A macroecological perspective on genetic diversity in the human gut microbiome

William R. Shoemaker[1,2,*]

1 Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California, USA
2 Current Affiliation: The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, 34151, Italy

*** Contact:** `williamrshoemaker@gmail.com`

Abstract

While the human gut microbiome has been intensely studied, we have yet to obtain a sufficient understanding of the genetic diversity it harbors. This difficulty is compounded by the confluence of forces operating within the gut, where the dynamics of a given allele can be driven by evolutionary forces (drift, selection, etc.) or the ecology of the strain on which said allele is a constituent. To make progress towards this goal, I identified and characterized statistically invariant patterns of genetic diversity in a large cohort of unrelated hosts across 22 prevalent microbial species. The existence of said patterns allows for the set of potential models to be pared down to those that can reproduce observed patterns without the use of free parameters. Using a proposed distribution of allele frequencies, I evaluated the extent that a model of ecological dynamics could predict the fraction of hosts harboring a given allele (i.e., prevalence), a statistic that allows for the effect of sampling to be taken into account. I found that the Stochastic Logistic Model (SLM), an ecological model that successfully predict the dynamics of microbial strains, succeeded across species. The accuracy of the SLM was correlated with independent estimates of strain structure, validating the ecological implications of the SLM as a model of strain dynamics. This work suggests that a large number of detectable genetic variants in the human gut are subject to ecological, as opposed to evolutionary, dynamics due to the existence of strain structure.

Introduction

The human gut microbiome harbors astounding levels of genetic diversity. Hundreds to thousands of species continually reproduce in a typical host, accruing a total of $\sim 10^9$ *de novo* mutations each day [1]. Due to the comparatively brief generation time of microbes in the human gut [2], those mutations that are beneficial can rapidly fix on a timescale of days to months [1, 3–8]. Such evolutionary dynamics have the capacity to alter the genetic composition of a species within a given host. However, while all genetic diversity ultimately arises due to mutation, this actuality does not mean that all the genetic variants observed in the human gut are necessarily subject to evolutionary dynamics.

It is becoming increasingly apparent that for many species a large number of genetic variants do not fix or become extinct within a given host. Instead, these variants fluctuate at intermediate frequencies on timescales ranging from months to

years [1, 3, 9–12]. This pattern arises due to the co-occurrence of a few ($\sim \mathcal{O}(1-4)$) genetically diverged strains belonging to the same species, a process known as oligocolonization [3, 10]. This form of ecology is more than a descriptive detail, as it has been proposed that strains are the relevant scale at which interactions and dynamics occur in microbial systems [13, 14]. Thus, the dynamics of the genetic variants that comprise a given strain are subject to exogenous and endogenous ecological forces [15–17]. Such dynamics are a clear departure from those captured by standard origin-fixation models from population genetics, suggesting that measures of genetic diversity estimated within the human gut are shaped by ecology alongside evolution.

This confluence of ecological and evolutionary dynamics calls into question the feasibility of characterizing genetic diversity in the human gut. Many studies tackle such complexity by examining individual species [6, 18–20] or by searching for genetic differences between species [10, 11, 21–25]. While such approaches are useful for identifying individual species that are potential contributors towards specific conditions such as disease or the conference of the ability to metabolize certain resources, it is difficult to translate isolated observations into general patterns. This disproportionate focus on individual species and differences between species can lead to idiosyncratic notions about the typical dynamics of a genetic variant in the microbiome.

Instead, it is reasonable to identify genetic patterns that are statistically invariant across species. Such an approach provides the empirical motivation necessary to identify mathematical models that can explain said patterns and aid in the subsequent identification of feasible ecological or evolutionary dynamics [9, 26–28]. In recent years, substantial progress has been made towards characterizing the typical microbial evolutionary dynamics across species that operate within the human gut [3, 4, 7, 8, 29–31]. Using this approach it is possible to leverage the richness of species in the human gut microbiome, where each species can be viewed as a draw from an unknown distribution and, as an ensemble, be used to identify patterns that are statistically invariant [32, 33].

To identify such patterns, it is useful to examine prior attempts that successfully pared down the complexity of the gut. Notable recent examples come from the discipline of macroecology, which has succeeded at predicting patterns of microbial diversity and abundance at the species level across disparate environments, including the gut microbiome [34–38]. This approach emphasizes the benefits of identifying patterns of diversity that are statistically invariant, motivating the development of quantitative predictions derived from first principles. Recent work suggests that this approach may hold across scales of organization in the human gut, as species-level macroecological patterns have been successfully extended to strain-level ecology [38]. Such consistency suggests that the effect of widespread strain structure on genetic diversity may not be as convoluted as a surface level examination would suggest. Namely, models of the ecological dynamics of strains may aid in the identification and interpretation of universal patterns of genetic diversity in the human gut microbiome.

In this study, I sought to identify and explain patterns of genetic diversity that are statistically invariant across species in the human gut microbiome. The existence of said patterns provides evidence of universality across species, the impetus necessary to identify a plausible model of allele frequency dynamics. That model is the Stochastic Logistic Model (SLM) of growth, a model that describes the ecological dynamics of microbial communities at the level of species as well as strains [36, 38]. To evaluate the feasibility of the SLM while simultaneously accounting for the effects of sampling, I obtained predictions of the fraction of hosts harboring a given allele (i.e., prevalence) using zero free parameters. To confirm that the success of the SLM was due to the presence of strains, I evaluated the extent that alternative evolutionary models could explain the success of the SLM and determined that the existence of strain structure within a species was correlated with the accuracy of my predictions.

Results

Patterns of genetic diversity are statistically invariant across species

In order to determine whether it is possible to predict patterns of genetic diversity in the human gut, it is necessary to first investigate the degree of similarity in measures of genetic diversity across phylogenetically distant species. This manner of visualization, known as a data collapse, allows one to assess whether it is reasonable to assume that similar dynamics underlie different systems [39–41]. Such an analysis also provides the benefit of allowing for the identification of previously unknown empirical patterns for subsequent investigation. To determine whether there is evidence that the distributions of measures of genetic diversity have qualitative similar forms across species, I compiled allele frequency data for 22 bacterial species across human hosts using a quality control pipeline that explicitly accounted for the rate of sequencing error (Materials and Methods). The total number of processed hosts ranged from $108 - 371$ across species, with a median of 182 (Fig. 1a). The total number of sites ranged from $39 - 37,204$ across species, with a median of 10,269 synonymous and 5,204 nonsynonymous sites (Fig. 1b, S2b).

First, I examined the empirical distribution of within-host allele frequencies across all hosts. If the typical allele is present due to evolutionary forces, then the empirical distribution can be viewed as an ensemble of single-site frequency spectra [42, 43]. If the typical allele is present because it is on the background of a given strain, then the macroecological view of this object is that it captures the distribution of relative abundances of a given strain across hosts, also known as the Abundance Fluctuation Distribution [36]. Regardless of the underlying forces, the degree of similarity across species allows one to assess whether it is reasonable to predict that a single probability distribution is capable of explaining the empirical distributions of multiple species.

In order to determine whether different distributions share a single form it is useful to rescale them by key parameters [39]. Inspired by prior work in macroecology [36], I rescaled the distribution of a variable for a given species by its mean and variance using the standard score (i.e., z-score). This approach allows one to determine whether distributions are qualitatively different across species or whether they simply differ in their moments. The distribution of within-host allele frequencies has a similar qualitative form across species in the human gut for synonymous (Fig. 1c) and nonsynonymous sites (Fig. S2c), suggesting that a single distribution may be sufficient to characterize all species. Indeed, the distribution is fairly well-captured by a previously derived form of the gamma distribution [36], suggesting that it would likely be informative to examine models grounded in ecological and evolutionary first principles that predict a gamma distribution of allele frequencies.

Beyond the shape of the distribution of within-host allele frequencies, the statistical moments of within-host frequencies across hosts also exhibit qualitatively similar forms. The distribution of the mean frequency of a given allele across hosts ($\bar{f}$) tends to overlap across species for both synonymous (Fig. 1d) and nonsynonymous sites (Fig. S2d). Using this observation, one can examine the relationship between $\bar{f}$ and central statistical moments of a higher-order. The mean and variance of random variables frequently follow linear relationships on a logarithmic scale across biological systems, most notably in patterns of biodiversity in ecological communities [36, 44, 45] but also among population genetic patterns [46–48]. The existence of this relationship, known as Taylor's Law [49], would imply in the context of this study that the mean and variance of allele frequencies across hosts are not independent among species, reducing the number of parameters necessary to characterize the dynamics of the system [36]. Furthermore, if the variance scales quadratically with the mean then the existence of

the relationship implies that the coefficient of variation of f is constant across sites, an observation that can considerably reduce the difficulty of characterizing the dynamics of the system.

By examining the relationship between $\bar{f}$ and the variance of f (σ_f^2), it is clear that the two moments follow a linear relationship on a logarithmic scale for low values of $\bar{f}$ ($0 < \bar{f} \lesssim 0.35$). The exponent of this relationship is ~ 1.96 (bootstrapped 95% CI from 10,000 samples: [1.85, 2.06]) for synonymous sites, a value that is remarkably close to two, implying that the coefficient of variation in f can be viewed as a constant for the range of $\bar{f}$ where the relationship is linear (Fig. 1e). The exponent is slightly reduced for nonsynonymous sites (~ 1.83, 95% CI [1.72, 1.93]; Fig. S2e), suggesting that the variance increases with the mean at a slower rate relative to synonymous sites. Given that purifying selection is pervasive across species within the human gut [3, 11, 31, 50], it is likely that the typical allele at a nonsynonymous site confers a deleterious fitness effect, reducing its variance across hosts for low values of $\bar{f}$. However, the linear relationship does not extend to high values of $\bar{f}$. Given that f is, by definition, a bounded quantity (i.e., $0 \leq f \leq 1$), it is possible that the relationship between $\bar{f}$ and σ_f^2 for values of $\bar{f} \approx 1$ is governed by the upper bound on f. To determine whether this is the case, I plotted the maximum possible value of σ_f^2 for a given value of $\bar{f}$ constrained on the lower and upper bounds of f. This relationship, known as the Bhatia–Davis inequality, is defined as $\sigma_f^2 \leq (\max(f) - \bar{f})(\bar{f} - \min(f)) = (1 - \bar{f})\bar{f}$ [51]. The empirical relationship between $\bar{f}$ and σ_f^2 follows the inequality across species for $\bar{f} \gtrsim 0.35$, suggesting that the empirical relationship can be explained solely by the mathematical constraints on f, making the relationship uninformative for the purpose of identifying universal evolutionary or ecological patterns at a certain scale. It is worth noting that an exponent of two can emerge if the underlying distribution is sufficiently skewed [52, 53]. However, similar to ecological analyses of the relationship between the mean and variance of species abundances [36, 45, 54], the fact that the observed mean allele frequency varies by close to two orders of magnitude suggests that this patterns reflects a true scaling relationship.

Finally, I turned my attention to the fraction of hosts where a given allele is present. I found that the number of hosts in which a typical allele is present is small for synonymous (Fig. 1f) and nonsynonymous sites (Fig. S2f). Alternatively stated, the fraction of hosts harboring a given allele (i.e., prevalence) is typically low. These results, and their presence for synonymous and nonsynonymous sites, provides the empirical basis necessary to formulate quantitative predictions.

Predicting the prevalence of an allele across hosts

The existence of multiple patterns of allelic diversity that are universal across evolutionarily distant bacterial species suggests that comparable dynamics are ultimately responsible. The next task is to identify candidate models capable of explaining said patterns. The shape of the rescaled distribution of allele frequencies suggests that a gamma distribution is a suitable candidate, reducing the range of feasible models to those that are capable of predicting said distribution or a distribution of similar form. Different approaches can be used to identify such a model. However, given the consistency of the patterns, it is appropriate to focus on models that solely contain parameters that can be measured from empirical data rather than relying on estimates of free parameters via statistical inference. This approach allows one to make quantitative predictions that can serve as a first step towards the construction of a comprehensive theory [55].

We begin with the assumption that the frequency dynamics of a typical allele are primarily driven by the ecological dynamics of the strain on which said allele resides.

Appropriate Langevin equations (i.e., Stochastic Differential Equations) that capture ecological dynamics can be identified, though regardless of the underlying dynamics, the form of the available data constrains the ways in which the appropriateness of a given model can be evaluated. Given that temporal observational metagenomic data is at-present restricted to a small number of hosts for the human gut microbiome, I chose to focus on static samples taken at a single timepoint across a large number of unrelated hosts. This detail means that time-dependent solutions of the probability distribution of f cannot be empirically evaluated, so I instead focused on stationary solutions and limiting cases where time-dependence can be captured by parameters that can, in principle, be estimated.

To generate quantitative predictions, I examined a Langevin describing the dynamics of an allele that is on the background of a colonizing strain that is subject to self-limiting growth. This equation, known as the Stochastic Logistic Model of growth (SLM), has recently been shown to describe a range of macroecological patterns for microbial communities across disparate environments [36, 56, 57] as well as the temporal dynamics of strains within a single human host [38]. The SLM is defined as

$$\frac{\partial f}{\partial t} = \frac{f}{\tau_i}\left(1 - \frac{f}{K_i}\right) + \sqrt{\frac{\sigma_{\tau_i}}{\tau_i}} f \cdot \eta(t) \tag{1}$$

The terms $\frac{1}{\tau_i}$, K_i, and σ_{τ_i} represent the intrinsic growth rate of the strain on which the allele resides, the relative carrying capacity ($0 \leq K_i \leq 1$), and the coefficient of variation of growth rate fluctuations of the ith strain. The term $\eta(t)$ is a Brownian noise term where $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$ [58]. By obtaining the equivalent partial differential equation of the probability of f (i.e., the Fokker-Planck equation) and deriving the stationary solution (Materials and Methods), one finds that the SLM predicts that the frequency of a given allele on the background of a strain follows a gamma distribution

$$f_{\text{SLM}} \sim \text{Gamma}\left(\frac{2}{\sigma_{\tau_i}} - 1, \frac{2}{K_i \sigma_{\tau_i}}\right) \tag{2}$$

This distribution is fully characterized by the mean frequency and the squared inverse of the coefficient of variation ($\beta = \frac{\bar{f}^2}{\sigma_f^2}$) across hosts [36, 59]

$$f \sim \text{Gamma}\left(\beta, \frac{\beta}{\bar{f}}\right) \tag{3}$$

The similarity in the shape of the distribution of $\bar{f}$ across species and the relationship between $\bar{f}$ and σ_f^2 suggests that the mean and variance are appropriate quantities to evaluate the predictive capacity of each model (Fig. 1d,e, S2d,e). However, $\bar{f}$ and σ_f^2 can be interpreted as parameters of the SLM (i.e., empirical inputs), meaning that the SLM cannot be used to generate predictions of $\bar{f}$ and σ_f^2. To test the applicability of the SLM, I chose to examine the fraction of hosts harboring a given allele (i.e., prevalence), a quantity that has been used to examine microbial ecology and evolution across systems [36, 60], including the human gut microbiome [3, 7, 36]. Furthermore, this statistic explicitly incorporates the sampling effort of each site in each host (i.e., total depth of sequencing coverage), allowing for the SLM to be evaluated using experimental details. The predicted value of prevalence is defined as

$$\langle \varrho \rangle = 1 - \frac{1}{M} \sum_{m=1}^{M} \Pr[0 | D_m, \beta, \beta/\bar{f}] \tag{4}$$

where $\Pr[0 | D_m, \beta, \beta/\bar{f}]$ is the probability of observing zero reads assigned to a given allele within the mth host given a total depth of coverage D_m. It is important to

emphasize that $\Pr[0|D_m, \beta, \beta/\bar{f}]$ represents the convolution of the probability of observing a given allele frequency and the probability of obtaining zero copies of said allele given a sampling depth of D_m, meaning that the success of this distribution reflects the true underlying distribution of f. While Eq.4 is correct, the pipeline used for sequence data enacted a cutoff for the total depth of coverage, resulting in a coverage cutoff for a given minor allele ($A_{\text{cutoff}} = 10$; Fig. S1). This cutoff truncates the sampling distribution of minor allele read counts, meaning that read counts for a given allele less than the specified cutoff are effectively observed as zeros. This inferential detail can be explicitly accounted for by summing over the probabilities of observing alternative allele read counts up to and excluding the cutoff

$$\langle \varrho \rangle = 1 - \frac{1}{M} \sum_{m=1}^{M} \sum_{A_i=0}^{A_{\text{cutoff}}-1} \Pr[A_i|D_m, \beta, \beta/\bar{f}] \tag{5}$$

The choice of prevalence also allows one to evaluate the extent that the gamma distribution can recapitulate empirical relationships between genetic quantities. One such relationship is that the prevalence of a species (equivalently known as *occupancy* in macroecology) should increase with its mean abundance across communities [61], a pattern that has been found to exhibit statistically similar forms at the species level across microbial systems [62–64] and can be quantitatively explained through the existence of macroecological laws [36].

The Stochastic Logistic Model succeeds at predicting allelic prevalence

By examining the relationship between observed and predicted allele prevalence, I found that the SLM generally succeeded among synonymous and nonsynonymous sites (Figs. 2a, S3, S4). The fraction of all sites with relative errors ≤ 0.1 ($\leq 10\%$) ranged from $0.19 - 0.6$ across species, suggesting that a considerable fraction of genetic variants in the human gut are driven by ecological dynamics (Figs. 2b, S5, S6). Furthermore, the SLM generally succeeded in recapitulating the relationship between $\bar{f}$ and prevalence, a strain-level analogue of abundance-prevalence relationships in macroecology (Figs. 2b, S7, S8). While the SLM on its own cannot be used to predict Taylor's Law since the mean and variance were used as empirical inputs, the empirical evidence for the existence of Taylor's Law constrains the parameterization of the SLM and its subsequent interpretation [36, 65]. For the sites that the SLM is able to predict with a high degree of accuracy, the existence of Taylor's Law implies that β is constant across sites ($\sigma_{\tau_i} = \sigma_\tau$; Figs. 1e, S10, S11). Thus, the function for the expected allele frequency reduces to the proportionality $\langle f \rangle \propto K_i$.

These results suggest that it is worth investigating the accuracy of the SLM across observed estimates of prevalence. Given that a considerable fraction of sites have prevalence values close to one (e.g., the dot in the top-right corner of Fig. 2b), where the SLM has its highest degree of accuracy, it is necessary to remove these sites in order to examine the full distribution of prediction errors. By focusing on sites with observed prevalences < 0.9, we can see that observed and predicted prevalence values followed a one-to-one relationship across a wide range of observed prevalence values for many species (Fig. 3a). From these results, one can glean a few insights into the appropriateness of the SLM. First, for most observed prevalence values, when the predicted prevalence differs from the observed value it is generally below the one-to-one line. This pattern does not deviate as the observed prevalence increases, suggesting that the SLM is generally able to capture the distribution of f and its relation to prevalence for the range of observed prevalence values. When the SLM is inaccurate it generally underpredicts the true prevalence of an allele. Alternatively stated, the SLM can predict

an excess of zero observations ($f = 0$). However, there is a large uptick in the predicted prevalence for alleles with high observed values of prevalence, where the predictions are effectively on the one-to-one line. This is the case, as there is a large drop in the error as observed prevalence increases (Fig. 3b), suggesting a negative relationship between the observed prevalence of an allele and the relative error of the SLM. I found that the correlation between these two quantities was negative for all species (Fig. 3c. By permuting the order of these quantities I obtained null distributions of correlation coefficients for all species, where the majority of coefficients (16/22) clearly fall below the bounds of their null 95% confidence intervals.

The dependence of the predictive success of the SLM on the observed prevalence of an allele is a curious pattern. Ideally, one expects that the SLM is capable of explaining the dynamics of alleles if they primarily exist on the genetic background of strains, as the SLM has been found to explain the temporal dynamics of strain frequencies within a single host [38]. The fact that the SLM, when it is inaccurate, tends to underpredict true prevalence allows one to rule out models that predict an excess proportion of values of $f = 0$ than expected from a given distribution (e.g., zero-inflated gamma), as they would further reduce the predicted prevalence and increase the error of the prediction. Such models are often representative of competitive exclusion [36], an ecological outcome where the presence of a given species precludes the existence of another species. In the context of this study, competitive exclusion corresponds to the hypothesis that a strain is not found in a given host because it is unable to compete, resulting in alternative stable states. This connection allows one to rule out competitive exclusion as probably contributor towards the predictive error.

Alternatively, the comparatively poor performance of an ecological model among low prevalence alleles provides room for evolutionary explanations. Low prevalence alleles may be present due to the evolutionary dynamics, where a given allele can arise in a host due to mutation, increasing the observed prevalence to a value higher than that predicted by an ecological model of strain dynamics. Without a Langevin describing the dynamics of both the ecological and evolutionary dynamics of f it is difficult to parse alleles into discrete "ecological" and "evolutionary" categories. Regardless, the survival distributions of prevalence prediction errors have a strikingly bimodal form for several species (Fig. S5,S6), suggesting that there may be some truth to the claim that evolution, rather than ecology, can disproportionately contribute to the prevalence of alleles across hosts.

Strain-level ecology determines the accuracy of the gamma distribution

While the gamma distribution succeeds at explaining patterns of genetic diversity for an appreciable number of sites across microbial species, the ability to connect quantitative predictions to empirical distributions ultimately rests on the fact that the SLM predicts a distribution that is fully parameterized by the mean and variance of f (Eq. 3). Alternatively stated, one does not directly estimate the carrying capacity, they estimate the mean frequency, the expectation of which is a function of the carrying capacity under the SLM that reduces to a proportionality when Taylor's law holds [36]. This reliance on statistical moments estimated from observational data suggests that models of alternative dynamics that also predict a gamma distribution are equivalent candidates in the absence of additional evidence. Here I briefly examine two such evolutionary models. Starting with a Langevin equation, the frequency dynamics of a single allele are governed by forward and backward mutation (μ, υ), selection (s), and

random genetic drift for a population of size N.

$$\frac{\partial f}{\partial t} = sf(1-f) + \mu(1-f) + \upsilon f + \sqrt{\frac{f(1-f)}{N}} \cdot \eta(t) \tag{6}$$

To remove the number of free parameters and remove nonlinear terms, one can examine Eq. 6 in the low frequency limit ($f \ll 1$) and obtain Langevin equations for evolution under positive ($s > 0$) and purifying selection ($s < 0$)

$$\frac{\partial f}{\partial t} = -sf + \mu + \sqrt{\frac{f}{N}} \cdot \eta(t) \tag{7a}$$

$$\frac{\partial f}{\partial t} = sf + \mu + \sqrt{\frac{f}{N}} \cdot \eta(t) \tag{7b}$$

A gamma-distributed stationary solution can be derived in the case of purifying selection ($s < 0$; [66, 67]), whereas the stationary solution of Eq. 7ab is trivial. For the $s > 0$ case, a gamma distribution of allele frequencies can be derived where the time-dependence is captured by the maximum frequency that an allele can reach ($f_{max} \equiv \frac{e^{st}-1}{2Ns}$) which, in principle, can be estimated from empirical data (Materials and Methods). This dynamic form of mutation-selection balance is a gamma distribution that is solely parameterized by the maximum obtainable frequency of said allele (f_{max}) and the population scaled mutation rate ($2N\mu$) [68]. Together, these selection regimes provide two forms of mutation-selection balance that predict the gamma distribution

$$f_{\text{Evo},s<0} \sim \text{Gamma}(2N\mu, 2N|s|) \tag{8a}$$

$$f_{\text{Evo},s>0} \sim \text{Gamma}(2N\mu, f_{max}^{-1}) \tag{8b}$$

Both of these distributions can be parameterized using the mean and variance, where $\langle f \rangle = \frac{\mu}{|s|}$ and $\beta = \frac{1}{2N\mu}$ for $s < 0$ and $\langle f \rangle = 2N\mu f_{max}$ and $\beta = (2N\mu)^2$ for $s > 0$. These distributions provide alternative explanations for the predictive capacity of the gamma distribution, and it is worth investigating their feasibility.

There is evidence that purifying selection is widespread in the human gut microbiome [3, 31], though it is unlikely responsible for the allele frequency fluctuations I was able to observe. After accounting for sequencing error using `MAPGD`, the mean of the lowest inferred allele frequency across all species was ~ 0.02. Give that microbial populations are typically large in size, it is unlikely that the same allele independently reached frequencies $\gtrsim 0.02$ in multiple hosts. Assuming that independence among sites holds, the frequencies of individual alleles should not exceed $\frac{1}{N|s|}$ [69], meaning that it is unlikely that a substantial fraction of the alleles with non-zero inferred frequencies were driven by purifying selection. This explanation becomes even less likely when one considers the negative relationship between observed allelic prevalence and the accuracy of the gamma.

The dynamic mutation-selection balance parameterization of the gamma distribution also contains forward mutation, a useful feature given that mutation can contribute towards the total observed frequency of a beneficial allele that is increasing in frequency within a given host (Eq. 8ab). Such a feature is appealing, as, when the SLM fails, it does so by under-predicting observed prevalence. However, while the inclusion of forward mutation may increase the predicted value of prevalence, it is unlikely that positive selection substantially contributes towards the typical across-host single-site frequency spectra. Furthermore, the explicit time-dependence in the parameterization of f_{max} implies that a given allele is increasing in frequency in the $f \ll 1$ regime in all hosts where non-zero frequencies were observed. This dynamic is highly useful for

describing the dynamics of an ensemble of populations where the initial state is known, but it is likely inapplicable to single-timepoint samples across unrelated human hosts. To the extent that this parameterization holds, it likely does so for alleles that are observed at low frequencies, restricting the model's applicability to low prevalence alleles. Because the SLM tends to fail for this prevalence regime, it is possible that a distribution derived from a model of evolutionary dynamics, gamma or otherwise, is necessary to predict genetic diversity among low prevalence alleles.

Beyond the consideration of evolutionary models, it is ultimately necessary to evaluate the extent that the success of the gamma is due to the existence of strain structure. This is a difficult task given that there is no guarantee that a given allele is present in multiple hosts because it is on the background of genetically identical strains. This limitation arises due to the difficulty inherent in determining whether a given host harbors a given strain from a single static metagenomic sample. Such difficulty persists in part due to technical and statistical limitations, but also due to the lack of practical strain definitions [70–73]. It is difficult to phase strains from short-read sequencing data where physical linkage between variants cannot be established, making it necessary to instead identify a single haplotype within a given host [3, 10]. So while it is currently possible to identify the prevalence of dominant lineages across hosts, there is no straightforward approach that allows one to assign an observed allele to a given strain.

Given these methodological constraints, I identified the existence of strain-level structure for a given species within a given host using `StrainFinder`, a program that infers the number of strains using the shape of the site-frequency spectrum (Materials and Methods; [74]). I found that the fraction of hosts harboring strain structure ranged from $0.089 - 0.64$ among the species I examined, with a median of ~ 0.27 (Fig. 4a). If alleles with higher prevalence were driven by the presence of strain structure, then one would expect a positive correlation between the accuracy of the predicted allele prevalence and the fraction of hosts containing strain structure across species, equivalent to observing a negative correlation between relative error and the fraction of hosts containing strain structure. This prediction held, as the correlation between these two variables was typically negative within a given range of prevalence values, but tended to become more negative when only alleles with high prevalence estimates were included (Fig. 4b). To resolve this relationship, I examined the degree of correlation between the relative error of the gamma and the fraction of hosts with strain structure across a wide range of observed prevalence thresholds (Fig. 4c). The correlation tended to increase with prevalence, though there was a substantial decrease once a prevalence threshold of ~ 0.1 was reached. The fraction of hosts with strain structure for a given species provide context as they place lower bounds on the range of allelic prevalences that can be driven by strain-level ecology, meaning that alleles with prevalence values lower than the lowest observed fraction of hosts with strain structure are unlikely to be driven by ecology. Given that the species with the lowest observed fraction of hosts with strain structure were close to this descent, it is possible that this value is where the ecological dynamics of strains began to dominate the trajectories of observed alleles.

Discussion

Identifying and explaining universal patterns is a difficult task in biology. Once they are identified, the extent that mathematical models grounded in first principles can be used to generate successful quantitative predictions is a reflection of one's understanding of their system. In this study, I identified patterns of genetic diversity in the human gut microbiome that were statistically invariant across evolutionarily distant species. Motivated by these results, and the prominence of strain structure in the human gut, I identified a prospective model of ecological dynamics (i.e., the Stochastic Logistic

Model [36]) that could explain said patterns. Using this model, I was able to predict the fraction of hosts that harbored a given allele (i.e., prevalence) through the sole use of measurable parameters (i.e., zero statistical fitting) for a considerable fraction of sites across species. The accuracy of this prediction was correlated with independent estimates of strain structure, providing additional empirical evidence that patterns of genetic diversity across human hosts are driven by strain-structure for a considerable fraction of sites.

The success of the SLM for common alleles suggests that true strain absences within a host are rare enough that they do not exert a detectable influence on allelic prevalence. A lack of true absences subsequently implies that competitive exclusion is rare at the strain level [36]. This conclusion seemingly contrasts with the observation that strains exhibit varying frequencies across hosts [3], though fluctuations alone are not demonstrative of competitive exclusion. Like any model with demographic stochasticity, fluctuations are expected under the SLM and some number of absences will inevitably arise due to the finite nature of sampling. However, it is possible that the carrying capacity of a given strain could vary from host-to-host, widening the across-host distribution of frequencies (Fig. 1c). This detail can readily be incorporated into the SLM if one assumes that the carrying capacity in a given host is an independently drawn random variable from some unknown distribution [57]. Thus, given certain assumptions, the success of the SLM is reconcilable with the view that the carrying capacity of a strain is host-dependent.

While the SLM successfully predicted the prevalence of common alleles across hosts and has been shown to describe the temporal dynamics of strains within a host [38], all models have limitations, and it is useful to briefly discuss those that are applicable. A fundamental limitation of the SLM is that it is phenomenological in nature. It is unclear how microscopic details that are relevant to the ecological dynamics of strains, namely, consumer-resource dynamics [13, 32, 75, 76], map onto the SLM. Models that incorporate such dynamics are capable of recapitulating the temporal dynamics of microbial communities at the species level [77], and are necessary to model the emergence of new strains and their subsequent eco-evolutionary dynamics [76], suggesting that these microscopic details may be necessary to describe certain macroecological patterns at the strain level. As a contrast, the phenomenological nature of the SLM could be viewed as an asset when one wants to capture and predict multiple empirical patterns using an analytic solution without the use of fitted parameters. Indeed, the SLM succinctly captures the dynamics of a constrained random walk [77] and it is likely that alternative models of strain-level ecology that capture the same stochastic process are equally applicable.

Regardless of the specific model, the conclusion that a substantial fraction of alleles are subject to ecological, as opposed to evolutionary, dynamics is of consequence to studies of genetic diversity in the human gut. Measures of genetic diversity within a single host (e.g., nucleotide diversity) are often used to assess the genetic content of a microbial species [11, 78–81]. Recent efforts to characterize patterns of genetic diversity within a single host have found that the temporal dynamics of nucleotide diversity are primarily driven by fluctuations in strain frequencies [38]. In addition, the contribution of strain structure to estimates of genetic diversity from unrelated human hosts has been previously reported [3]. This study builds on past results by specifying that strain structure shapes patterns of genetic diversity across hosts. The implications of strain-level ecology likely extends to measures of genetic differentiation between populations (e.g., fixation index, F_{ST}) that have been used to assess the degree of structure of a given species across human hosts [11]. The results presented in this study suggest that measures of genetic diversity that do not account for strain ecology are unlikely to be informative of evolutionary processes operating within the human gut.

Finally, it is worth noting that the success of any prediction is contingent on one's range of observation. After accounting for sequencing error, the lowest inferred allele frequencies ranged from $0.006 - 0.06$ across species with an average of ~ 0.02. A straightforward calculation suggests that this range is higher than the true minimum frequency (i.e., $1/N$) by several orders of magnitude. The mean relative abundances of a given species across hosts ranged from $0.02 - 0.1$. Previously established order-of-magnitude estimates of the typical number of cells in the human gut range from $\sim 10^{13} - 10^{14}$, from which one can use the mean relative abundance distribution across species to calculate a first-pass range of empirical abundances of $\sim 10^{11} - 10^{13}$ [1, 82]. This range suggests that the true minimum frequency of an allele is at least eight orders of magnitude lower than the minimum inferred allele frequency. It is clear from this calculation that, at present, researchers are only able to examine a narrow range of possible allele frequencies within the human gut microbiome.

The implications of such a relatively narrow observational window are that the success of predictions derived from ecological principles, and the feasibility of alternative single-locus models of evolution, are likely contingent on present limitations on the depth and error rates of shotgun metagenomic sequencing. As advances in sequencing technology and statistical inference continue to permit lower observational thresholds and provide information about the physical linkage between variants [5, 83–86], one expects that an increasingly higher fraction of observed alleles will be subject to evolutionary forces rather than the ecological forces affecting the strain harboring said allele, reducing the aptness of ecological models. Succinctly stated, the existence of strain structure suggests that the dynamics of allele frequencies are likely dependent on the frequency scale at which observations can be made. By expanding said range, it may be possible to identify a frequency threshold where observable alleles are primarily driven by evolutionary dynamics rather than the ecological dynamics of strains, providing the means to test quantitative predictions of recent developments in population genetic theory [68, 69]. Recognition of the possibility of such scale-dependence has the potential to shape future studies and rigorously assess the purported universality of empirical patterns of genetic diversity in the human gut microbiome.

Materials and Methods

Data acquisition and processing

To investigate patterns of genetic diversity within the human gut microbiome, I used shotgun metagenomic data from 468 healthy North American individuals sequenced by the Human Microbiome Project [17, 87]. I first processed the data using a previously developed analysis pipeline to identify the set of sites in core genes [3]. This pipeline uses a standard reference-based approach (`MIDAS` v1.2.2 [21]) to map reads from each metagenomic sample to reference genes across a panel of prevalent species and filter reads based on quality scores and read mapping criteria. The relative abundance of each species in a given host was inferred using `merge_midas.py species`. Then, the command `merge_midas.py genes` was run with the following flags: `--sample_depth 10`, `--min_samples 1`, and `--max_species 150`. Using this processed gene output, the command `merge_midas.py snps` was run with the following flags: `--sample_depth 5`, `--site_depth 3`, `--min_samples 1`, `--max_species 150`, and `--site_prev 0.0`. I only processed samples from unrelated hosts and removed all temporal samples, leaving us with a set of samples that correspond to observations from unique hosts. I retained alleles in sites that were present in a core gene (i.e., a gene that was present in $\geq 90\%$ of hosts) with a minimum total within-host depth of coverage of 20 ($D \geq 20$) in at least 20

hosts.

After identifying the set of sites in core genes that passed the quality control thresholds, I obtained appropriate BAM files so that allele frequencies could ultimately be estimated. First, using the reference genomes used by `MIDAS` and the BAM file generated by `MIDAS` for all species in a single host, I split the BAM file containing all species into separate BAM files for each species using the command `samtools view` with default settings [88]. Each BAM file was sorted using the `samtools sort` command with default settings, from which a `.header` file was using `samtools view`.

Sequencing errors can obfuscate naïve estimates of low frequency alleles ($f \ll 1$). To account for such errors, I elected to use the full maximum likelihood estimator `MAPGD` v0.4.40 [89]. This estimator was chosen due its comparative high sensitivity to low frequency alleles without sacrificing its false discovery rate [90]. Using BAM and `.header` files processed from `MIDAS` output, `MAPGD` was run with a log-likelihood ratio polymorphism cutoff of 20 (`-a 20`). Samples with insufficient coverage for `MAPGD` to be run were removed from all downstream analyses. I then polarized alleles based on the major allele across hosts.

As an independent estimate of strain structure, I estimated strain frequencies for all species using all sites with ≥ 20 fold coverage within each host by applying `StrainFinder` v1.0 [74] to the frequency spectra obtained from the upstream pipeline [3]. The program `StrainFinder` was run on each sample for each species using 10 initial conditions using local convergence criteria with the following flags: `--dtol 1`, `--ntol 2`, `--max_time 20000`, `--converge`. The program was run for strain numbers ranging from one to four the estimates with the top five log-likelihoods were retained. I then selected strain frequencies with the lowest Bayesian Information Criterion for each species in each sample. Joint density plots (e.g., Fig. 2) were made using functions from `macroecotools` v0.4.0 [91].

Predicting allele prevalence using the SLM

Deriving the distribution of allele frequencies from the SLM

We begin with the assumption that the typical polymorphism observed within a given host for a given species is present because it is on the background of a colonizing strain. In such a scenario, the dynamics of an allele are not determined by its evolutionary attributes (i.e., fitness effect, mutation rate, etc.) but by the ecological dynamics of the strain. There is increasing evidence that the stochastic logistic model of growth (SLM) is a suitable null model of microbial ecological dynamics at the species level [36, 56, 57] and recent evidence indicates that the SLM sufficiently fits the temporal dynamics of strains within a human host for the vast majority of microbial species [38]. A non-trivial application of the SLM to strain-level ecology requires there to be more than one strain within a given host, giving the allele a range of frequencies of $0 < f < 1$

$$\frac{\partial f}{\partial t} = \frac{f}{\tau_i}\left(1 - \frac{f}{K_i}\right) + \sqrt{\frac{\sigma_{\tau_i}}{\tau_i}} f \cdot \eta(t) \tag{9}$$

The terms $\frac{1}{\tau_i}$, K_i, and σ_{τ_i} represent the intrinsic growth rate of the strain, carrying capacity, and the coefficient of variation of growth rate fluctuations. The term $\eta(t)$ is a Brownian noise term where $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$ [58]. By definition, strain frequencies within a species must be between zero and one, so $0 < K_i < 1$.

Using the Itô $\leftrightarrow$ Fokker-Planck equivalence [58], one can formulate a partial differential equation for the probability $p(f, t)$ that an allele has frequency f at time t

$$\frac{\partial p(f,t)}{\partial t} = -\frac{\partial}{\partial f}\left[\left(\frac{f}{\tau_i}\left(1 - \frac{f}{K_i}\right)\right) p(f,t)\right] + \frac{\sigma_{\tau_i}}{2\tau_i}\frac{\partial^2}{\partial f^2}\left(f^2 p(f,t)\right) \tag{10}$$

From which one sets $\frac{\partial p}{\partial t} = 0$ to obtain the stationary probability distribution of allele frequencies

$$p(f) = \frac{1}{\Gamma\left(2\sigma_{\tau_i}^{-1} - 1\right)} \left(\frac{2}{K_i \sigma_{\tau_i}}\right)^{2\sigma_{\tau_i}^{-1}-1} \exp\left[-\frac{2}{K_i \sigma_{\tau_i}} f\right] f^{2\sigma_{\tau_i}^{-1}-2} \tag{11}$$

This distribution, known as the abundance fluctuation distribution in macroecology [36], is a gamma distribution with the following mean and squared coefficient of variation

$$\langle f \rangle = K_i \left(1 - \frac{\sigma_{\tau_i}}{2}\right) \tag{12a}$$

$$\frac{\langle f^2 \rangle - \langle f \rangle^2}{\langle f \rangle^2} = \frac{\sigma_{\tau_i}}{2 - \sigma_{\tau_i}} \tag{12b}$$

Defining empirical estimates of $\langle f \rangle$ and $\frac{\langle f^2 \rangle - \langle f \rangle^2}{\langle f \rangle^2}$ as $\bar{f}$ and β^{-1}, I obtained a form of the gamma (represented in its shape/rate parameterization form) that can be used to generate ecological predictions of measures of genetic diversity with zero free parameters

$$p\left(f|\beta, \beta/\bar{f}\right) = \frac{1}{\Gamma(\beta)} \left(\frac{\beta}{\bar{f}}\right)^{\beta} \exp\left[-f\frac{\beta}{\bar{f}}\right] f^{\beta-1} \tag{13}$$

Deriving prevalence predictions using the SLM

The probability of detecting an allele of a given frequency within a host depends on one's sampling effort. The impact of finite sampling from gamma distributed random variables has been previously examined within macroecology [36], results that I apply and extend in this section. To model this process, I start by letting (A, D) denote the number of reads of the alternate allele and total sequencing depth at a given site. I estimate the frequency of the alternate allele within a given host as $\hat{f} = A/D$. I assume that sampling distribution of A is binomial

$$\Pr[A|D, f] = \binom{D}{A} f^A (1-f)^{D-A} \tag{14}$$

When $D \gg 1$ and $f \ll 1$ while $D \cdot f$ remains finite, the binomial sampling process can be approximated by the Poisson distribution

$$\Pr[A|D, f] = \frac{(D \cdot f)^A}{A!} e^{-D \cdot f} \tag{15}$$

Using this approximation, one can solve the integral for the probability of observing A reads assigned to the alternate allele out of D total reads when f is a gamma distributed random variable

$$\begin{aligned}
\Pr[A|D, \beta, \beta/\bar{f}] &= \int_0^1 \Pr[A|D, f] \cdot p(f|\beta, \beta/\bar{f})\, df && \text{(16a)} \\
&= \frac{D^A}{A!\Gamma(\beta)\frac{\beta}{\bar{f}}^{\beta}} \int_0^1 f^{A+\beta-1} e^{-f(D+\frac{\beta}{\bar{f}}^{-1})}\, df && \text{(16b)} \\
&= \frac{\Gamma(\beta + A)}{A!\Gamma(\beta)} \left(\frac{\bar{f}D}{\beta + \bar{f}D}\right)^A \left(\frac{\beta}{\beta + \bar{f}D}\right)^{\beta} && \text{(16c)}
\end{aligned}$$

By setting $A = 0$, one can calculate the probability of not detecting the alternate allele (i.e., absence) with a sampling depth of D reads [36, 60]

$$\Pr[0|D, \beta, \beta/\bar{f}] = \left(1 + D \cdot \frac{\beta}{\bar{f}}\right)^{-\beta} \tag{17}$$

From which one can calculate the expected prevalence of the allele $\langle \varrho \rangle$ over M hosts as

$$\langle \varrho \rangle = \frac{1}{M} \sum_{m=1}^{M} (1 - \Pr[0|D_m, \beta, \beta/\bar{f}]) \tag{18}$$

The distribution of allele frequencies under a linearized single-locus model of evolution

Deriving the distribution of allele frequencies in the linear purifying selection regime

The derivation of the gamma distribution for a single-locus model of purifying selection in the low frequency limit has previously been described [66]. We begin with a Langevin equation describing evolution on a single-locus with forward and backward mutation (μ, υ) and selection (s) in a population of N individuals

$$\frac{\partial f}{\partial t} = -sf(1-f) + \mu(1-f) + \upsilon f + \sqrt{\frac{f(1-f)}{N}}\eta(t) \tag{19}$$

where $\eta(t)$ is a Gaussian noise term. Assuming that $f \ll 1$, the nonlinear selection and back mutation terms drop out and our equation reduces to

$$\frac{\partial f}{\partial t} = -sf + \mu + \sqrt{\frac{f}{N}}\eta(t) \tag{20}$$

Where we again use the Itô $\leftrightarrow$ Fokker-Planck equivalence [58] to formulate a PDE for the probability $p(f, t)$ that an allele has frequency f at time t

$$\frac{\partial p(f,t)}{\partial t} = -\frac{\partial}{\partial f}\left[(-sf + \mu)\, p(f,t)\right] + \frac{1}{2N}\frac{\partial^2}{\partial f^2}\left(f p(f,t)\right) \tag{21}$$

Solving for the stationary distribution, we obtain a gamma distribution of allele frequencies

$$p(f) = \frac{1}{\Gamma(2N\mu)} (2N|s|)^{2N\mu} \exp\left[-2N|s|f\right] f^{2N\mu - 1} \tag{22}$$

Deriving the distribution of allele frequencies in the linear positive selection regime

We again take the low-frequency limit of the single-locus Langevin, but with $s > 0$. This derivation has previously been derived [68, 92], which I rederive below for the convenience of the reader.

$$\frac{\partial f}{\partial t} = sf + \mu + \sqrt{\frac{f}{N}}\eta(t) \tag{23}$$

This SDE can be solved using the method of characteristics. The solution to this problem has previously been derived [68, 92], which I rederive below for the sake of completeness. To gain a more comprehensive perspective, the unfamiliar reader should

refer to the source material. To work with Eq.23, it will be convenient to define the moment generating function

$$H(z,t) \equiv \left\langle e^{-zf(t)} \right\rangle \tag{24}$$

which obeys the following PDE

$$\frac{\partial H}{\partial t} = \left[sz - \frac{z^2}{2N} \right] \frac{\partial H}{\partial z} - \mu z H \tag{25}$$

We then define the log of the moment generating function along the characteristic curve in the absence of mutation ($\mu = 0$) z^* as

$$\Psi(t) \equiv \log \left[H(z^*(t_f - t), t_f - t) \right] \tag{26}$$

where tf is the current time and $H(z^*(t_f - t), t_f - t)$ is a reverse time function. The function $\Psi(t)$ satisfies the following ordinary differential equation and respective initial conditions, where $\varphi(t) \equiv z^*(t_f - t)$

$$\begin{aligned} \frac{\mathrm{d}\Psi(t)}{\mathrm{d}t} &= \mu\varphi(t) && \text{(27a)} \\ \Psi(0) &= \log \left[H(\varphi(0), t_f) \right] && \text{(27b)} \\ \Psi(t) &= \log \left[H(\varphi(t), 0) \right] = 0 && \text{(27c)} \end{aligned}$$

which gives us the following function and corresponding moment generating function

$$\begin{aligned} \Psi(t) &= \Psi(0) + \int_0^t \mu\varphi(t^{'})\mathrm{d}t^{'} && \text{(28a)} \\ H(z,t) &= \exp\left[-\mu \int_0^t \varphi(t^{'})\mathrm{d}t^{'} \right] && \text{(28b)} \end{aligned}$$

At this point, it is necessary to solve $\varphi(t)$, which requires us to look at Eq. 25 for the $\mu = 0$ case. To derive a solution for this case one must identify the family of curves, z^*, where $\frac{\mathrm{d}H(z^*,t)}{\mathrm{d}t} = 0$. The line $z^* = 2Ns$ is one characteristic curve. Using the chain rule, one can derive an expression for the ordinary differential equation of H

$$\frac{\mathrm{d}H(z^*,t)}{\mathrm{d}t} = \frac{\partial H}{\partial z^*}\frac{\mathrm{d}z^*}{\mathrm{d}t} + \frac{\partial H}{\partial t} = \frac{\partial H}{\partial z^*}\left[\frac{\mathrm{d}z^*}{\mathrm{d}t} + sz^* - \frac{(z^*)^2}{2N} \right] \tag{29}$$

Solving then the function is equal to zero, one obtains an ordinary differential equation for z^*

$$\frac{\mathrm{d}z^*}{\mathrm{d}t} = -sz^* + \frac{(z^*)^2}{2N} \tag{30}$$

We can then use our definition of φ and solve

$$\begin{aligned} \frac{\mathrm{d}\varphi(t)}{\mathrm{d}t} &= s\varphi(t) - \frac{\varphi(t)^2}{2N} && \text{(31a)} \\ \varphi(t) &= \frac{ze^{st}}{1 + \frac{z}{2Ns}\left(e^{st} - 1\right)} && \text{(31b)} \end{aligned}$$

Finally, one can substitute in the solution for $\varphi(t)$ into Eq.28a and solve the integral

$$\begin{aligned} H(z,t) &= \exp\left[-\mu\int_0^t \frac{ze^{st'}}{1+\frac{z}{2Ns}\left(e^{st'}-1\right)}\mathrm{d}t'\right] && \text{(32a)} \\ &= \exp\left[-2N\mu\log\left[1+\frac{z}{2Ns}\left(e^{st}-1\right)\right]\right]\Bigg|_0^t && \text{(32b)} \\ &= \exp\left[-2N\mu\log\left[1+\frac{z}{2Ns}\left(e^{st}-1\right)\right]\right] && \text{(32c)} \\ &= \left(1+z\cdot\frac{e^{st}-1}{2Ns}\right)^{-2N\mu} && \text{(32d)} \end{aligned}$$

Which is the moment generating function for the gamma distribution, providing us with a probability distribution of allele frequencies

$$p(f)df = \frac{df}{f_{max}^{2N\mu}\,\Gamma(2N\mu)} f^{2N\mu-1}e^{-f/f_{max}} \tag{33}$$

where $f_{max} \equiv \frac{e^{st}-1}{2Ns}$ [93]. In the neutral limit $s \to 0$, f_{max} reduces to $t/2N$. This result is corroborated by previous work, where that the maximum size that a neutral mutation can reach over $\sim t$ generations is $\sim t/N$ [93]. This result means that the presence or absence of positive selection alone does not invalidate the gamma distribution as a model for $p(f)$ in the $f \ll 1$ limit.

Evaluating prevalence predictions

Full derivations of the predicted prevalence of each model can be found in the Materials and Methods. The predicted values of prevalence were compared to the following estimate of observed prevalence.

$$\hat{\varrho} = \frac{1}{M}\sum_{m=1}^{M}(1-\delta_{f_m,0}) = 1-\frac{1}{M}\sum_{m=1}^{M}\delta_{f_m,0} \tag{34}$$

where f_m is the frequency of the alternative allele in the mth host and the Kronecker delta $\delta_{f_m,0}$ is equal to 1 if $f_m = 0$ and zero otherwise. To evaluate the success of the predictions I calculated the relative error of a given prediction

$$\varepsilon = \left|\frac{\text{Obs.} - \text{Pred.}}{\text{Obs.}}\right| \tag{35}$$

We performed permutation tests to determine whether the SLM had higher success among alleles with higher prevalence. By permuting all values of ε for a given species and calculating the Pearson correlation coefficient between it and the observed prevalence, I obtained a null distribution of correlation coefficients from which I calculated 95% intervals.

To determine whether there was a relationship between the error of the SLM and the fraction of hosts with strain structure across species, I implemented a permutation approach. First, for each species, I calculated the number of alleles in a given prevalence threshold (T total thresholds). I then permuted all values of ε and calculated the mean ε using the number of alleles that were found in each prevalence threshold, $(\bar{\varepsilon}_1, \cdots, \bar{\varepsilon}_T)$. The correlation coefficient was then calculated between $\bar{\varepsilon}_t$ and the fraction of hosts containing strains among all species for each prevalence threshold, allowing us to obtain a null distribution of correlation coefficients for all values of t. I only retained a prevalence threshold for a given species if there were at least 10 sites within the threshold.

Data and code availability

All raw sequencing data for the metagenomic samples used in this study were downloaded from the HMP [17, 94] (URL: `https://aws.amazon.com/datasets/human-microbiome-project/`). `MIDAS` output was processed using publicly available code [3] and all code written for this study is publicly available on GitHub: `https://github.com/garudlab/snv_prevalence`.

Acknowledgments

I thank S. Bald and R.W. Wolff for their assistance with `StrainFinder` and both N.R. Garud and R.W. Wolff for their comments at various stages of the project. Thanks to B.H. Good for pivotal discussions, sharing their insights, and for making their lecture notes available to the public. Thanks to S. Bubnovich, J. Grilli, and D. Reyes-González for their feedback on the manuscript. Finally, thanks to M.S. Ackerman for their assistance with `MAPGD`. This work was supported by the NSF Postdoctoral Research Fellowships in Biology Program under Grant No. 2010885 (W.R.S.). This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

Author contributions

W.R.S. conceptualized the project, performed all derivations and analyses, and wrote the manuscript.

References

1. Zhao, S. *et al.* Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host & Microbe* **25**, 656–667.e8 (2019).

2. Ghosh, O. M. & Good, B. H. Emergent evolutionary forces in spatial models of luminal growth in the human gut microbiota. Tech. Rep. (2021). URL `https://www.biorxiv.org/content/10.1101/2021.07.15.452569v1`. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

3. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biology* **17**, e3000102 (2019). URL `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000102`. Publisher: Public Library of Science.

4. Yaffe, E. & Relman, D. A. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nature Microbiology* **5**, 343–353 (2020).

5. Roodgar, M. *et al.* Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. *Genome Research* **31**, 1433–1446 (2021). URL `https://genome.cshlp.org/content/31/8/1433`. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

6. Ghalayini, M. *et al.* Evolution of a Dominant Natural Isolate of Escherichia coli in the Human Gut over the Course of a Year Suggests a Neutral Evolution with Reduced Effective Population Size. *Applied and Environmental Microbiology* **84**, e02377–17 (2018). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5835743/`.

7. Chen, D. W. & Garud, N. R. Rapid evolution and strain turnover in the infant gut microbiome. Tech. Rep. (2021). URL `https://www.biorxiv.org/content/10.1101/2021.09.26.461856v1`. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

8. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053–2067.e18 (2021). URL `https://www.sciencedirect.com/science/article/pii/S0092867421002415`.

9. Good, B. H. & Hallatschek, O. Effective models and the search for quantitative principles in microbial evolution. *Current Opinion in Microbiology* **45**, 203–212 (2018).

10. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research* **27**, 626–638 (2017).

11. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013). URL https://www.nature.com/articles/nature11711. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7430 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation;Metagenomics;Microbiota Subject_term_id: genetic-variation;metagenomics;microbiota.

12. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science (New York, N.Y.)* **341**, 1237439 (2013).

13. Goyal, A., Bittleston, L. S., Leventhal, G. E., Lu, L. & Cordero, O. X. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *eLife* **11**, e74987 (2022). URL https://doi.org/10.7554/eLife.74987. Publisher: eLife Sciences Publications, Ltd.

14. Wang, Z., Fridman, Y., Maslov, S. & Goyal, A. Fine-scale diversity of microbial communities due to satellite niches in boom-and-bust environments (2022). URL https://www.biorxiv.org/content/10.1101/2022.05.26.493560v2. Pages: 2022.05.26.493560 Section: New Results.

15. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011). URL https://www.nature.com/articles/nature09944. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7346 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Metagenomics;Microbiota Subject_term_id: metagenomics;microbiota.

16. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* **6**, 776–788 (2008). URL https://www.nature.com/articles/nrmicro1978. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group.

17. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).

18. Baumgartner, M., Bayer, F., Pfrunder-Cardozo, K. R., Buckling, A. & Hall, A. R. Resident microbial communities inhibit growth and antibiotic-resistance evolution of Escherichia coli in human gut microbiome samples. *PLOS Biology* **18**, e3000465 (2020). URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000465. Publisher: Public Library of Science.

19. Tett, A., Pasolli, E., Masetti, G., Ercolini, D. & Segata, N. Prevotella diversity, niches and interactions with the human host. *Nature Reviews Microbiology* **19**, 585–599 (2021). URL https://www.nature.com/articles/s41579-021-00559-y. Number: 9 Publisher: Nature Publishing Group.

20. Karcher, N. *et al.* Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biology* **21**, 138 (2020). URL https://doi.org/10.1186/s13059-020-02042-y.

21. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research* **26**, 1612–1625 (2016).

22. Niu, J. *et al.* Evolution of the Gut Microbiome in Early Childhood: A Cross-Sectional Study of Chinese Children. *Frontiers in Microbiology* **11** (2020). URL https://www.frontiersin.org/article/10.3389/fmicb.2020.00439.

23. Kundu, P., Blacher, E., Elinav, E. & Pettersson, S. Our Gut Microbiome: The Evolving Inner Self. *Cell* **171**, 1481–1493 (2017). URL https://www.sciencedirect.com/science/article/pii/S0092867417313715.

24. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012). URL https://www.nature.com/articles/nature11053. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7402 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Epidemiology;Genomics;Microbial ecology;Microbial genetics Subject_term_id: epidemiology;genomics;microbial-ecology;microbial-genetics.

25. Tierney, B. T. *et al.* The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host & Microbe* **26**, 283–295.e8 (2019). URL https://www.sciencedirect.com/science/article/pii/S193131281930352X.

26. Prosser, J. I. *et al.* The role of ecological theory in microbial ecology. *Nature Reviews Microbiology* **5**, 384–392 (2007). URL https://www.nature.com/articles/nrmicro1643. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group.

27. Marquet, P. A. *et al.* On Theory in Ecology. *BioScience* **64**, 701–710 (2014). URL https://doi.org/10.1093/biosci/biu098.

28. Good, B. H. & Rosenfeld, L. B. Eco-evolutionary feedbacks in the human gut microbiome. Tech. Rep., bioRxiv (2022). URL https://www.biorxiv.org/content/10.1101/2022.01.26.477953v1. Section: New Results Type: article.

29. Garud, N. R. & Pollard, K. S. Population Genetics in the Human Microbiome. *Trends in Genetics* **36**, 53–67 (2020). URL https://www.sciencedirect.com/science/article/pii/S0168952519302215.

30. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature Medicine* **25**, 1442–1452 (2019).

31. Shoemaker, W. R., Chen, D. & Garud, N. R. Comparative Population Genetics in the Human Gut Microbiome. *Genome Biology and Evolution* (2021). URL https://doi.org/10.1093/gbe/evab116.

32. Cui, W., Marsland, R. & Mehta, P. Diverse communities behave like typical random ecosystems. *Physical Review E* **104**, 034416 (2021). URL https://link.aps.org/doi/10.1103/PhysRevE.104.034416. Publisher: American Physical Society.

33. Advani, M., Bunin, G. & Mehta, P. Statistical physics of community ecology: a cavity solution to MacArthur's consumer resource model. *Journal of Statistical Mechanics (Online)* **2018**, 033406 (2018).

34. Descheemaeker, L. & de Buyl, S. Stochastic logistic models reproduce experimental time series of microbial communities. *eLife* **9**, e55650 (2020). URL https://doi.org/10.7554/eLife.55650. Publisher: eLife Sciences Publications, Ltd.

35. Shoemaker, W. R., Locey, K. J. & Lennon, J. T. A macroecological theory of microbial biodiversity. *Nature Ecology & Evolution* **1**, 1–6 (2017). URL https://www.nature.com/articles/s41559-017-0107. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biodiversity;Community ecology;Macroecology;Microbial ecology Subject_term_id: biodiversity;community-ecology;macroecology;microbial-ecology.

36. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nature Communications* **11**, 4743 (2020). URL https://www.nature.com/articles/s41467-020-18529-y. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biodiversity;Community ecology;Ecological modelling;Macroecology;Microbial ecology Subject_term_id: biodiversity;community-ecology;ecological-modelling;macroecology;microbial-ecology.

37. Ji, B. W., Sheth, R. U., Dixit, P. D., Tchourine, K. & Vitkup, D. Macroecological dynamics of gut microbiota. *Nature Microbiology* **5**, 768–775 (2020). URL https://www.nature.com/articles/s41564-020-0685-1. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Microbial communities;Ecological modelling;Microbial ecology Subject_term_id: communities;ecological-modelling;microbial-ecology.

38. Wolff, R., Shoemaker, W. & Garud, N. Ecological Stability Emerges at the Level of Strains in the Human Gut Microbiome. Tech. Rep. (2021). URL https://www.biorxiv.org/content/10.1101/2021.09.30.462616v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

39. Phillips, R. Theory in Biology: Figure 1 or Figure 7? *Trends in Cell Biology* **25**, 723–729 (2015). URL https://www.cell.com/trends/cell-biology/abstract/S0962-8924(15)00194-4. Publisher: Elsevier.

40. Bhattacharjee, S. M. & Seno, F. A measure of data collapse for scaling. *Journal of Physics A: Mathematical and General* **34**, 6375–6380 (2001). URL https://doi.org/10.1088/0305-4470/34/33/302. Publisher: IOP Publishing.

41. Stanley, H. E. Scaling, universality, and renormalization: Three pillars of modern critical phenomena. *Reviews of Modern Physics* **71**, S358–S366 (1999). URL https://link.aps.org/doi/10.1103/RevModPhys.71.S358. Publisher: American Physical Society.

42. Theys, K. *et al.* Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. *PLOS Genetics* **14**, e1007420 (2018). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007420. Publisher: Public Library of Science.

43. Vogl, C. & Bergman, J. Computation of the Likelihood of Joint Site Frequency Spectra Using Orthogonal Polynomials. *Computation* **4**, 6 (2016). URL https://www.mdpi.com/2079-3197/4/1/6. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

44. Marquet, P. A. *et al.* Scaling and power-laws in ecological systems. *Journal of Experimental Biology* **208**, 1749–1769 (2005). URL https://journals.biologists.com/jeb/article/208/9/1749/9378/Scaling-and-power-laws-in-ecological-systems.

45. Ramsayer, J., Fellous, S., Cohen, J. E. & Hochberg, M. E. Taylor's Law holds in experimental bacterial populations but competition does not influence the slope. *Biology Letters* **8**, 316–319 (2012). URL https://royalsocietypublishing.org/doi/10.1098/rsbl.2011.0895. Publisher: Royal Society.

46. Kendal, W. S. & Jørgensen, B. Taylor's power law and fluctuation scaling explained by a central-limit-like convergence. *Physical Review E* **83**, 066115 (2011). URL https://link.aps.org/doi/10.1103/PhysRevE.83.066115.

47. Kendal, W. S. A scale invariant clustering of genes on human chromosome 7. *BMC Evolutionary Biology* 10 (2004).

48. Kendal, W. S. An Exponential Dispersion Model for the Distribution of Human Single Nucleotide Polymorphisms. *Molecular Biology and Evolution* **20**, 579–590 (2003). URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msg057.

49. Taylor, L. R. Aggregation, Variance and the Mean. *Nature* **189**, 732–735 (1961). URL https://www.nature.com/articles/189732a0. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4766 Primary_atype: Research Publisher: Nature Publishing Group.

50. He, M. *et al.* Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 7527–7532 (2010).

51. Bhatia, R. & Davis, C. A Better Bound on the Variance. *The American Mathematical Monthly* **107**, 353–357 (2000). URL https://doi.org/10.1080/00029890.2000.12005203. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00029890.2000.12005203.

52. Cohen, J. E. & Xu, M. Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. *Proceedings of the National Academy of Sciences* **112**, 7749–7754 (2015). URL https://www.pnas.org/content/112/25/7749. Publisher: National Academy of Sciences Section: Biological Sciences.

53. Giometto, A., Formentin, M., Rinaldo, A., Cohen, J. E. & Maritan, A. Sample and population exponents of generalized Taylor's law. *Proceedings of the National Academy of Sciences* **112**, 7755–7760 (2015). URL https://www.pnas.org/content/112/25/7755. Publisher: National Academy of Sciences Section: Biological Sciences.

54. Xiao, X., Locey, K. J. & White, E. P. A Process-Independent Explanation for the General Form of Taylor's Law. *The American Naturalist* **186**, E51–E60 (2015). URL https://www.journals.uchicago.edu/doi/full/10.1086/682050. Publisher: The University of Chicago Press.

55. Transtrum, M. K. *et al.* Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of Chemical Physics* **143**, 010901 (2015). URL https://aip.scitation.org/doi/10.1063/1.4923066. Publisher: American Institute of Physics.

56. Zaoli, S. & Grilli, J. The stochastic logistic model with correlated carrying capacities reproduces beta-diversity metrics of microbial communities. *PLOS Computational Biology* **18**, e1010043 (2022). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010043. Publisher: Public Library of Science.

57. Zaoli, S. & Grilli, J. A macroecological description of alternative stable states reproduces intra- and inter-host variability of gut microbiome. *Science Advances* **7**, eabj2882 (2021). URL https://www.science.org/doi/10.1126/sciadv.abj2882. Publisher: American Association for the Advancement of Science.

58. Gardiner, C. W. *Stochastic methods: a handbook for the natural and social sciences*. No. 13 in Springer series in synergetics (Springer, Berlin Heidelberg, 2009), 4th ed edn.

59. Engen, S. & Lande, R. Population Dynamic Models Generating Species Abundance Distributions of the Gamma Type. *Journal of Theoretical Biology* **178**, 325–331 (1996). URL https://www.sciencedirect.com/science/article/pii/S0022519396900284.

60. Shoemaker, W. R. & Lennon, J. T. Predicting Parallelism and Quantifying Divergence in Microbial Evolution Experiments. *mSphere* (2022). URL https://journals.asm.org/doi/abs/10.1128/msphere.00672-21. Publisher: American Society for Microbiology 1752 N St., N.W., Washington, DC.

61. Gaston, K. J. *et al.* Abundance–occupancy relationships. *Journal of Applied Ecology* **37**, 39–59 (2000). URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2664.2000.00485.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2664.2000.00485.x.

62. Shade, A. *et al.* Macroecology to Unite All Life, Large and Small. *Trends in Ecology & Evolution* **33**, 731–744 (2018).

63. Sloan, W. T., Woodcock, S., Lunn, M., Head, I. M. & Curtis, T. P. Modeling Taxa-Abundance Distributions in Microbial Communities using Environmental Sequence Data. *Microbial Ecology* **53**, 443–455 (2007). URL https://doi.org/10.1007/s00248-006-9141-x.

64. Burns, A. R. *et al.* Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *The ISME Journal* **10**, 655–664 (2016). URL https://www.nature.com/articles/ismej2015142. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Microbial ecology Subject_term_id: microbial-ecology.

65. Cohen, J. E. Every variance function, including Taylor's power law of fluctuation scaling, can be produced by any location-scale family of distributions with positive mean and variance. *Theoretical Ecology* **13**, 1–5 (2020). URL https://doi.org/10.1007/s12080-019-00445-7.

66. Nei, M. The frequency distribution of lethal chromosomes in finite populations. *Proceedings of the National Academy of Sciences of the United States of America* **60**, 517–524 (1968). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC225078/.

67. Ewens, W. J. *Mathematical population genetics* (Springer, 2010). OCLC: 778838329.

68. Good, B. H. Linkage disequilibrium between rare mutations. *Genetics* **220**, iyac004 (2022). URL https://doi.org/10.1093/genetics/iyac004.

69. Cvijović, I., Good, B. H. & Desai, M. M. The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics* **209**, 1235–1278 (2018). URL https://doi.org/10.1534/genetics.118.301058.

70. Yan, Y., Nguyen, L. H., Franzosa, E. A. & Huttenhower, C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine* **12**, 71 (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7427293/.

71. Dijkshoorn, L., Ursing, B. M. & Ursing, J. B. Strain, clone and species: comments on three basic concepts of bacteriology. *Journal of Medical Microbiology* **49**, 397–401 (2000).

72. Zhu, A., Sunagawa, S., Mende, D. R. & Bork, P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biology* **16**, 82 (2015). URL https://doi.org/10.1186/s13059-015-0646-9.

73. Shapiro, B. J. & Polz, M. F. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology* **22**, 235–247 (2014).

74. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host & Microbe* **23**, 229–240.e5 (2018). URL https://www.cell.com/cell-host-microbe/abstract/S1931-3128(18)30038-6. Publisher: Elsevier.

75. Chesson, P. MacArthur's consumer-resource model. *Theoretical Population Biology* **37**, 26–38 (1990). URL https://www.sciencedirect.com/science/article/pii/004058099090025Q.

76. Good, B. H., Martis, S. & Hallatschek, O. Adaptation limits ecological diversification and promotes ecological tinkering during the competition for substitutable resources. *Proceedings of the National Academy of Sciences* **115**, E10407–E10416 (2018). URL https://www.pnas.org/content/115/44/E10407. Publisher: National Academy of Sciences Section: PNAS Plus.

77. Ho, P.-Y., Good, B. H. & Huang, K. C. Competition for fluctuating resources reproduces statistics of species abundance over time across wide-ranging microbiotas. *eLife* **11**, e75168 (2022). URL https://doi.org/10.7554/eLife.75168. Publisher: eLife Sciences Publications, Ltd.

78. Madi, N., Chen, D., Wolff, R., Shapiro, B. J. & Garud, N. R. Community diversity is associated with intra-species genetic diversity and gene loss in the human gut microbiome. Tech. Rep., bioRxiv (2022). URL https://www.biorxiv.org/content/10.1101/2022.03.08.483496v1. Section: New Results Type: article.

79. Li, J. *et al.* Antibiotic Treatment Drives the Diversification of the Human Gut Resistome. *Genomics, Proteomics & Bioinformatics* **17**, 39–51 (2019). URL https://www.sciencedirect.com/science/article/pii/S1672022919300609.

80. N'Guessan, A., Brito, I. L., Serohijos, A. W. R. & Shapiro, B. J. Mobile Gene Sequence Evolution within Individual Human Gut Microbiomes Is Better Explained by Gene-Specific Than Host-Specific Selective Pressures. *Genome Biology and Evolution* **13**, evab142 (2021). URL https://doi.org/10.1093/gbe/evab142.

81. Simonet, C. & McNally, L. Kin selection explains the evolution of cooperation in the gut microbiota. *Proceedings of the National Academy of Sciences* **118**, e2016046118 (2021). URL https://www.pnas.org/doi/10.1073/pnas.2016046118. Publisher: Proceedings of the National Academy of Sciences.

82. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology* **14**, e1002533 (2016). URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002533. Publisher: Public Library of Science.

83. Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature Biotechnology* **34**, 64–69 (2016). URL https://www.nature.com/articles/nbt.3416. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Metagenomics Subject_term_id: metagenomics.

84. Zlitni, S. *et al.* Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Medicine* **12**, 50 (2020). URL https://doi.org/10.1186/s13073-020-00747-0.

85. DeMaere, M. Z. & Darling, A. E. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biology* **20**, 46 (2019). URL https://doi.org/10.1186/s13059-019-1643-1.

86. Press, M. O. *et al.* Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. Tech. Rep. (2017). URL https://www.biorxiv.org/content/10.1101/198713v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

87. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486**, 215–221 (2012). URL https://www.nature.com/articles/nature11209. Bandiera_abtest: a Cc_license_type: cc_y Cg_type: Nature Research Journals Number: 7402 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Medical research;Metagenomics;Microbiome Subject_term_id: medical-research;metagenomics;microbiome.

88. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10** (2021). URL https://doi.org/10.1093/gigascience/giab008. Giab008, https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf.

89. Lynch, M., Bost, D., Wilson, S., Maruki, T. & Harrison, S. Population-Genetic Inference from Pooled-Sequencing Data. *Genome Biology and Evolution* **6**, 1210–1218 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4040993/.

90. Guirao-Rico, S. & González, J. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. *Molecular Ecology Resources* **21**, 1216–1229 (2021). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13343. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13343.

91. Xiao, X., Thibault, K., Harris, D. J., Baldridge, E. & White, E. weecology/macroecotools: v0.4.0 (2016). URL https://zenodo.org/record/166721.

92. Good, B. H. APHYS 237/BIO251: Quantitative evolutionary dynamics and genomics (2021). "URL: https://bgoodlab.github.io/courses/apphys237/.

93. Fisher, D. S. Course 11 Evolutionary dynamics. In *Les Houches*, vol. 85, 395–446 (Elsevier, 2007). URL https://linkinghub.elsevier.com/retrieve/pii/S0924809907800187.

94. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012). URL https://www.nature.com/articles/nature11234. Number: 7402 Publisher: Nature Publishing Group.

Figures

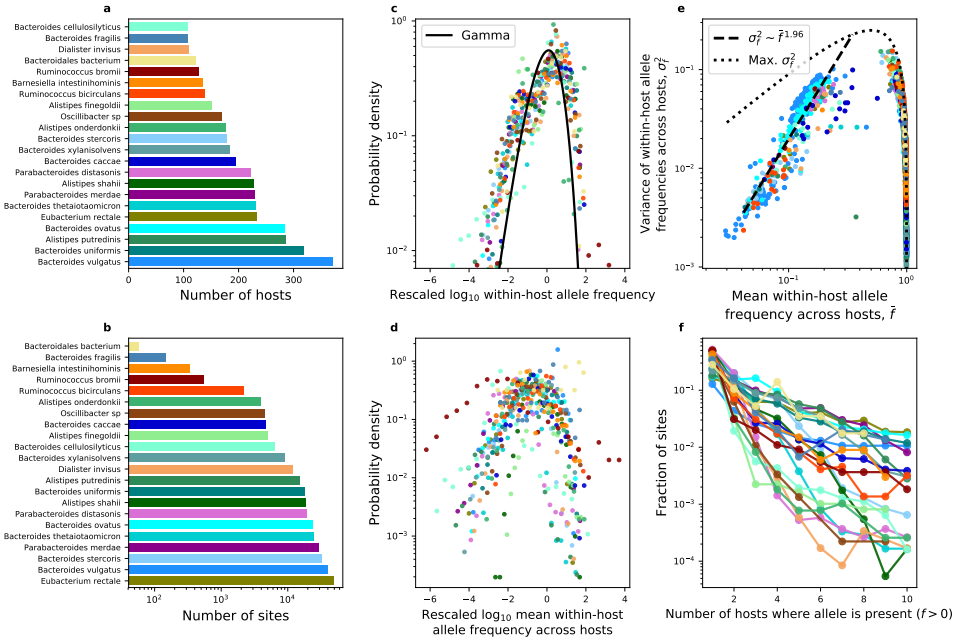

Figure 1. Distributions of genetic diversity exhibit similar statistical forms across phylogenetically distant species in the human gut. a,b) Using the set of hosts and alleles that passed the quality control criteria, I was able to evaluate the degree of similarity in patterns of genetic diversity for the 22 most prevalent bacterial species. **c)** The distribution of within-host frequencies for all alleles for all hosts as well as **(d)** the distribution of mean allele frequencies are qualitatively similar across species, where the logarithm of all values for each species has been rescaled using the standard score (i.e., z-score) to facilitate visual comparison. The similarity in the form of each distribution across species suggests that it is reasonable to identify a single model of allelic dynamics. **e)** The relationship between statistical moments of within-host allele frequencies is consistent across species, as there is a strong linear relationship between the mean frequency of an allele and its the variance on a log-log scale (i.e., Taylor's Law). To reduce the contribution of an excess number of zeros towards estimates of $\bar{f}$ and σ_f^2, I only examined alleles with non-zero values of f in $\geq 35\%$ of hosts. **f)** Finally, the number of hosts where an allele is present can contain information about its ecological or evolutionary dynamics. I found that the typical allele is present in a low number of hosts and that the fraction of alleles in a given number of hosts tends to decay as the number of hosts increases across species. All sites in this analysis are synonymous, identical analyses were performed on alleles at nonsynonymous sites (Fig. S2). Species within the same genus were assigned the same primary/secondary color with varying degrees of saturation.

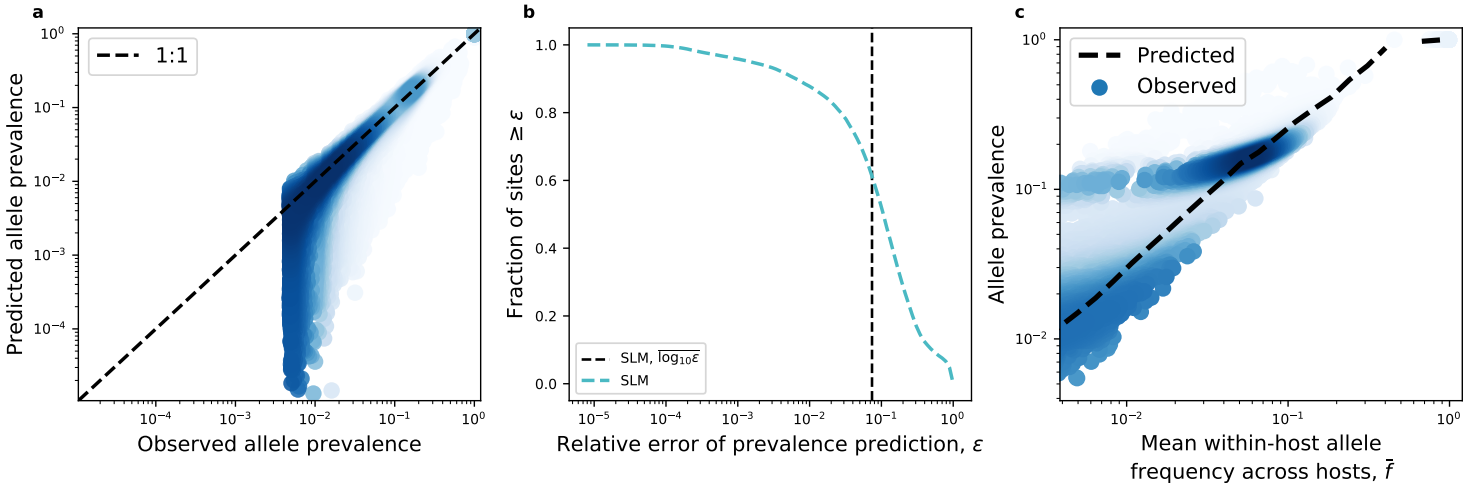

Figure 2. The SLM successfully predicts genetic patterns for prevalent alleles. To demonstrate a particular instance of this pattern, I examined the degree of success for each model on a common commensal gut species, *B. vulgatus*. **a)** The SLM generally succeeds in predicting the prevalence of alleles. This assessment is particularly true for alleles with high observed estimates of prevalence, all of which fall on a one-to-one line indicating perfect accuracy. **b)** The visual evidence of the predictive success of the SLM is confirmed by quantifying the relative error of the prevalence predictions. The mean of the logarithm of the relative error of the SLM ($\overline{\log_{10}\varepsilon}$, dashed black line) is ~ 0.1. **c)** This success is also present in the relationship between patterns of within-host vs. among-host allelic diversity, as the SLM generally succeeded in predicting the relationship between the mean frequency of a allele ($\bar{f}$) and its prevalence with zero statistical fitting. All analyses here were performed on alleles at synonymous sites. Visualizations of the predictions in this plot for all species for nonsynonymous and synonymous sites can be found in the Supplemental Text.

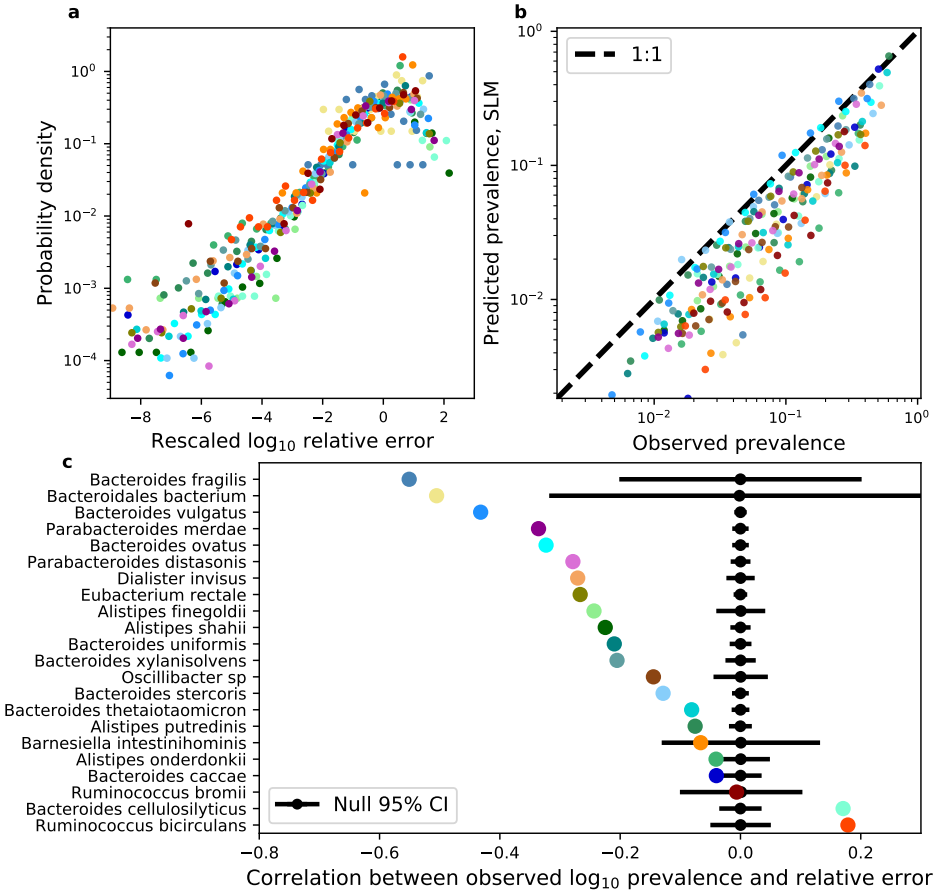

Figure 3. The SLM succeeds and fails in a consistent manner across phylogenetically distant species. a) Distributions of logarithmic relative errors rescaled by their mean and standard deviation exhibit qualitatively similar forms across species. **b)** When we examine the predictions of the SLM across a range of observed prevalences, the values tend to follow a one-to-one relationship across species. However, high prevalence alleles tend to be closer to the one-to-one line, suggesting that there is a negative relationship between the relative error of the SLM and observed prevalence. **c)** A permutation test confirms that there is a negative relationship between the accuracy the SLM and the observed prevalence of an allele, as the correlation between the two quantities is far more negative than expected by chance for the majority of species (16/22). All analyses in this plot were performed using synonymous sites. Identical analyses and equivalent results for nonsynonymous sites can be found in the Supplemental Text (Fig. S9).

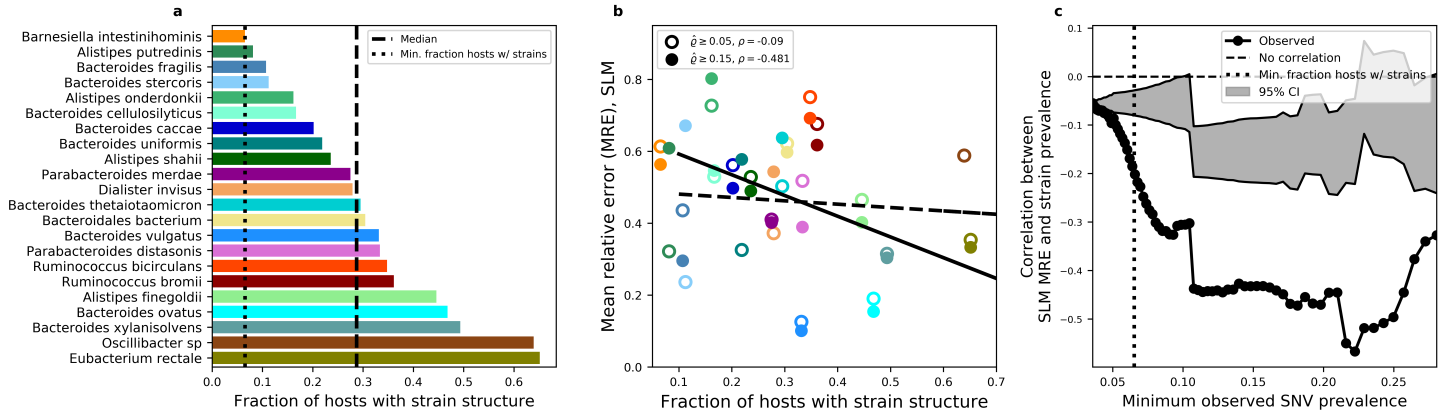

Figure 4. The presence of strain structure is correlated with the accuracy of the SLM. a) The presence of absence of strain structure can be inferred for a given species within a given host, providing the means to determine whether the degree of deviation from perfect accuracy (i.e., relative error) of the SLM is correlated with the fraction of hosts that harbor strain structure. **b)** By examining the relationship between error and strain structure across species within a low ($\hat{\varrho} \geq 0.05$: dashed line) and high ($\hat{\varrho} \geq 0.15$: solid line) threshold value of observed allele prevalence, I found that the correlation is higher for alleles with a high prevalence threshold. **c)** One can resolve this dependence using a range of observed prevalence bins and an appropriate permutation test, where the correlation between the error and the fraction of hosts with strain structure tends to become increasingly negative (Materials and Methods). Notably, the correlation drops off after the minimum fraction of hosts where strain structure was inferred (> 1 strain) and continues as allele prevalence increases. In order to maintain a high number of species across values where the existence of strain structure could feasibly contribute, I restricted my analysis to prevalence values less than the median fraction of hosts with strain structure across species.