\omega^{\text{focal}}_m}{\omega^{\text{total}}_m}\right\rangle_m\,, \tag{S26}$$

where n_{modules} is the total number of modules, ω^{focal}_m denotes the total target size of module m within the focal set, and the average in the angle brackets is now taken over all modules m that have at least one site in the focal set. Substituting this result into Eq. S23 yields a related expression for dN/dS:

$$\frac{d_N}{d_S} = \alpha_0 + \alpha_r\cdot\frac{n_{\text{modules}}}{\alpha_r|\mathcal{L}_{\text{focal}}|}\left\langle\frac{\omega^{\text{focal}}_m}{\omega^{\text{total}}_m}\right\rangle_m\left(\frac{2r_{\text{on}}}{r_{\text{on}}+r_{\text{off}}}\right)\frac{1-e^{-r_{\text{off}}d_S/\mu}}{d_S}\,, \tag{S27}$$

which reduces to the expression in Ref. 30 in the special case where $r_{\text{on}} \approx r_{\text{off}} \approx 1/\tau_{\text{flip}}$ and $\langle\omega^{\text{focal}}_m/\omega^{\text{total}}_m\rangle_m \approx |\mathcal{L}_{\text{focal}}|/3L_{\text{syn}}$.

In contrast to the purifying selection model in Eq. 1, all of the time-dependence in Eqns. S23 and S27 is generated by reversions of adaptive sites. This places strong constraints on the underlying parameters, which we can start to understand by considering the large and small d_S limits of this expression.

Comparisons to data. Like the purifying selection model in Eq. 1, the long-time behavior of Eqns. S23 and S27 is governed by the the fraction of effectively neutral nonsynonymous sites α_0. In order to match the empirical observations of $d_N/d_S \approx 0.1$ between unrelated strains, we must continue to have $\alpha_0 \lesssim 0.1$. Similarly, the timescale of the decay of d_N/d_S suggests that the typical reversion rate must be on the order of $r_{\text{off}}/\mu \sim 10^4$, analogous to the effective selection coefficient $s/\mu \sim 10^4$ inferred under the purifying selection model in Fig. 3.

One important difference between the two pictures is that while the purifying selection model was mechanistically constrained to approach one at short times, the adaptive reversion model

possesses a different short-time limit,

$$\begin{aligned}\frac{d_N}{d_S} &\approx \alpha_0 + 2\alpha_r \langle \overline{P}_{\mathrm{M,M}} \rangle \cdot \frac{r_{\mathrm{off}}}{\mu} \\ &\approx \alpha_0 + \alpha_r \cdot \frac{n_{\mathrm{modules}}}{\alpha_r |\mathcal{L}_{\mathrm{focal}}|} \left\langle \frac{\omega_m^{\mathrm{focal}}}{\omega_m^{\mathrm{tot}}} \right\rangle_m \left(\frac{2r_{\mathrm{on}}}{r_{\mathrm{on}} + r_{\mathrm{off}}} \right) \frac{r_{\mathrm{off}}}{\mu} \,,\end{aligned} \tag{S28}$$

which functions as an additional fitting parameter. This extra degree of freedom allows the adaptive reversion model to produce dN/dS values that are substantially greater than one, which can be useful in cases where adaptive evolution is observed (8). However, for empirical curves like Figs. 3 and 4, where dN/dS remains comparatively close to one at short times, this additional degree of freedom must be precisely tuned in order to match this feature of the data. In particular, given the constraint that $\alpha_0 \lesssim 0.1$, matching the empirical values of $d_N/d_S \sim O(1)$ at short times requires that the reversible component of Eq. S28 must also be order one, yielding a corresponding condition

$$\alpha_r \cdot \underbrace{\frac{n_{\mathrm{modules}}}{\alpha_r |\mathcal{L}_{\mathrm{focal}}|} \left\langle \frac{\omega_m^{\mathrm{focal}}}{\omega_m^{\mathrm{tot}}} \right\rangle_m}_{\langle \epsilon \rangle} \cdot \frac{r_{\mathrm{on}}}{\mu} \sim \mathcal{O}(1) \,. \tag{S29}$$

The left-hand side of Eq. S29 is a product of several independent parameters (α_r, $\langle \epsilon \rangle$, r_{on}, and μ) which can differ from each other by several orders of magnitude in real biological systems. Achieving a dN/dS value of order one requires that these different terms must precisely balance one another. Since there is not an *a priori* reason why these different evolutionary parameters should be related to each other in this way, we argue that the adaptive reversion model requires an additional degree of fine-tuning to match this feature of the data.

Notably, this fine-tuning constraint must hold for any collection of sites within the genome that exhibit $d_N/d_S \sim \mathcal{O}(1)$ at short times. For example, Fig. 4 shows that Eq. S29 must separately hold within the subsets of missense and nonsense mutations, even though the underlying parameters could differ between them. Such agreement is not automatically guaranteed by the model and would require an additional degree of fine-tuning.

To illustrate this idea, we recall that the speed of the dN/dS decline in Fig. 4 was ∼4x faster for nonsense versus missense mutations, which we interpreted as a 4-fold higher fitness cost of nonsense mutations under the purifying selection model. In the adaptive reversion picture, this same difference in the dN/dS curves implies that the environmental switching rates must be 4-fold faster for nonsense vs missense mutations, which would require in turn that the nonsense mutations must arise from a systematically different set of modules than their missense counterparts. However, in order for the condition in Eq. S29 to remain satisfied in this new setting, the larger value of r_{off}/μ must be balanced by a 4-fold reduction in the remaining terms (e.g. the number of modules per focal site or the relative on-vs-off rates) to ensure that $d_N/d_S \sim \mathcal{O}(1)$ at short times. At present, the mechanisms that would allow for such fine-tuning remain unclear.

Even stronger tests of the adaptive reversion model are possible if one can identify the subset of reversible adaptive loci. In this case, the remaining, non-adaptive portion of the genome should provide a direct readout of the effects of purifying selection, which is assumed to be very strong

in the version of the model in Eqns. S23 and S27. While it is difficult to fully enumerate the set of adaptive loci, several previous studies have performed conceptually related analyses by leveraging recurrently mutated genes in collections of isolates sampled from individual hosts (8, 64).

In one study in *Bacteroides fragilis* (8) (one of the species included in Fig. 2), this subset of recurrently mutated genes exhibited a dN/dS ratio of 6.0, indicating strong positive selection (8). However, the remaining portion of the genome retained a dN/dS indistinguishable from one, similar to the genome-wide patterns in Fig. 3. A similar analysis in the pathogenic bacterium *Burkholderia dolosa* (64) observed a dN/dS ratio of 18 in the most recurrently mutated genes, and a residual dN/dS that was measurably less than one (dN/dS≈0.63), but still larger than the long-term value of $\alpha_0 \sim 0.1$ estimated from Ref. 65. In both species, these dN/dS values imply that either the vast majority of mutations in the non-parallel genes ($1 - \alpha_0/(d_N/d_S) \gtrsim 80\%$) must also be locally adaptive — and subject to the same fine-tuning constraints above — or that they must be eliminated on longer timescales through a purifying-selection-like process similar to Eq. 1. By contrast, no studies (to our knowledge) have identified a suitably large collection of sites with a short-term dN/dS $\lesssim$ 0.1, as predicted by the strong form of the adaptive reversion model in Eq. S27.

Together, these analysis show how comparing dN/dS curves for different classes of sites can provide additional information about the underlying selective forces that operate within a population. Extending this approach to other subsets of the genome would be an interesting avenue for future work.

Constraints from parallel evolution. An additional set of parameter constraints can be obtained by considering signatures of parallel evolution — i.e. the probability that the same adaptive mutations are independently sampled in unrelated lineages. In the model defined above, the degree of parallel evolution depends on the equilibrium frequencies $\overline{P}_{\mathrm{M,M}}$ in Eq. S18, and how they vary across the different sites in our focal set.

To make this intuition more concrete, we recall that the short-time limit of Eq. S28 implies that $1 - \alpha_0/(d_N/d_S) \approx 90\%$ of all nonsynonymous substitutions at short times must be locally adaptive mutations. This suggests that any signatures of parallel evolution observed among these recent clonal mutations will provide additional information about the realized values of $\overline{P}_{\mathrm{M,M}}$. We implemented this idea using clonal strain pairs in *P. vulgatus* in Fig. 5 and found that the vast majority ($> 90\%$) of nonsynonymous differences are private to a single strain pair (SI Section 6). For the adaptive reversion model to be consistent with this observation, the average steady-state probability $\langle \overline{P}_{\mathrm{M,M}} \rangle$ must be less than $\sim 1/n$, where n is the total size of the cohort. For our cohort size of $n \sim 100$, this implies that

$$\langle \overline{P}_{\mathrm{M,M}} \rangle \lesssim 10^{-2}. \tag{S30}$$

This constraint introduces a tension: if adaptive substitutions are rare because $\langle \overline{P}_{\mathrm{M,M}} \rangle$ is small, then the fraction of adaptive reversible sites (α_r) must be large to maintain the necessary nonsynonymous substitution rate. Combining this result with Eq. S28 and the estimated value of

$r_{\text{off}}/\mu \sim 10^4$ from Fig. 3, we find that α_r must satisfy the condition

$$\alpha_r \sim \left(\frac{r_{\text{off}}}{\mu}\right)^{-1} \cdot \langle \overline{P}_{\text{M,M}} \rangle^{-1} \gtrsim 10^{-2}\,. \tag{S31}$$

In other words, at least one percent of all nonsynonymous sites in the genome — corresponding to tens of thousands of sites — would need to participate in reversible adaptive dynamics to generate the observed d_N/d_S pattern in Fig. 3. If recurrent clonal SNVs continue to remain as rare in cohorts of increasing sizes, the upper bound on $\langle \overline{P}_{\text{M,M}} \rangle$ would become more stringent, requiring an even larger fraction of nonsynonymous sites to be reversible in order to explain the observed d_N/d_S trend.

Similar arguments hold for more general distributions of $\rho(r_{\text{on}}, r_{\text{off}}, \epsilon)$. Echoing the purifying selection model in Eq. 1, the timescale of the d_N/d_S curve requires that the vast majority of adaptive reversions have $\overline{P}_{\text{M,M}} \lesssim 10^{-2}$ and $10^2 \cdot \mu \lesssim r_{\text{off}} \lesssim 10^5 \cdot \mu$. These constraints suggest that the adaptive reversion model can account for the empirical d_N/d_S patterns only if a substantial fraction of nonsynonymous sites experience reversible selection on environmental timescales much longer than the human generation time.

Limitations and future directions. Our analysis above has focused on a particularly simple form of the adaptive reversion model (Eq. S27), similar to the one employed in Ref. 30. Extensions of this model that incorporate more elaborate forms of epistasis, compensatory mutations (30), or mixtures of adaptation and purifying selection (59) could potentially some of the constraints identified above. Efforts to explore this space of models — similar to Ref. 30 — remain an important avenue for future work.

We have also neglected the effects of recombination. While this was motivated by our ability to focus on the clonal regions of the genome where the recombined segments have been removed, recombination can still affect this clonal diversity in an indirect manner by removing portions of the genome (and any variants they might contain) from both the numerator and denominator of dN/dS. If recombination events affect synonymous and nonsynonymous variants in an unbiased manner, their effects will cancel out when computing the clonal dN/dS ratio in Figs. 3 and 4. However, recombination could also be an important mechanism for generating the reversion events considered above (30), which could in principle lead to systematic biases in the realized recombination rates at synonymous versus nonsynonymous sites. In the simple models considered above, these biases would effectively contribute to the overall reversion rate r_{off}. However, recombination can also be an important source of forward mutations, which would reduce the effective value of $\langle \epsilon \rangle$ in the presence of modular epistasis, and further limit the contribution of reversible adaptive mutations to the clonal dN/dS curves in Figs. 3 and 4. Accounting for this behavior with more explicit recombination-aware models is another important avenue for future work.

6 Analysis of clonal versus population-wide SNVs

To examine the distribution of clonal SNVs across the broader host population, we focused on the subset of strain pairs that were predicted to be fully clonal (i.e. those with no inferred recombination

events). Because multiple pairs of strains can all be closely related to each other, clonal SNVs can occur in more than one pair by common descent rather than by homoplasy. To avoid such redundancy, we filtered the initial strain set so that the remaining genomes were either clonal relatives or nearly fully recombined with respect to one another, defined as sharing less than 50% of their genome in identical blocks (SI Section 2). Within each group of strains connected by clonal relationships, we retained a single representative clonal pair. This procedure ensured that clonal variants were assessed only against otherwise unrelated genomic backgrounds.

In the case of *P. vulgatus*, our metagenomic dataset contains two major clades that are sometimes classified as separate species (*P. vulgatus* and *P. dorei*; 66). We restricted our analysis to the major *P. vulgatus* clade. After filtering, this procedure yielded a total of 120 strains, including 17 independent pairs of clonal strains. We catalogued all SNV differences at non-degenerate (1D) and fourfold degenerate (4D) coding sites between clonal pairs and computed the frequency with which each such variant appeared across the full set of 120 strains. For comparison, we also tabulated the prevalence of all SNVs present in at least one of the 120 strains, irrespective of whether they were originally identified in clonal pairs.

We repeated this procedure for each species with at least five clonal pairs, applying the same strain filtering and allele prevalence analysis. The resulting distributions of clonal versus population-wide SNVs, stratified by degeneracy class and allele prevalence, are shown in Fig. S8.

7 Analysis of *Staphylococcus aureus*

To assess whether our results extend beyond human gut commensals, we also analyzed a collection of 110 publicly available *Staphylococcus aureus* genomes obtained from Ref. 42 and originally sequenced by Ref. 67. These assemblies had already been aligned to the reference genome of strain MRSA252 in Ref. 42, and we annotated site degeneracy using the same reference–based in SI Section 1. When estimating d_N/d_S between genome pairs, we utilized all sites jointly covered between the two genomes, rather than restricting to a predefined core genome set.

Unlike most gut bacteria, the recent evolutionary history of *S. aureus* is dominated by clonal expansion, with relatively few recombination events (42). However, three large hybridization events have been documented previously, corresponding to the formation of sequence types (STs) 34, 239, and 582. These strains arose through the replacement of $\sim$200–600 kb regions in ST30, ST8, and ST25 genetic backgrounds, respectively (42–44). We used these previously identified tracts to stratify sites into "clonal" versus "recombined" categories and then estimated d_N/d_S separately within each class using the same Poisson–thinning procedures described above (SI Section 2).

Pairs of genomes that did not belong to these hybrid sequence types were included in the genome-wide d_N/d_S analysis, but were omitted from the stratified clonal versus recombined comparison. Despite the limited number of recombination events, the resulting dynamics again revealed distinct trajectories in d_N/d_S between clonal and recombined regions, consistent with the predictions of our recombination model.

Supplementary Figures

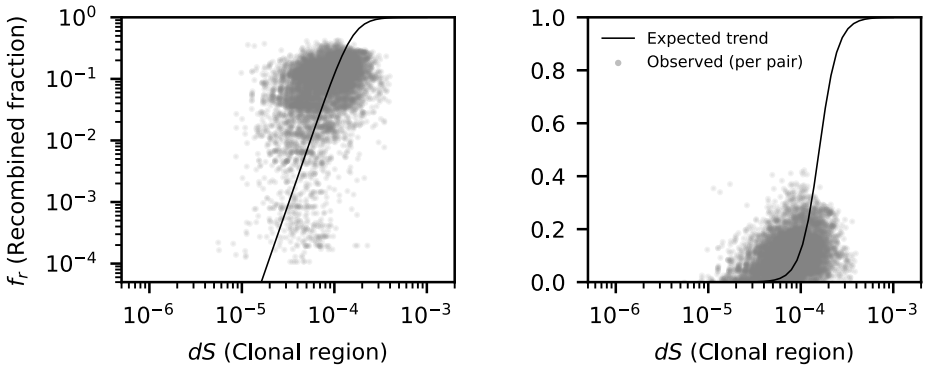

Figure S1: Empirical fit of recombination fraction versus clonal divergence. Each gray point shows the estimated recombination fraction (f_r) and the associated clonal divergence (dS_c) for each strain pair. The solid black line shows the logistic fit described in Eq. S5, with parameters $k = 10$ and $dS^* = 1.5 \times 10^{-4}$ chosen by visual inspection of the data. This logistic form captures the nonlinear increase of recombination fraction with divergence, providing an empirical model for $\bar{f}_r(dS_c)$ used in subsequent analyses. The two panels show the same data and fit, with the left axis on a logarithmic scale and the right axis on a linear scale.

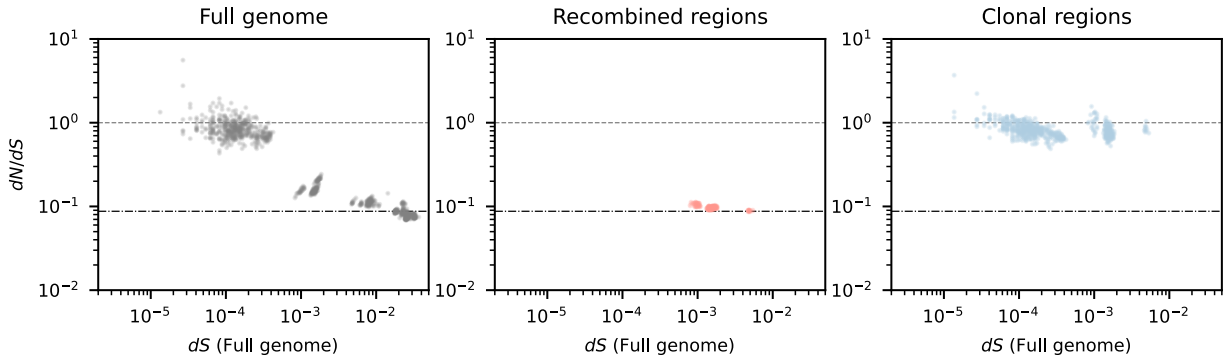

Figure S2: Analogous version of Fig. 2 for a collection of *Staphylococcus aureus* strains. Left: Genome-wide dN/dS dynamics across pairs of *S. aureus* strains, estimated from 1D and 4D sites. Middle: dN/dS values computed only within recombined regions corresponding to the large recombination events in ST34, ST239, and ST582 (SI Section 7). Right: dN/dS values computed within the remaining clonal regions.

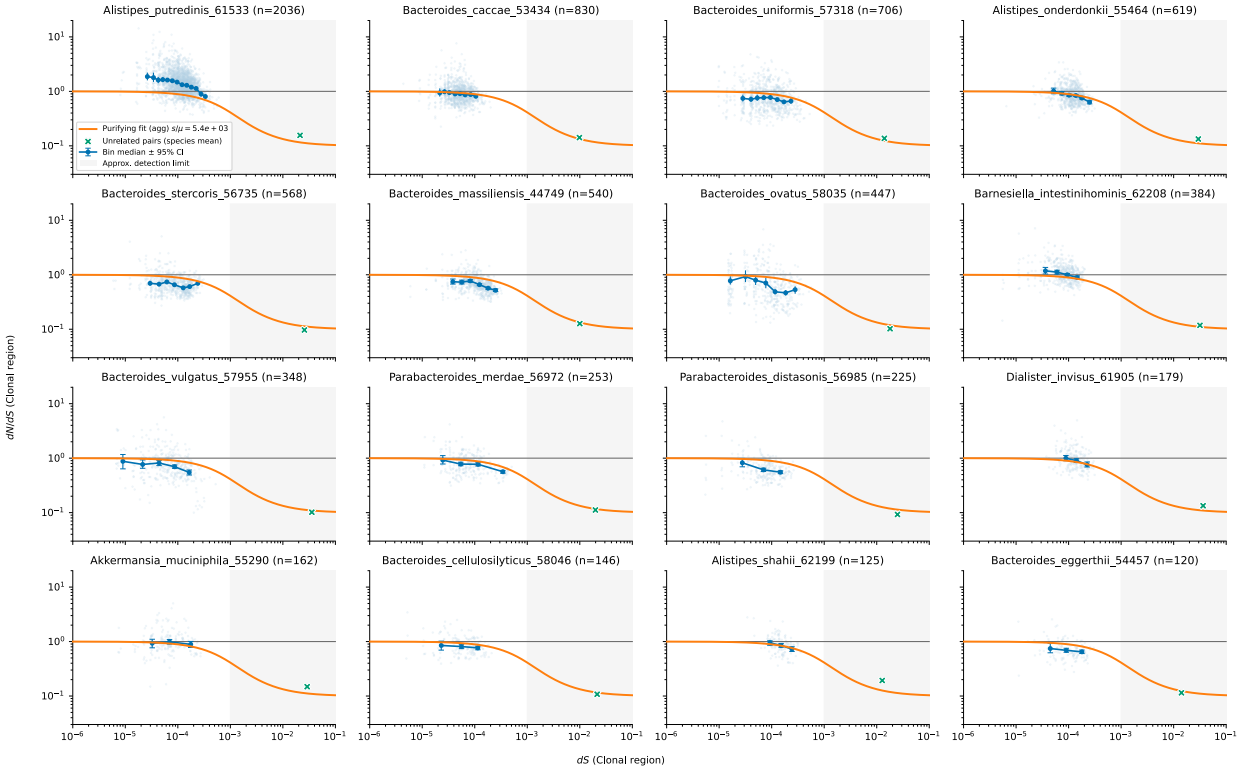

Figure S3: Species-level clonal d_N/d_S dynamics (all nonsynonymous sites). Light-blue points show pairwise clonal d_N/d_S versus clonal d_S estimated by Poisson-thinning; dark-blue markers with error bars denote the median of each divergence bin with 95% CIs from pooled counts. The orange curve is the same aggregate two-class purifying-selection fit used in Fig. 3, drawn identically in every panel for reference. Green "X" marks indicate the mean values of d_N/d_S and d_S between unrelated, fully recombined pairs for that species; the gray line marks $d_N/d_S = 1$; the shaded band denotes the approximate detection limit ($d_S \gtrsim 10^{-3}$). Species and bins with too few pairs are omitted; panels are ordered by the number of available pairs in each species.

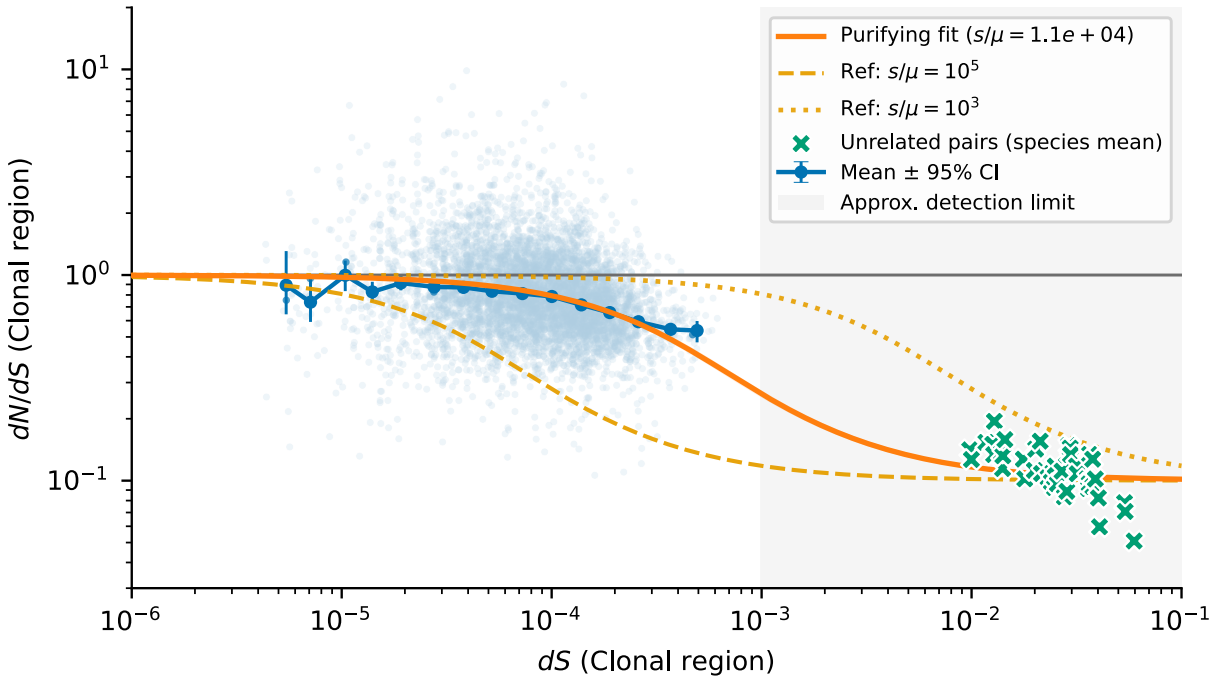

Figure S4: An analogous version of Fig. 3 excluding *Alistipes putredinis*. The overall results remain unchanged, with only a slightly higher estimated selection coefficient.

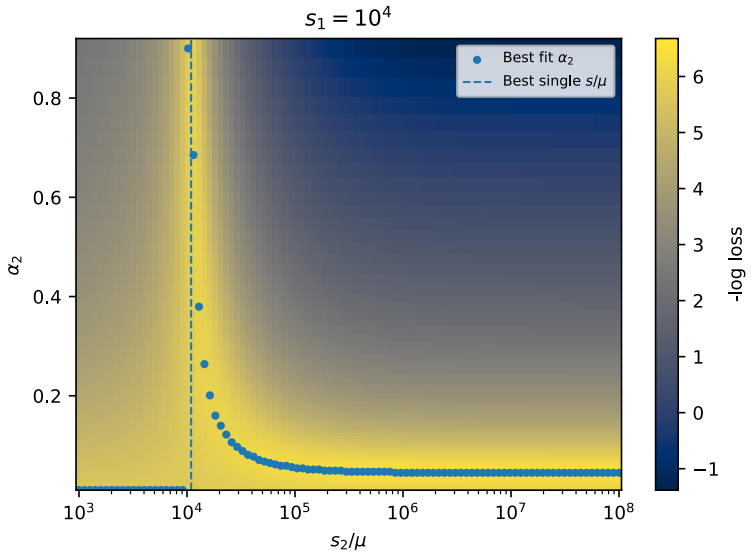

Figure S5: Loss surface of the three-class purifying selection model in SI Section 4. Each pixel shows the model fit quality ($-\log(\mathrm{wMSE})$, Eq. S11) as a function of the fraction of strongly deleterious sites (α_2) and their scaled selection coefficient (s_2/μ). Models were fit to all nonsynonymous sites, excluding *Alistipes putredinis*. The loss surface was evaluated by varying α_2 and s_2/μ, with $\alpha_0 = 0.1$ and $s_1/\mu = 10^4$ held fixed at the values obtained from the best-fit two-class model. The dashed vertical line marks the best-fit single-class value of s/μ, while dots indicate the best-fit α_2 value at each value of s_2/μ. The shape of the loss surface suggests that introducing an additional strongly deleterious mutation class ($s_2/\mu \gtrsim 10^5$) modestly improves the fit, but the inferred fraction of such sites remains very small ($\alpha_2 < 0.05$).

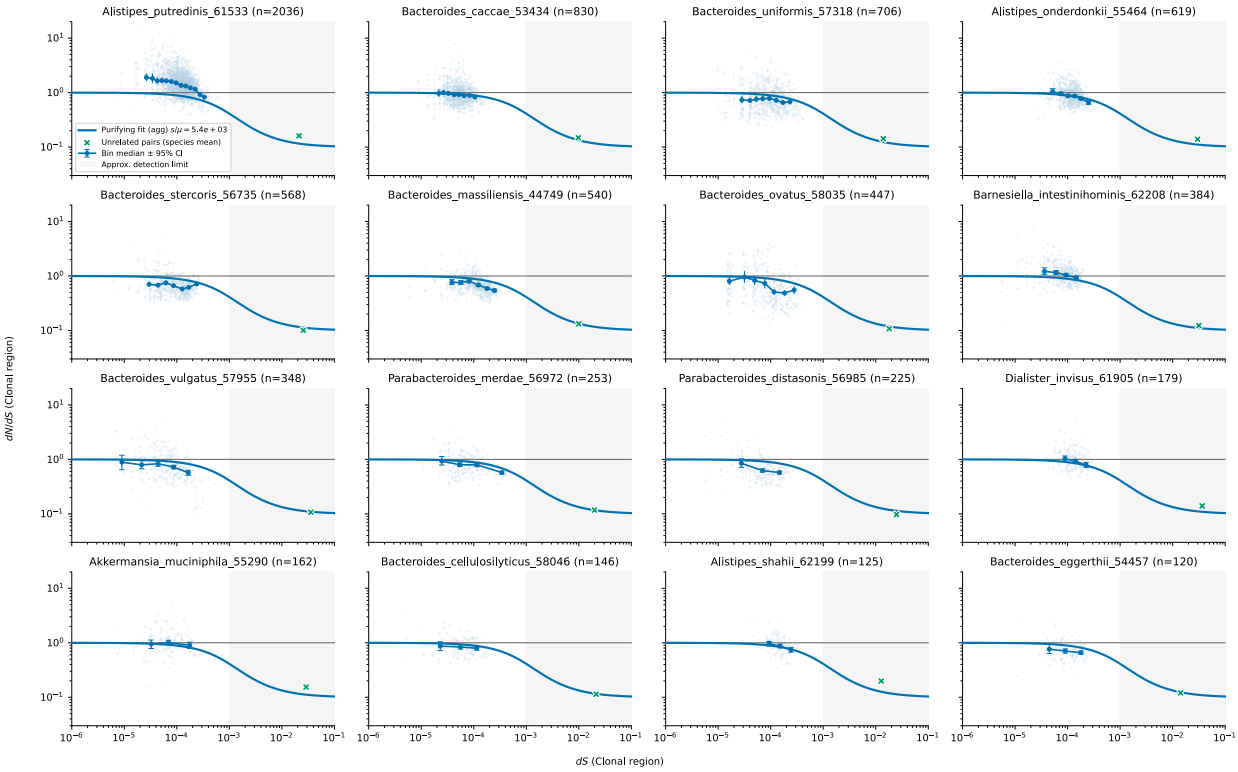

Figure S6: An analogous version of Fig. S3 computed for missense mutations only. The blue curve shows the aggregate two-class missense fit from Fig. 4 (left), held fixed across species.

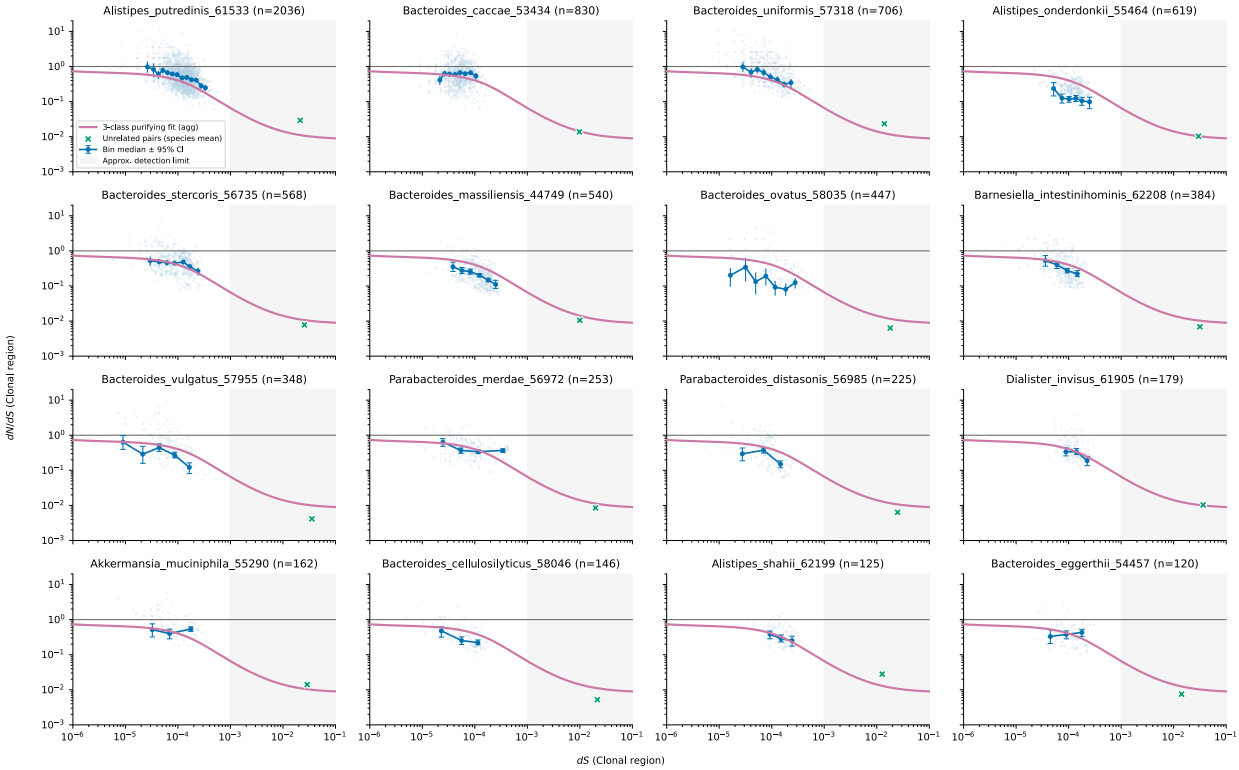

Figure S7: An analogous version of Fig. S3 computed for nonsense mutations only. The purple curve shows the aggregate three-class nonsense fit from Fig. 4 (right), held fixed across species.

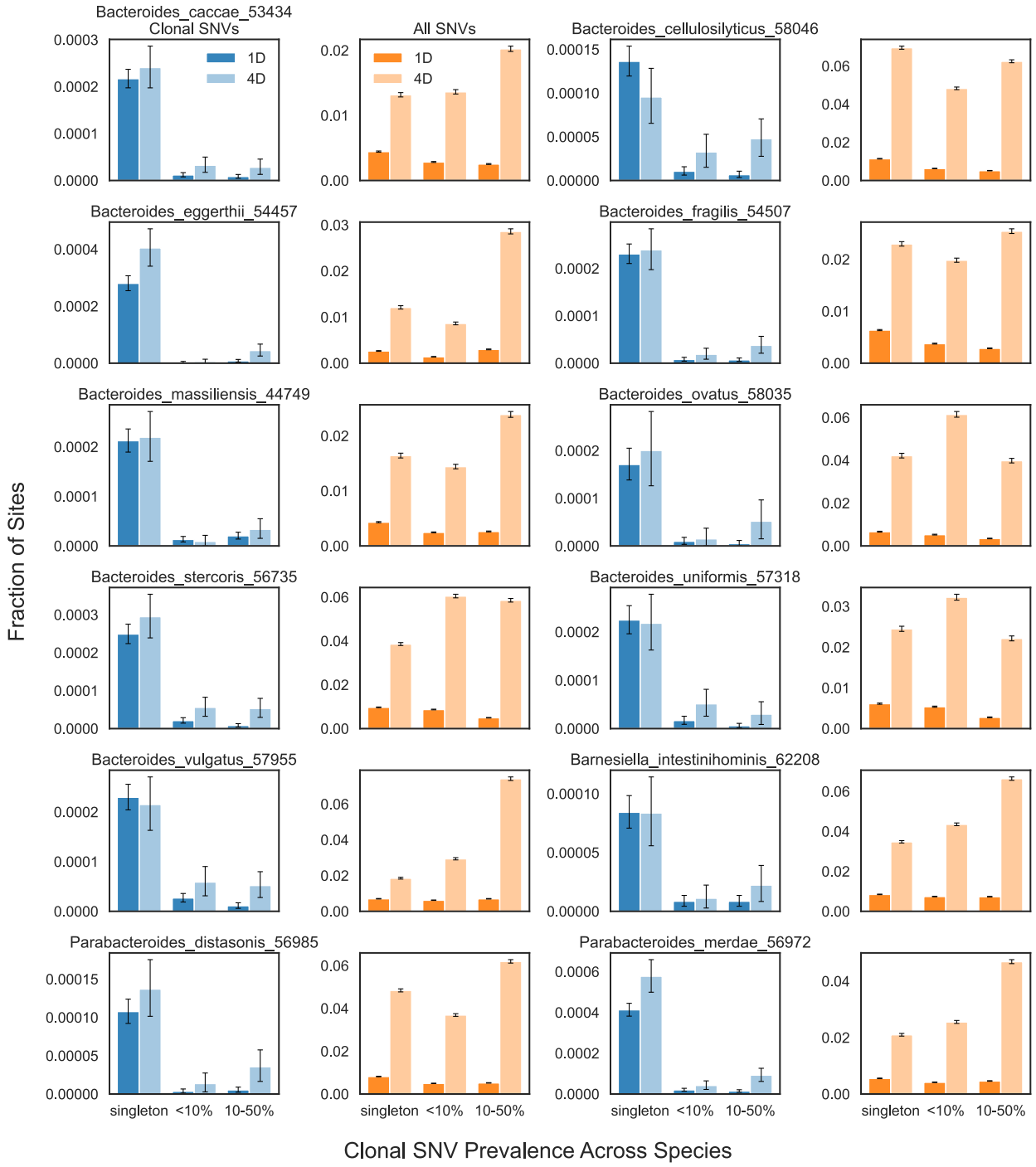

Figure S8: An analogous version of Fig. 5 for other commensal gut species. Left panels showing clonal SNVs (1D sites, dark blue; 4D sites, light blue) and right panels showing all SNVs for each species. "Singleton" denotes variants observed in only one host genome within each species. Error bars are 95% Poisson confidence intervals.