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: Quantitative Life Sciences, 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 that it harbors. Research efforts have demonstrated that a considerable fraction of within-host genetic variation in the human gut is driven by the ecological dynamics of co-occurring strains belonging to the same species, suggesting that an ecological lens may provide insight into empirical patterns of genetic diversity. Indeed, an ecological model of self-limiting growth and environmental noise known as the Stochastic Logistic Model (SLM) was recently shown to successfully predict the temporal dynamics of strains within a single human host. However, its ability to predict patterns of genetic diversity across human hosts has yet to be tested. In this manuscript I determine whether the predictions of the SLM explain patterns of genetic diversity across unrelated human hosts for 22 common microbial species. Specifically, the stationary distribution of the SLM explains the distribution of allele frequencies across hosts and predicts the fraction of hosts harboring a given allele (i.e., prevalence) for a considerable fraction of sites. The accuracy of the SLM was correlated with independent estimates of strain structure, suggesting that patterns of genetic diversity in the gut microbiome follow statistically similar forms across human hosts due to the existence of strain-level ecology.

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, 3–9]. 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 over time on timescales ranging from months to years [1, 10–13]. Such within-host genetic structure is also reflected in the shape of

phylogenetic trees constructed from microbial isolates [11, 14–19]. This pattern arises due to the co-occurrence of a few ($\sim \mathcal{O}(1-4)$) genetically and ecologically diverged strains that belong to the same species, a process known as oligocolonization [3, 11]. This sub-species ecological structure that can occur within a host 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 [20, 21]. Thus, the dynamics of the genetic variants that comprise a given strain are subject to exogenous and endogenous ecological forces [22–24]. However, evolution within a strain does not stop, as genetic variants continue to be acquired and segregate over time within a given strain [5]. Such dynamics are a clear departure from those captured by standard population genetic models, where genetic variants either arise in a population due to mutation or are introduced by migration and then proceed towards extinction or fixation (i.e., origin-fixation models), suggesting that measures of genetic diversity estimated within the human gut are shaped by the ecology of strains alongside evolutionary forces such as low recombination rates generating physical linkage between alleles [3, 5, 10, 25].

This confluence of ecological and evolutionary dynamics requires new approaches and theory for characterizing genetic diversity in the human gut. Many studies tackle such complexity by examining individual species [6, 26–28] or by searching for genetic differences between species [11, 12, 29–33]. While such approaches are useful for identifying individual species that are potential contributors towards specific conditions such as disease or the ability to metabolize certain resources, it is difficult to translate isolated observations into general patterns. By focusing on individual species and differences between species it is plausible that uncharacterized patterns of genetic diversity that are generalizable across species may have been overlooked.

As an alternative, it is reasonable to first identify genetic patterns that are similar across species (i.e., statistical invariance). Such an approach may provide the empirical motivation necessary to identify mathematical models that can explain said patterns and aid in the identification of underlying ecological or evolutionary dynamics [10, 34–36]. In recent years, substantial progress has been made towards characterizing the typical microbial evolutionary dynamics across species that operate within and across human hosts [3, 4, 7, 8, 37–39]. An example of such a pattern is the observation that the relationship between synonymous nucleotide divergence (a proxy for evolutionary time) and the ratio of nonsynonymous and synonymous divergence (dS vs. dN/dS) falls on a single curve across microbial species in the human gut, representing 20 genera, 14 families, 7 orders, 6 classes, and 5 phyla [3, 39]. While this approach can often be limited by the number of observations and measurement error, modern data curation methods can often alleviate these limitations. 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 [40, 41].

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 [42–47]. This approach emphasizes the benefits of identifying patterns of diversity that are statistically invariant, motivating the development of quantitative predictions derived from ecological 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 extended to temporal patterns of strain-level ecology within a single host [46]. This consistency in strain-level patterns provided the motivation to apply an established model of ecology to predict macroecological quantities within a single host over time, the Stochastic Logistic Model

of growth (SLM). In macroecology, the SLM has been found to successfully characterize the distribution of species relative abundance across hosts and over time within a host (i.e., the Abundance Fluctuation Distribution (AFD)), the relationship between the mean abundance of a species and the fraction of hosts where it is present (i.e., the abundance-prevalence relationship [48]), and the relationship between the mean and variance of the abundance of a species (i.e., Taylor's Law [49]) [44]. Inspired by the success of the SLM, it was recently applied to the strain-level to explain the temporal form of the AFD and Taylor's Law within a single human host [46]. In this study it was found that the temporal dynamics of strains within a single healthy human host invariant with respect to time (i.e., stationary). Motivated by this result, it was determined that the empirical distribution of strain frequencies over time followed the distribution of the SLM at stationarity, a gamma distribution. The results of this study suggest that the SLM, a model that succeeded in predicting patterns of strains within a single human host, may also succeed in predicting patterns of genetic diversity across unrelated hosts due to the existence of strain structure.

In this study, I sought to determine whether the SLM as a model of ecology was capable of quantitatively predicting patterns of genetic diversity across hosts due to the existence of strain structure. I identified patterns of diversity that remained statistically invariant among phylogenetically distant species, providing the motivation necessary to identify the SLM as a plausible model of across-host patterns of diversity. To evaluate the feasibility of the SLM while accounting for the effects of sampling, I obtained predictions for the fraction of hosts harboring an allele at a given nucleotide site (i.e., prevalence) using zero free parameters. I identified evolutionary models of allele frequencies that predict the same stationary probability distribution as the SLM and found that their assumptions are unrealistic to explain the data. To confirm that the success of the SLM was due to the presence of strains, I inferred whether strain structure was present in each host for each species using an established computational approach, finding that the presence of strain structure was correlated with the accuracy of the SLM in predicting allelic prevalence.

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 [50–52]. 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 using the Maximum-likelihood Analysis of Population Genomic Data `MAPGD` program (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, S1b). These results, and their existence for both synonymous and nonsynonymous sites, provides the empirical basis necessary to formulate quantitative predictions.

First, I obtained the distribution of across-host allele frequencies for each nucleotide site and then pooled the frequencies of all sites. If the typical allele was present due to evolutionary forces, then the empirical distribution of within-host allele frequencies across hosts would be the equivalent to the ensemble of single-site frequency spectra expected from within-host evolution [53, 54]. If the typical allele was present because it was on the background of a strain, then the macroecological view of this distribution is that it captures the distribution of relative abundances of strains across hosts, the AFD [44]. Furthermore, 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 distribution of within-host allele frequencies across hosts for phylogenetically distant species.

In order to determine whether different distributions share a single form it is useful to rescale them by key parameters [50]. Inspired by prior work [44], I rescaled the distribution of within-host frequencies across hosts by 1) pooling all non-zero frequencies for a given species, 2) log-transforming all frequencies, 3) calculating the mean and standard deviation of the frequency, and 4) calculating each log-transformed frequency as a standard score (i.e., z-score). By repeating this process for each species, one can determine whether the form of the distribution qualitatively varies across species or whether they simply differ in their statistical moments. The distribution of within-host allele frequencies had a similar qualitative form across species in the human gut for synonymous (Fig. 1c) and nonsynonymous sites (S1c), suggesting that a single distribution may be sufficient to characterize all species. Regions of the distribution are well-captured by the gamma distribution [44], suggesting that it would be informative to examine models that lead to a gamma distribution and then assess the gamma's capacity to predict quantities calculated from individual alleles. As a point of comparison, I fit the distribution in Fig. 1c using a lognormal distribution. This distribution was previously used to evaluate the AFD at the species level in disparate ecosystems [44]. The lognormal clearly deviates from the bulk of the distribution, a result that is even more apparent when the probability density is plotted as a survival probability (Figs. S2a, S3a). An Akaike Information Criterion (AIC) test supports this conclusion (Synonymous: $\mathrm{AIC}_{\mathrm{gamma}} = 6,277,330$, $\mathrm{AIC}_{\mathrm{lognormal}} = 6,618,492$; Nonsynonymous: $\mathrm{AIC}_{\mathrm{gamma}} = 3,359,295$, $\mathrm{AIC}_{\mathrm{gamma}} = 3,490,025$).

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. I repeated the same standard score rescaling procedure for the logarithm of the mean within-host allele frequency across hosts ($\bar{f}$). The distribution of $\bar{f}$ tended to overlap across species for both synonymous (Figs. 1d, S2b) and nonsynonymous sites (Figs. S1d, S2b). While here I am not explicitly interested in the processes that shape the mean distribution, as the mean will be used as an empirical input for calculating predictions in the subsequent section, the result does suggest that

The mean and variance of random variables frequently follows linear relationships on logarithmic scales across biological systems, most notably in patterns of biodiversity in ecological communities [44, 55, 56] but also among population genetic patterns [57–59]. 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 [44]. 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. S1e), 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, 12, 39, 60], 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 $\sigma_f^2 \leq (\max(f) - \bar{f})(\bar{f} - \min(f)) = (1 - \bar{f})\bar{f}$ [61]. The empirical relationship between $\bar{f}$ and σ_f^2 follows the inequality across species for $\bar{f} \gtrsim 0.35$, suggesting that the 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 [62, 63]. However, similar to ecological analyses of the relationship between the mean and variance of species abundances [44, 56, 64], 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. While the existence of this relationship may not be system-specific [64], it does allow us to make a claim about the relationship between statistical moments.

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. S1f). Alternatively stated, the fraction of hosts harboring a given allele (i.e., prevalence) is typically low.

Predicting the prevalence of an allele across hosts

The existence of multiple patterns of genetic diversity that are universal across evolutionarily distant bacterial species suggests that comparable dynamics are ultimately responsible. The next task is to identify a candidate model 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 (i.e., no statistical fitting) rather than relying on estimates of free parameters via statistical inference (i.e., statistical fitting) [65].

We begin with the assumption that the frequency dynamics of a typical allele within a host 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 relevant ecological dynamics can be used. However, regardless of the underlying dynamics, the available data constrains the ways in which a given model can be evaluated. Given that temporal metagenomic data for the human gut microbiome remains restricted to a small number of hosts, I focused on 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 (i.e., $p(f,t)$), so I instead focused on stationary probability distributions and limiting cases where time-dependence is captured by parameters that can, in principle, be estimated (i.e., $p(f)$). This assumption of stationarity (i.e., time-invariance) is supported by previous research efforts that examined macroecological patterns that were stationary with respect to time at the strain level [46].

To model the dynamics of an allele that is on the genetic background a strain, it is necessary to identify essential features of growth. There are two main features that are necessary to consider the deterministic dynamics of strain dynamics: 1) that the rate of growth is often exponential when a species or strain is far from its carrying capacity and 2) that growth is self-limiting. There is also the need to consider stochasticity in growth that can be driven by environmental noise. To capture these features I examined a Langevin equation known as the Stochastic Logistic Model of growth (SLM), a model that has recently been shown to describe a range of macroecological patterns for microbial communities across disparate environments [44, 66, 67] as well as the temporal dynamics of strains within a single human host [46]. The SLM is defined as

$$\underbrace{\frac{\partial f}{\partial t} = \frac{f}{\tau_i}\left(1 - \frac{f}{K_i}\right)}_{\text{Self-limiting growth}} + \underbrace{\sqrt{\frac{\sigma_{\tau_i}}{\tau_i}} f \cdot \eta(t)}_{\text{Environmental noise}} \tag{1}$$

Here I am assuming that the allele I observed within a given host was present because it was on the background of a strain. This interpretation means that the fluctuations of the allele over time are due to the ecological fluctuations of the strain, where the terms $\frac{1}{\tau_i}$, K_i, and σ_{τ_i} respectively represent the intrinsic growth rate, the carrying capacity within a given species in terms of relative abundance ($0 \leq K_i \leq 1$), and the coefficient of variation of growth rate fluctuations of the ith strain.

Environmental noise is captured by the product of a linear frequency term (as opposed to demographic noise, which would be captured by the term $\sqrt{f}$), the compound parameter $\sqrt{\frac{\sigma_{\tau_i}}{\tau_i}}$, and a Brownian noise term $\eta(t)$ that introduces stochasticity into the equation. Using standard definitions of stochastic differential equations, the expected value of $\eta(t)$ is $\langle \eta(t) \rangle = 0$ [68]. The dependence of $\eta(t')$ at time t' on an earlier time $\eta(t)$ is defined as $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$ [68]. This standard definition means that if the noise term is shifted in time it has zero correlation with itself, otherwise it is identical to itself.

This definition of a stochastic differential equation is convenient in that it is possible to obtain a partial differential equation describing how the probability distribution of f changes with time (i.e., the Fokker-Planck equation) [68]. Once this equation is obtained, the probability distribution of f at stationarity (i.e., no dependence on time) can be obtained. One finds that the SLM predicts that the frequency of a given allele on the background of a strain follows a gamma distribution (additional detail provided in Materials and Methods)

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

This distribution is fully characterized by the mean frequency and the squared inverse of the coefficient of variation across hosts ($\beta = \frac{\bar{f}^2}{\sigma_f^2}$) across hosts [44, 69]

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

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, S1d,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 predict $\bar{f}$ 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 [44, 70], including the human gut microbiome [3, 7, 44].

In order to test prevalence predictions it is necessary to account for the sampling effort at a given site in a given host (i.e., total depth of sequencing coverage). This can be accomplished by deriving the sampling distribution of the gamma, providing the probability of observing A reads of a gamma distributed allele with D coverage.

$$\Pr[A|D, \beta, \beta/\bar{f}] = \frac{\Gamma(\beta + A)}{A!\Gamma(\beta)} \left(\frac{\bar{f}D}{\beta + \bar{f}D} \right)^A \left(\frac{\beta}{\beta + \bar{f}D} \right)^\beta \tag{4}$$

A value $A = 0$ represents the absence of an allele, which can be used to define presence as the complement, providing a natural definition of the prevalence of an allele across hosts.

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

where I have defined prevalence as the average of the probability of presence over M hosts. While Eq. 5 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 ($A_{\text{cutoff}} = 10$; S4). 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

$$\langle \varrho \rangle = 1 - \frac{1}{M} \sum_{m=1}^{M} \sum_{A_i=0}^{A_{\text{cutoff}}-1} \Pr[A_i|D_m, \beta, \beta/\bar{f}] \tag{6}$$

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 [71], a pattern that has been found to exhibit statistically similar forms at the species level across microbial systems [48, 72, 73] and can be quantitatively explained through the existence of macroecological laws [44].

The Stochastic Logistic Model succeeds at predicting allelic prevalence

By examining the relationship between observed and predicted allelic prevalence, I found that the SLM generally succeeded in predicting this relationship for both synonymous and nonsynonymous sites using zero free parameters (Figs. 2a, S5, S6). Alternatively stated, predictions were obtained by computing the predicted prevalence using Eq. 6 without the need to perform statistical fitting. 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 within the human gut are driven by ecological dynamics (Figs. 2b, S7, S8). 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, S9, S10). 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 evidence for the existence of Taylor's Law constrains the parameterization of the SLM and its subsequent interpretation [44, 74]. 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, S12, S13). 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, one 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 across hosts 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 [46]. 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 [44], 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 operating within a host, 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 model describing the dynamics of both the ecological and evolutionary dynamics of f within a host it is difficult to parse alleles into discrete "ecological" and "evolutionary" categories. Regardless, the distributions of prevalence prediction errors have a strikingly

bimodal form for several species (Figs. S7,S8), 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 [44]. This reliance on statistical moments estimated from observational data suggests that alternative models that also predict a gamma distribution are equivalent candidates in the absence of additional evidence. In contrast, models of ecology and evolution that predict alternative forms of the distribution presented in Fig. 1c are inappropriate on the outset. For example, models of neutral evolution $\frac{\partial f}{\partial t} \propto \sqrt{\frac{f(1-f)}{N}}\eta(t)$ and neutral ecology $\frac{\partial f}{\partial t} \propto \sqrt{f}\eta(t)$ predict that Fig. 1c should resemble a Gaussian distributions over short timescales, a prediction that is incompatible with what we observe [75, 76].

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.

$$\frac{\partial f}{\partial t} = sf(1-f) + \mu(1-f) + \upsilon f + \sqrt{\frac{f(1-f)}{N}} \cdot \eta(t) \tag{7}$$

To reduce the number of free parameters and remove nonlinear terms, one can examine Eq. 7 in the low frequency limit ($f \ll 1$) and obtain Langevin equations for evolution under positive ($s > 0$) and purifying selection ($s < 0$)

$$\frac{\partial f}{\partial t} = -sf + \mu + \sqrt{\frac{f}{N}} \cdot \eta(t) \tag{8a}$$

$$\frac{\partial f}{\partial t} = sf + \mu + \sqrt{\frac{f}{N}} \cdot \eta(t) \tag{8b}$$

A gamma-distributed stationary solution can be derived in the case of purifying selection ($s < 0$; [75, 77]), whereas the stationary solution of Eq. 8a 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 ($2N\mu$) [25]. Together, these selection regimes provide two forms of mutation-selection balance that predict the gamma distribution (Supporting Information).

$$f_{\text{Evo},s<0} \sim \text{Gamma}(2N\mu, 2N|s|) \tag{9a}$$

$$f_{\text{Evo},s>0} \sim \text{Gamma}(2N\mu, f_{max}^{-1}) \tag{9b}$$

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, 39], though it is unlikely responsible for the allele frequency fluctuations across hosts that I observed in this study. 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 $\frac{1}{N|s|}$ [78], 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. 9ab). 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 model is highly useful for describing the dynamics of an ensemble of populations where the initial frequency is known and of low frequency, 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 [79–82]. 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, 11]. 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; [83]). 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 predominantly influence across-host patterns of genetic diversity.

Discussion

This study demonstrated that a model of ecology that explained the dynamics of strains within a single host can be successfully applied to explain across-host patterns of diversity at individual nucleotide sites, the constituent of strains. 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 [44]) 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) using measurable parameters (i.e., zero statistical fitting) for a considerable fraction of sites across species. Prediction accuracy tended to improve among more prevalent alleles, a result that is consistent with the conceptual picture that both ecology and evolution are operating within a given species in the human gut [10]. The accuracy of prevalence predictions were 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 implies that one's level of sampling (i.e., sequencing coverage) is the primary determinant of whether or not strain structure can be detected within a healthy human host for several species [44]. A lack of genuine absences of strain structure (i.e., extinction) in healthy human hosts would subsequently imply that competitive exclusion is rare at the strain level. This is a strong claim and proving it is beyond the scope of this study. Instead it is worth noting that the claim seemingly contrasts with the observation that strains exhibit varying frequencies across hosts for several species [3], though fluctuations across hosts alone are not demonstrative of competitive exclusion. Like any model with stochasticity, fluctuations across hosts are expected under the SLM and some number of absences will inevitably arise due to the finite nature of sampling. However, it is also possible that the carrying capacity of a given strain could vary from host-to-host for several species, 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 [67]. 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 [46], 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 [20, 40, 84, 85], map onto the SLM. Models that incorporate such dynamics are capable of recapitulating the temporal dynamics of microbial communities at the species level [86], and are necessary to model the emergence of new strains and their subsequent eco-evolutionary dynamics [85], 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 [86] and it is likely that alternative models of strain-level ecology that capture the same stochastic process are equally applicable.

The conclusion that properties of a substantial fraction of alleles can be predicted across hosts using a model of ecology, as opposed to evolutionary, dynamics is of consequence to studies of 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 [12, 87–90]. 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 over time [46]. 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 [12]. The results presented in this study suggest that single-sample measures of genetic diversity that do not account for strain ecology are unlikely to be informative of evolutionary processes operating within the human gut.

The success of a prediction is often 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, 91]. 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, a range that is likely primarily driven by strain structure.

The implication of this relatively narrow observational window is 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 physical linkage between variants [5, 92–95], 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 [25, 78]. 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.

Finally, it worth commenting on the applicability of these results towards evaluating the ecological effects of environmental perturbations, namely, host-induced changes. Across-host patterns are often the consequence of within-host dynamics, so, in principle, it should be possible to use the SLM to predict across-host changes in response to perturbations. However, it is difficult to know *a priori* whether a host is in a perturbed state unless it is administrated as part of a controlled study (e.g., major diet change, drug trials, etc.). As a concrete example, one could compile data from studies on different human hosts where courses of antibiotics were administrated and metagenomic sequencing was performed over time. Knowing the inferred frequency of a strain or the frequencies of its constituent alleles at the start of the trial f_0, one could then leverage the SLM and its stationary solution to determine when statistical quantities calculated across hosts such as the mean frequency or prevalence approach their stationary values as the system relaxes away from its perturbed state (i.e., $\langle f|t, f_0\rangle \to \langle f\rangle$ or $\langle \varrho|t, \varrho_0\rangle \to \langle \varrho\rangle$) .

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 [24, 96]. 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 [29]) 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. Definitions of "species" vary across disciplines in biology. To avoid ambiguity, I opted for a direct operational definition provided by the resolution of the reference genome panel used by `MIDAS`, a definition that has been used in many studies of the human gut microbiome [3, 5, 25, 36, 39, 46, 97].

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. These parameter settings are effectively identical to the settings used in prior studies that have examined the HMP [3, 7, 39, 46]. The (non)synonymous status of sites were determined using `MIDAS` reference genomes. While members of a species can differ in genomic content and

can alter the accuracy of calling the status of a site, this pipeline was previously run on the same dataset used in this manuscript to test population genetic predictions on measures of genetic diversity using nonsynonymous and synonymous sites, finding that the empirical data matched theoretical predictions [3], implying that any ambiguity in site (non)synonymous assignment did not shape our results.

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 be reliably 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 [98]. 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 ($f \ll 1$). This effect is particularly on measures of genetic diversity, as sequencing error-induced noise can easily swamp real biological signal when statistics are calculated over a large number of sites. Studies often attempt to control for errors by restricting their analyses to sites that pass a particular threshold for the total depth of coverage and/or the coverage of the minor allele. However, the effectiveness of an arbitrary cutoff will not be identical across sites and hosts due to variance in the total depth of coverage. For this specific analysis, reliable frequency estimates are also necessary since an allele of any non-zero frequency can contribute towards measures of prevalence, so it is imperative for low frequency alleles to be estimated in a statistically justified and unbiased manner that balances sensitivity and specificity.

To account for sequencing errors, I elected to use the full maximum likelihood estimator `MAPGD` v0.4.40 [99], an unbiased estimator that accounts for the total depth of coverage and an unknown sequencing error. This estimator was chosen due its comparative high sensitivity to low frequency alleles without sacrificing its false discovery rate in real and simulated data [100], reflecting a balance between sensitivity and specificity. Using BAM and `.header` files processed from `MIDAS` output, `MAPGD` was run with a log-likelihood ratio polymorphism cutoff of 20 (`-a 20`), a choice informed by prior benchmarking studies and `MAPGD` recommendations [100]. The choice of log-likelihood ratio cutoff is unlikely to shape the results of this study, as the cutoff effectively establishes a coverage cutoff which can be incorporated into predictions derived using probability distributions that explicitly account for sampling (e.g., Eq. 4). 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.

Gamma and lognormal distributions were fit to the distribution of within-host allele frequencies across all hosts using `SciPy`. Because the distribution was rescaled using the logarithm of the data, the distributions were fit as if I was interested in the logarithm of the random variable (i.e., $\log_{10}(f)$). This detail translates to fitting the loggamma instead of the gamma and the Gaussian instead of the lognormal. AIC was calculated using custom scripts.

The `MAPGD` inference procedure makes no assumptions about the existence of strain structure. If strain structure is present in the data it will shape the distribution of allele frequencies, subsequently altering measures of genetic diversity and the predictive capacity of the SLM. To determine whether the error of the SLM was related to the existence of strains, I used an algorithm to determine whether strain structure was present for each species within each host. 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 [83] 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 [101].

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 [44, 66, 67] 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 [46]. 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 $0 < f < 1$

$$\frac{\partial f}{\partial t} = \frac{f}{\tau_i}\left(1 - \frac{f}{K_i}\right) + \sqrt{\frac{\sigma_{\tau_i}}{\tau_i}} f \cdot \eta(t) \tag{10}$$

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 $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$ [68]. 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 [68], one can formulate a partial differential equation for the probability $p(f, t)$ that an allele has frequency f at time t

$$\frac{\partial p(f,t)}{\partial t} = -\frac{\partial}{\partial f}\left[\left(\frac{f}{\tau_i}\left(1 - \frac{f}{K_i}\right)\right) p(f,t)\right] + \frac{\sigma_{\tau_i}}{2\tau_i}\frac{\partial^2}{\partial f^2}\left(f^2 p(f,t)\right) \tag{11}$$

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

$$p(f) = \frac{1}{\Gamma\left(2\sigma_{\tau_i}^{-1} - 1\right)}\left(\frac{2}{K_i \sigma_{\tau_i}}\right)^{2\sigma_{\tau_i}^{-1} - 1} \exp\left[-\frac{2}{K_i \sigma_{\tau_i}} f\right] f^{2\sigma_{\tau_i}^{-1} - 2} \tag{12}$$

This distribution, known as the abundance fluctuation distribution in macroecology [44], is a gamma distribution with the following mean and squared coefficient of variation

$$\langle f \rangle = K_i\left(1 - \frac{\sigma_{\tau_i}}{2}\right) \tag{13a}$$

$$\frac{\langle f^2 \rangle - \langle f \rangle^2}{\langle f \rangle^2} = \frac{\sigma_{\tau_i}}{2 - \sigma_{\tau_i}} \tag{13b}$$

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

$$p\left(f | \beta, \beta/\bar{f}\right) = \frac{1}{\Gamma(\beta)}\left(\frac{\beta}{\bar{f}}\right)^{\beta} \exp\left[-f\frac{\beta}{\bar{f}}\right] f^{\beta - 1} \tag{14}$$

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 [44], 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

$$\Pr[A|D, f] = \binom{D}{A} f^A (1-f)^{D-A} \tag{15}$$

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

$$\Pr[A|D, f] = \frac{(D \cdot f)^A}{A!} e^{-D \cdot f} \tag{16}$$

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

$$\begin{aligned}
\Pr[A|D, \beta, \beta/\bar{f}] &= \int_0^1 \Pr[A|D, f] \cdot p(f|\beta, \beta/\bar{f})\, df && \text{(17a)} \\
&= \frac{D^A}{A!\Gamma(\beta)\frac{\beta}{\bar{f}}^{\beta}} \int_0^1 f^{A+\beta-1} e^{-f(D+\frac{\beta}{\bar{f}}^{-1})}\, df && \text{(17b)} \\
&= \frac{\Gamma(\beta + A)}{A!\Gamma(\beta)} \left(\frac{\bar{f}D}{\beta + \bar{f}D}\right)^A \left(\frac{\beta}{\beta + \bar{f}D}\right)^{\beta} && \text{(17c)}
\end{aligned}$$

The distribution $\Pr[A|D, \beta, \beta/\bar{f}]$ now represents a gamma distribution that explicitly accounts for sampling. 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 [44, 70]

$$\Pr[0|D, \beta, \beta/\bar{f}] = \left(1 + D \cdot \frac{\beta}{\bar{f}}\right)^{-\beta} \tag{18}$$

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

$$\langle \varrho \rangle = \frac{1}{M} \sum_{m=1}^{M} (1 - \Pr[0|D_m, \beta, \beta/\bar{f}]) \tag{19}$$

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.

$$\hat{\varrho} = \frac{1}{M} \sum_{m=1}^{M} (1 - \delta_{f_m,0}) = 1 - \frac{1}{M} \sum_{m=1}^{M} \delta_{f_m,0} \tag{20}$$

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

$$\varepsilon = \left| \frac{\text{Obs.} - \text{Pred.}}{\text{Obs.}} \right| \tag{21}$$

I 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 [24, 102] (URLs: `https://portal.hmpdacc.org/`, `https://aws.amazon.com/datasets/human-microbiome-project/`). `MIDAS` output was processed using publicly available code [3]. Processed data is available on Zenodo: `https://doi.org/10.5281/zenodo.6793770`. All code written for this study is available on GitHub: `https://github.com/wrshoemaker/StrainPrevalence`.

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, D. Reyes-González, and N.I. Wisnoski for their feedback on the manuscript. Finally, thanks to M.S. Ackerman for their assistance with `MAPGD`. 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, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host & Microbe. 2019 May;25(5):656–667.e8.

2. Ghosh OM, Good BH. Emergent evolutionary forces in spatial models of luminal growth in the human gut microbiota; 2021. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. Available from: https://www.biorxiv.org/content/10.1101/2021.07.15.452569v1.

3. Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLOS Biology. 2019 Jan;17(1):e3000102. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000102.

4. Yaffe E, Relman DA. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. Nature Microbiology. 2020 Feb;5(2):343–353.

5. Roodgar M, Good BH, Garud NR, Martis S, Avula M, Zhou W, et al. Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. Genome Research. 2021 Aug;31(8):1433–1446. 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. Available from: https://genome.cshlp.org/content/31/8/1433.

6. Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat 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. 2018 Mar;84(6):e02377–17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5835743/.

7. Chen DW, Garud NR. Rapid evolution and strain turnover in the infant gut microbiome; 2021. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. Available from: https://www.biorxiv.org/content/10.1101/2021.09.26.461856v1.

8. Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell. 2021 Apr;184(8):2053–2067.e18. Available from: https://www.sciencedirect.com/science/article/pii/S0092867421002415.

9. Dapa T, Wong DP, Vasquez KS, Xavier KB, Huang KC, Good BH. Within-host evolution of the gut microbiome. Current Opinion in Microbiology. 2023 Feb;71:102258. Available from: https://www.sciencedirect.com/science/article/pii/S1369527422001424.

10. Good BH, Hallatschek O. Effective models and the search for quantitative principles in microbial evolution. Current Opinion in Microbiology. 2018 Oct;45:203–212.

11. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research. 2017 Apr;27(4):626–638.

12. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013 Jan;493(7430):45–50. 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. Available from: https://www.nature.com/articles/nature11711.

13. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science (New York, NY). 2013 Jul;341(6141):1237439.

14. Moeller AH. Metagenomic signatures of balancing selection in the human gut. Molecular Ecology. 2023;32(10):2582–2591. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.16474. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16474.

15. Valles-Colomer M, Blanco-Míguez A, Manghi P, Asnicar F, Dubois L, Golzato D, et al. The person-to-person transmission landscape of the gut and oral microbiomes. Nature. 2023 Feb;614(7946):125–135. Number: 7946 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41586-022-05620-1.

16. Turroni F, Foroni E, Pizzetti P, Giubellini V, Ribbera A, Merusi P, et al. Exploring the Diversity of the Bifidobacterial Population in the Human Intestinal Tract. Applied and Environmental Microbiology. 2009 Mar;75(6):1534–1545. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/full/10.1128/AEM.02216-08.

17. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nature Microbiology. 2019 Mar;4(3):470–479. Number: 3 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41564-018-0321-5.

18. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nature Biotechnology. 2019 Feb;37(2):186–192. Number: 2 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41587-018-0009-7.

19. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host & Microbe. 2018 Jul;24(1):133–145.e5.

20. Goyal A, Bittleston LS, Leventhal GE, Lu L, Cordero OX. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. eLife. 2022 Feb;11:e74987. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.74987.

21. Wang Z, Fridman Y, Maslov S, Goyal A. Fine-scale diversity of microbial communities due to satellite niches in boom-and-bust environments. bioRxiv;

2022. Pages: 2022.05.26.493560 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2022.05.26.493560v2.

22. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011 May;473(7346):174–180. 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. Available from: https://www.nature.com/articles/nature09944.

23. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nature Reviews Microbiology. 2008 Oct;6(10):776–788. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrmicro1978.

24. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017 Oct;550(7674):61–66.

25. Good BH. Linkage disequilibrium between rare mutations. Genetics. 2022 Apr;220(4):iyac004. Available from: https://doi.org/10.1093/genetics/iyac004.

26. Baumgartner M, Bayer F, Pfrunder-Cardozo KR, Buckling A, Hall AR. Resident microbial communities inhibit growth and antibiotic-resistance evolution of Escherichia coli in human gut microbiome samples. PLOS Biology. 2020 Apr;18(4):e3000465. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000465.

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

28. Karcher N, Pasolli E, Asnicar F, Huang KD, Tett A, Manara S, et al. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biology. 2020 Jun;21(1):138. Available from: https://doi.org/10.1186/s13059-020-02042-y.

29. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research. 2016 Nov;26(11):1612–1625.

30. Niu J, Xu L, Qian Y, Sun Z, Yu D, Huang J, et al. Evolution of the Gut Microbiome in Early Childhood: A Cross-Sectional Study of Chinese Children. Frontiers in Microbiology. 2020;11. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.00439.

31. Kundu P, Blacher E, Elinav E, Pettersson S. Our Gut Microbiome: The Evolving Inner Self. Cell. 2017 Dec;171(7):1481–1493. Available from: https://www.sciencedirect.com/science/article/pii/S0092867417313715.

32. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012 Jun;486(7402):222–227. 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. Available from: https://www.nature.com/articles/nature11053.

33. Tierney BT, Yang Z, Luber JM, Beaudin M, Wibowo MC, Baek C, et al. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. Cell Host & Microbe. 2019 Aug;26(2):283–295.e8. Available from: https://www.sciencedirect.com/science/article/pii/S193131281930352X.

34. Prosser JI, Bohannan BJM, Curtis TP, Ellis RJ, Firestone MK, Freckleton RP, et al. The role of ecological theory in microbial ecology. Nature Reviews Microbiology. 2007 May;5(5):384–392. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrmicro1643.

35. Marquet PA, Allen AP, Brown JH, Dunne JA, Enquist BJ, Gillooly JF, et al. On Theory in Ecology. BioScience. 2014 Aug;64(8):701–710. Available from: https://doi.org/10.1093/biosci/biu098.

36. Good BH, Rosenfeld LB. Eco-evolutionary feedbacks in the human gut microbiome. bioRxiv; 2022. Section: New Results Type: article. Available from: https://www.biorxiv.org/content/10.1101/2022.01.26.477953v1.

37. Garud NR, Pollard KS. Population Genetics in the Human Microbiome. Trends in Genetics. 2020 Jan;36(1):53–67. Available from: https://www.sciencedirect.com/science/article/pii/S0168952519302215.

38. Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nature Medicine. 2019 Sep;25(9):1442–1452.

39. Shoemaker WR, Chen D, Garud NR. Comparative Population Genetics in the Human Gut Microbiome. Genome Biology and Evolution. 2021 May;(evab116). Available from: https://doi.org/10.1093/gbe/evab116.

40. Cui W, Marsland R, Mehta P. Diverse communities behave like typical random ecosystems. Physical Review E. 2021 Sep;104(3):034416. Publisher: American Physical Society. Available from: https://link.aps.org/doi/10.1103/PhysRevE.104.034416.

41. 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 Mar;2018:033406.

42. Descheemaeker L, de Buyl S. Stochastic logistic models reproduce experimental time series of microbial communities. eLife. 2020 Jul;9:e55650. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.55650.

43. Shoemaker WR, Locey KJ, Lennon JT. A macroecological theory of microbial biodiversity. Nature Ecology & Evolution. 2017 Apr;1(5):1–6. 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. Available from: https://www.nature.com/articles/s41559-017-0107.

44. Grilli J. Macroecological laws describe variation and diversity in microbial communities. Nature Communications. 2020 Sep;11(1):4743. 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. Available from: https://www.nature.com/articles/s41467-020-18529-y.

45. Ji BW, Sheth RU, Dixit PD, Tchourine K, Vitkup D. Macroecological dynamics of gut microbiota. Nature Microbiology. 2020 May;5(5):768–775. 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. Available from: https://www.nature.com/articles/s41564-020-0685-1.

46. Wolff R, Shoemaker W, Garud N. Ecological Stability Emerges at the Level of Strains in the Human Gut Microbiome. mBio. 2023 Feb;0(0):e02502–22. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/mbio.02502-22.

47. Shoemaker WR, Grilli J. Macroecological patterns in coarse-grained microbial communities. bioRxiv; 2023. Pages: 2023.03.02.530804 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2023.03.02.530804v1.

48. Shade A, Dunn RR, Blowes SA, Keil P, Bohannan BJM, Herrmann M, et al. Macroecology to Unite All Life, Large and Small. Trends in Ecology & Evolution. 2018 Oct;33(10):731–744.

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

50. Phillips R. Theory in Biology: Figure 1 or Figure 7? Trends in Cell Biology. 2015 Dec;25(12):723–729. Publisher: Elsevier. Available from: https://www.cell.com/trends/cell-biology/abstract/S0962-8924(15)00194-4.

51. Bhattacharjee SM, Seno F. A measure of data collapse for scaling. Journal of Physics A: Mathematical and General. 2001 Aug;34(33):6375–6380. Publisher: IOP Publishing. Available from: https://doi.org/10.1088/0305-4470/34/33/302.

52. Stanley HE. Scaling, universality, and renormalization: Three pillars of modern critical phenomena. Reviews of Modern Physics. 1999 Mar;71(2):S358–S366. Publisher: American Physical Society. Available from: https://link.aps.org/doi/10.1103/RevModPhys.71.S358.

53. Theys K, Feder AF, Gelbart M, Hartl M, Stern A, Pennings PS. Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. PLOS Genetics. 2018 Jun;14(6):e1007420. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007420.

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

55. Marquet PA, Quiñones RA, Abades S, Labra F, Tognelli M, Arim M, et al. Scaling and power-laws in ecological systems. Journal of Experimental Biology. 2005 May;208(9):1749–1769. Available from: https://journals.biologists.com/jeb/article/208/9/1749/9378/Scaling-and-power-laws-in-ecological-systems.

56. Ramsayer J, Fellous S, Cohen JE, Hochberg ME. Taylor's Law holds in experimental bacterial populations but competition does not influence the slope. Biology Letters. 2012 Apr;8(2):316–319. Publisher: Royal Society. Available from: https://royalsocietypublishing.org/doi/10.1098/rsbl.2011.0895.

57. Kendal WS, Jørgensen B. Taylor's power law and fluctuation scaling explained by a central-limit-like convergence. Physical Review E. 2011 Jun;83(6):066115. Available from: https://link.aps.org/doi/10.1103/PhysRevE.83.066115.

58. Kendal WS. A scale invariant clustering of genes on human chromosome 7. BMC Evolutionary Biology. 2004:10.

59. Kendal WS. An Exponential Dispersion Model for the Distribution of Human Single Nucleotide Polymorphisms. Molecular Biology and Evolution. 2003 Apr;20(4):579–590. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msg057.

60. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, 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. 2010 Apr;107(16):7527–7532.

61. Bhatia R, Davis C. A Better Bound on the Variance. The American Mathematical Monthly. 2000 Apr;107(4):353–357. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00029890.2000.12005203. Available from: https://doi.org/10.1080/00029890.2000.12005203.

62. Cohen JE, Xu M. Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. Proceedings of the National Academy of Sciences. 2015 Jun;112(25):7749–7754. Publisher: National Academy of Sciences Section: Biological Sciences. Available from: https://www.pnas.org/content/112/25/7749.

63. Giometto A, Formentin M, Rinaldo A, Cohen JE, Maritan A. Sample and population exponents of generalized Taylor's law. Proceedings of the National Academy of Sciences. 2015 Jun;112(25):7755–7760. Publisher: National Academy of Sciences Section: Biological Sciences. Available from: https://www.pnas.org/content/112/25/7755.

64. Xiao X, Locey KJ, White EP. A Process-Independent Explanation for the General Form of Taylor's Law. The American Naturalist. 2015 Aug;186(2):E51–E60. Publisher: The University of Chicago Press. Available from: https://www.journals.uchicago.edu/doi/full/10.1086/682050.

65. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. The Journal of Chemical Physics. 2015 Jul;143(1):010901. Publisher: American Institute of Physics. Available from: https://aip.scitation.org/doi/10.1063/1.4923066.

66. Zaoli S, Grilli J. The stochastic logistic model with correlated carrying capacities reproduces beta-diversity metrics of microbial communities. PLOS Computational Biology. 2022 Apr;18(4):e1010043. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010043.

67. Zaoli S, Grilli J. A macroecological description of alternative stable states reproduces intra- and inter-host variability of gut microbiome. Science Advances. 2021;7(43):eabj2882. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/sciadv.abj2882.

68. Gardiner CW. Stochastic methods: a handbook for the natural and social sciences. 4th ed. No. 13 in Springer series in synergetics. Berlin Heidelberg: Springer; 2009.

69. Engen S, Lande R. Population Dynamic Models Generating Species Abundance Distributions of the Gamma Type. Journal of Theoretical Biology. 1996 Feb;178(3):325–331. Available from: https://www.sciencedirect.com/science/article/pii/S0022519396900284.

70. Shoemaker WR, Lennon JT. Predicting Parallelism and Quantifying Divergence in Microbial Evolution Experiments. mSphere. 2022 Feb. Publisher: American Society for Microbiology 1752 N St., N.W., Washington, DC. Available from: https://journals.asm.org/doi/abs/10.1128/msphere.00672-21.

71. Gaston KJ, Blackburn TM, Greenwood JJD, Gregory RD, Quinn RM, Lawton JH. Abundance–occupancy relationships. Journal of Applied Ecology. 2000;37(s1):39–59. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2664.2000.00485.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2664.2000.00485.x.

72. Sloan WT, Woodcock S, Lunn M, Head IM, Curtis TP. Modeling Taxa-Abundance Distributions in Microbial Communities using Environmental Sequence Data. Microbial Ecology. 2007 Apr;53(3):443–455. Available from: https://doi.org/10.1007/s00248-006-9141-x.

73. Burns AR, Stephens WZ, Stagaman K, Wong S, Rawls JF, Guillemin K, et al. Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. The ISME Journal. 2016 Mar;10(3):655–664. 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. Available from: https://www.nature.com/articles/ismej2015142.

74. Cohen JE. 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. 2020 Mar;13(1):1–5. Available from: https://doi.org/10.1007/s12080-019-00445-7.

75. Ewens WJ. Mathematical population genetics. Springer; 2010. OCLC: 778838329.

76. Hubbell SP. The unified neutral theory of biodiversity and biogeography. No. 32 in Monographs in population biology. Princeton: Princeton University Press; 2001.

77. Nei M. The frequency distribution of lethal chromosomes in finite populations. Proceedings of the National Academy of Sciences of the United States of America. 1968 Jun;60(2):517–524. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC225078/.

78. Cvijović I, Good BH, Desai MM. The Effect of Strong Purifying Selection on Genetic Diversity. Genetics. 2018 Aug;209(4):1235–1278. Available from: https://doi.org/10.1534/genetics.118.301058.

79. Yan Y, Nguyen LH, Franzosa EA, Huttenhower C. Strain-level epidemiology of microbial communities and the human microbiome. Genome Medicine. 2020 Aug;12:71. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7427293/.

80. Dijkshoorn L, Ursing BM, Ursing JB. Strain, clone and species: comments on three basic concepts of bacteriology. Journal of Medical Microbiology. 2000 May;49(5):397–401.

81. Zhu A, Sunagawa S, Mende DR, Bork P. Inter-individual differences in the gene content of human gut bacterial species. Genome Biology. 2015 Apr;16(1):82. Available from: https://doi.org/10.1186/s13059-015-0646-9.

82. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends in Microbiology. 2014 May;22(5):235–247.

83. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host & Microbe. 2018 Feb;23(2):229–240.e5. Publisher: Elsevier. Available from: https://www.cell.com/cell-host-microbe/abstract/S1931-3128(18)30038-6.

84. Chesson P. MacArthur's consumer-resource model. Theoretical Population Biology. 1990 Feb;37(1):26–38. Available from: https://www.sciencedirect.com/science/article/pii/004058099090025Q.

85. Good BH, 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. 2018 Oct;115(44):E10407–E10416. Publisher: National Academy of Sciences Section: PNAS Plus. Available from: https://www.pnas.org/content/115/44/E10407.

86. Ho PY, Good BH, Huang KC. Competition for fluctuating resources reproduces statistics of species abundance over time across wide-ranging microbiotas. eLife. 2022 Apr;11:e75168. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.75168.

87. Madi N, Chen D, Wolff R, Shapiro BJ, Garud NR. Community diversity is associated with intra-species genetic diversity and gene loss in the human gut microbiome. bioRxiv; 2022. Section: New Results Type: article. Available from: https://www.biorxiv.org/content/10.1101/2022.03.08.483496v1.

88. Li J, Rettedal EA, van der Helm E, Ellabaan M, Panagiotou G, Sommer MOA. Antibiotic Treatment Drives the Diversification of the Human Gut Resistome. Genomics, Proteomics & Bioinformatics. 2019 Feb;17(1):39–51. Available from: https://www.sciencedirect.com/science/article/pii/S1672022919300609.

89. N'Guessan A, Brito IL, Serohijos AWR, Shapiro BJ. Mobile Gene Sequence Evolution within Individual Human Gut Microbiomes Is Better Explained by Gene-Specific Than Host-Specific Selective Pressures. Genome Biology and Evolution. 2021 Aug;13(8):evab142. Available from: https://doi.org/10.1093/gbe/evab142.

90. Simonet C, McNally L. Kin selection explains the evolution of cooperation in the gut microbiota. Proceedings of the National Academy of Sciences. 2021 Feb;118(6):e2016046118. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/10.1073/pnas.2016046118.

91. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLOS Biology. 2016 Aug;14(8):e1002533. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002533.

92. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nature Biotechnology. 2016 Jan;34(1):64–69. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Metagenomics Subject_term_id: metagenomics. Available from: https://www.nature.com/articles/nbt.3416.

93. Zlitni S, Bishara A, Moss EL, Tkachenko E, Kang JB, Culver RN, et al. Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. Genome Medicine. 2020 May;12(1):50. Available from: https://doi.org/10.1186/s13073-020-00747-0.

94. DeMaere MZ, Darling AE. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biology. 2019 Feb;20(1):46. Available from: https://doi.org/10.1186/s13059-019-1643-1.

95. Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo CC, et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions; 2017. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. Available from: https://www.biorxiv.org/content/10.1101/198713v1.

96. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature. 2012 Jun;486(7402):215–221. 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. Available from: https://www.nature.com/articles/nature11209.

97. Liu Z, Good BH. Dynamics of bacterial recombination in the human gut microbiome. bioRxiv; 2022. Pages: 2022.08.24.505183 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2022.08.24.505183v1.

98. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 02;10(2). Giab008. Available from: https://doi.org/10.1093/gigascience/giab008.

99. Lynch M, Bost D, Wilson S, Maruki T, Harrison S. Population-Genetic Inference from Pooled-Sequencing Data. Genome Biology and Evolution. 2014 Apr;6(5):1210–1218. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4040993/.

100. Guirao-Rico S, González J. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. Molecular Ecology Resources. 2021;21(4):1216–1229. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13343. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13343.

101. Xiao X, Thibault K, Harris DJ, Baldridge E, White E. weecology/macroecotools: v0.4.0. Zenodo; 2016. Available from: https://zenodo.org/record/166721.

102. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun;486(7402):207–214. Number: 7402 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature11234.

Figures

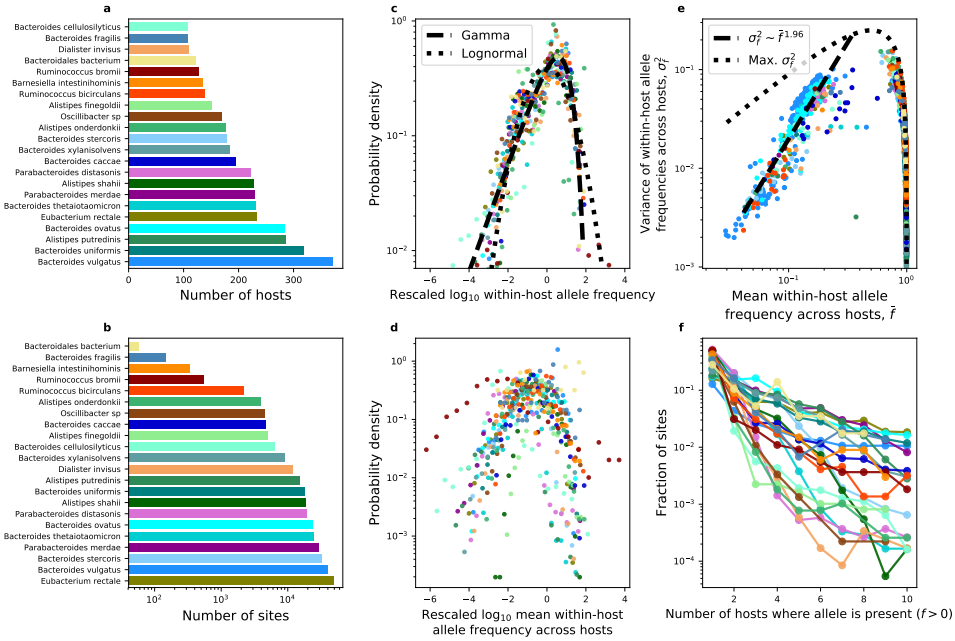

Figure 1. Distributions of genetic diversity exhibit similar statistical forms across phylogenetically distant species in the human gut. a,b) Similarity in patterns of genetic diversity was evaluated for sites obtained from the 22 most prevalent bacterial species. **c)** The distribution of within-host allele frequencies across all hosts as well as **d)** the distribution of mean allele frequencies were rescaled to determine whether they exhibited similar forms, specifically by rescaling their logarithm using the standard score (i.e., z-score). In **c**, statistical fits of a gamma (the distribution predicted by the SLM) and a lognormal (a point of comparison) are illustrated as black lines. To limit the effect of the bounded nature of allele frequencies on the distribution, mean frequencies containing observations of $f = 1$ were excluded from subplot **d**. **e)** The relationship between statistical moments of within-host allele frequencies was consistent across species, as there was a strong linear relationship between the mean frequency of an allele and its 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, alleles with non-zero values of f in $< 35\%$ of hosts were excluded. **f)** Finally, the fraction of sites harboring alleles present in a given number of hosts decreased in a similar manner across species. All sites in this analysis are synonymous, identical analyses were performed on alleles at nonsynonymous sites (Fig. S1). Species within the same genus were assigned the same primary or secondary color with different degrees of saturation.

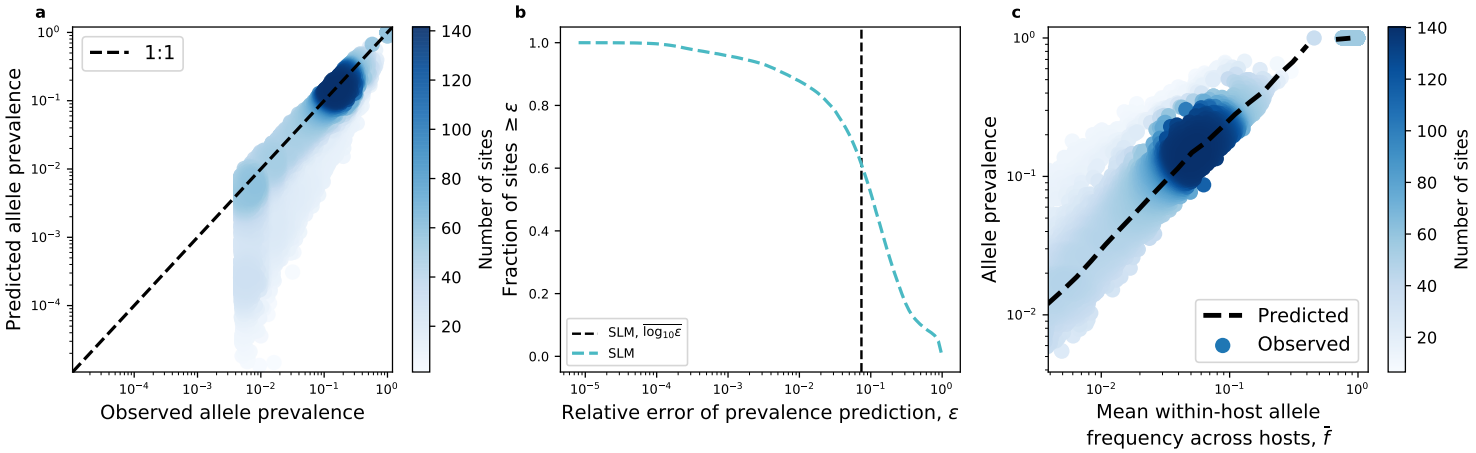

Figure 2. The SLM successfully predicts genetic patterns for prevalent alleles. a) Predicted values of prevalence were obtained (Eq. 6) and matched with their observed values, where the data generally fell on the one-to-one line (dashed black line) for sites with a prevalence $\gtrsim 0.01$. **b)** The pattern of the SLM performing better for higher values of prevalence was illustrated by quantifying the relative error of the prevalence predictions. The mean of the logarithm of the relative error of the SLM over all sites ($\overline{\log_{10}\varepsilon}$, dashed black line) was ~ 0.1. **c)** The contingency of the SLM's success was illustrated by examining the relationship between the mean frequency of an allele across hosts ($\bar{f}$) and its prevalence. The predictions of the SLM (not a statistical fit) succeed for high mean frequency alleles (dashed black line). All analyses here were performed on alleles at synonymous sites using the common commensal gut species *B. vulgatus*. The color of each datapoint is proportionate to the number of sites. Visualizations of the predictions in this plot for all species for nonsynonymous and synonymous sites can be found in the Supplemental Text.

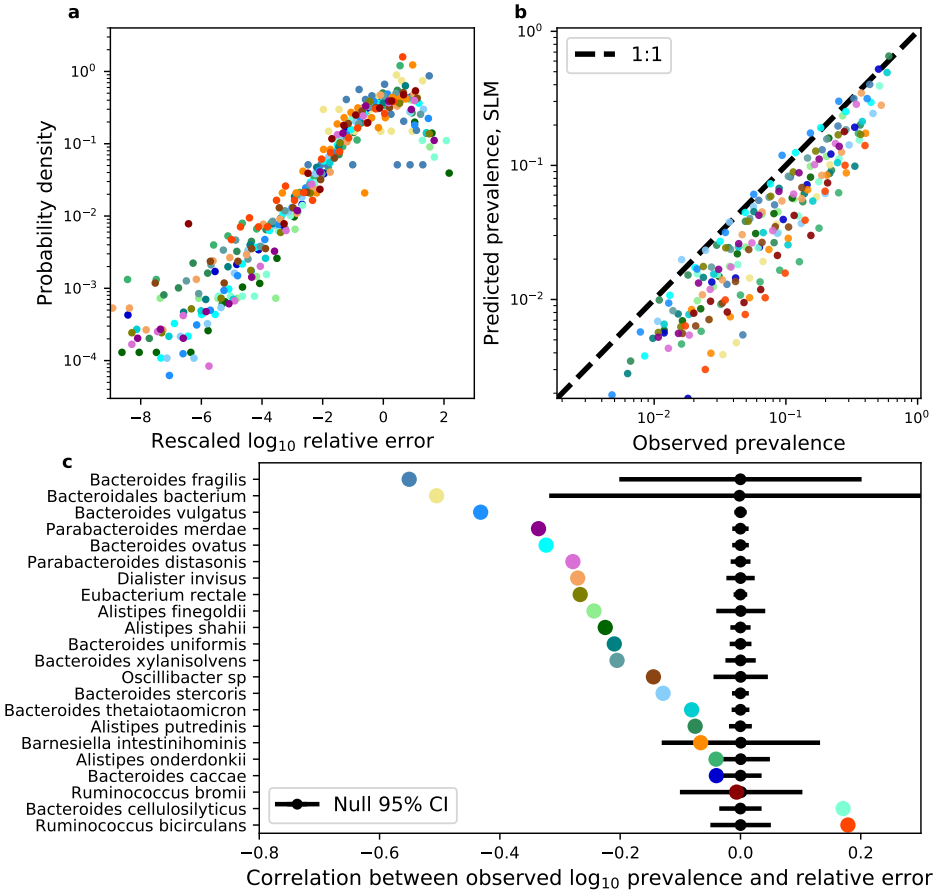

Figure 3. The SLM succeeds and fails in a consistent manner across phylogenetically distant species. a) Distributions of logarithmic relative errors of prevalence obtained from Eq. 6 and 21 were rescaled by their mean and standard deviation to illustrate their similarity across species. **b)** Binning observed prevalences reveals how the predicted values tend to follow a one-to-one relationship (dashed black line) across species, with variation among species. **c)** By calculating the correlation between $\mathbf{log}_{10}$-transformed observed prevalence and the relative error of our predictions, we find that the correlation is negative for the majority of species. A permutation test confirms that these negative correlations are significant (95% CIs as black lines) 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. S11).

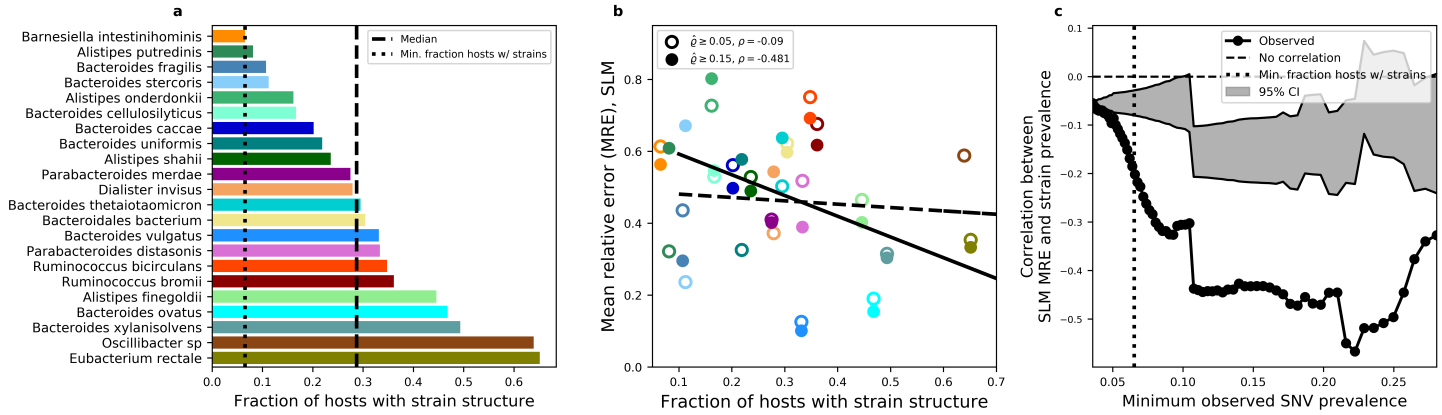

Figure 4. The presence of strain structure explains the accuracy of the SLM. **a)** The presence or absence of strain structure was inferred from the distribution of allele frequencies for each species within each host using `StrainFinder`, providing an estimate of the fraction of hosts that harbor strain structure. **b)** This per-species estimate of strain structure can be compared to the mean relative error of allelic prevalence predictions obtained using the SLM (Eq. 6) to determine whether the success of the SLM is correlated with the existence of strains. By examining this relationship when sites with rare alleles (i.e., low prevalence) are included ($\hat{\varrho} \geq 0.05$: dashed line) and excluded ($\hat{\varrho} \geq 0.15$: solid line), we see that the correlation is stronger for alleles with a high prevalence threshold. **c)** This trend can be systematically evaluated by calculating the correlation across a range of prevalence values (black dots). A permutation test establishes 95% confidence intervals of the null (grey window).