

Supplementary materials

List of Tables

S1	Host sampling metadata	2
S2	Reference genome length by species	3
S3	Number of strains per species: <i>am</i>	4
S4	Number of strains per species: <i>ao</i>	4
S5	Number of strains per species: <i>an</i>	5
S6	Number of strains per species: <i>ae</i>	5

List of Figures

S1	Oligo-colonization	6
S2	Read and sampling metadata	7
S3	Nucleotide diversity π	8
S4	F_{ST}' by species and host, colored by stationarity	9
S5	F_{ST}' by species and host	10
S6	Polarization when three strains are present	16
S7	Taylor's Law: multinomial compositional sampling	18
S8	Taylor's Law: variable intensity multinomial compositional sampling	19

Supplementary text

1	Data	11
2	Metagenomic pipeline	11
	2.1 Species abundance	11
	2.2 Building reference panels	12
	2.3 CNV content	12
	2.4 SNV content	12
3	Detecing strains	13
	3.1 Oligo-colonization	13
	3.2 Strain inference	14
4	SLM	15
5	Goodness of fit test	16
6	Taylor's Law	17

Supplementary Figures and Tables

Host alias	Samples	Unique sample days	Sampling period (days)	Mean samples per day	Median days between samples	Max days between samples
<i>am</i>	206	193	539	.35	1	51
<i>ao</i>	74	71	144	.49	1	14
<i>an</i>	63	61	375	.16	2	58
<i>ae</i>	59	58	375	.29	2	41

Table S1: Host sampling metadata for the four hosts considered in the study. When sample metadata indicated that multiple samples had been collected on the same day, or were technical replicates of one another, samples were merged by concatenating fastq files. Hence, the total number of samples reported (**Samples**) differed from the number of samples analyzed (**Unique sample days**). While gaps between successive successive samples were as long as 58 days (*an*), the median time between samples was either one or two days.

Species name	Genome length (bp)
<i>Alistipes onderdonkii</i>	3,864,643
<i>Alistipes putredinis</i>	2,550,678
<i>Bacteroides cellulosilyticus</i>	6,870,144
<i>Bacteroides clarus</i>	3,746,690
<i>Bacteroides fragilis</i>	5,723,061
<i>Bacteroides intestinalis</i>	6,052,596
<i>Phocaeicola massiliensis</i>	4,443,004
<i>Bacteroides ovatus</i>	7,880,760
<i>Bacteroides stercoris</i>	4,009,829
<i>Bacteroides thetaiotaomicron</i>	6,855,195
<i>Bacteroides uniformis</i>	4,719,097
<i>Phocaeicola vulgatus</i>	5,163,189
<i>Bacteroides xylanisolvens</i>	5,986,762
<i>Barnesiella intestinihominis</i>	3,433,706
<i>Bifidobacterium adolescentis</i>	2,304,613
<i>Dialister invisus</i>	1,895,960
<i>Lachnospira eligens</i>	2,831,389
<i>Eubacterium rectale</i>	3,344,951
<i>Faecalibacterium prausnitzii</i> (57453)	3,127,383
<i>Faecalibacterium prausnitzii</i> (61481)	3,090,349
<i>Faecalibacterium prausnitzii</i> (62201)	3,321,367
<i>Parabacteroides distasonis</i>	4,887,173
<i>Parabacteroides merdae</i>	4,434,377
<i>Paraprevotella clara</i>	4,187,245
<i>Phascolarctobacterium sp</i>	2,369,100
<i>Ruminococcus bicirculans</i>	2,968,500
<i>Ruminococcus bromii</i>	2,249,085
<i>Sutterella wadsworthensis</i>	2,949,098

Table S2: Reference genome lengths used by MIDAS in aligning shotgun reads for the purpose of calling SNVs.

<i>am</i>	
Species name	Number of strains
<i>Bacteroides fragilis</i>	1
<i>Bacteroides ovatus</i>	2
<i>Bacteroides stercoris</i>	1
<i>Bacteroides uniformis</i>	1
<i>Phocaeicola vulgatus</i>	3
<i>Bacteroides xylanisolvens</i>	2
<i>Barnesiella intestinihominis</i>	1
<i>Eubacterium rectale</i>	2
<i>Faecalibacterium prausnitzii</i> (57453)	1
<i>Phascolarctobacterium</i> sp.	1
<i>Ruminococcus bicirculans</i>	1
<i>Ruminococcus bromii</i>	1

Table S3: Number of strains inferred for each species analyzed in host *am*.

<i>ao</i>	
Species name	Number of strains
<i>Bacteroides fragilis</i>	1
<i>Bacteroides ovatus</i>	2
<i>Bacteroides uniformis</i>	1
<i>Bacteroides xylanisolvens</i>	2
<i>Bifidobacterium adolescentis</i>	1
<i>Dialister invisus</i>	1
<i>Lachnospira eligens</i>	1
<i>Eubacterium rectale</i>	1
<i>Faecalibacterium prausnitzii</i> (61481)	2
<i>Faecalibacterium prausnitzii</i> (62201)	1
<i>Parabacteroides distasonis</i>	1
<i>Ruminococcus bicirculans</i>	1

Table S4: Number of strains inferred for each species analyzed in host *ao*.

<i>an</i>	
Species name	Number of strains
<i>Alistipes putredinis</i>	1
<i>Bacteroides clarus</i>	1
<i>Phocaeicola massiliensis</i>	1
<i>Bacteroides uniformis</i>	2
<i>Phocaeicola vulgatus</i>	3
<i>Bacteroides xylanisolvens</i>	1
<i>Barnesiella intestinhominis</i>	1
<i>Eubacterium rectale</i>	2
<i>Paraprevotella clara</i>	1
<i>Ruminococcus bromii</i>	1
<i>Sutterella wadsworthensis</i>	2

Table S5: Number of strains inferred for each species analyzed in host *an*.

<i>ae</i>	
Species name	Number of strains
<i>Alistipes onderdonkii</i>	2
<i>Alistipes putredinis</i>	1
<i>Bacteroides cellulosilyticus</i>	1
<i>Phocaeicola massiliensis</i>	2
<i>Bacteroides ovatus</i>	2
<i>Bacteroides thetaiotaomicron</i>	1
<i>Bacteroides uniformis</i>	1
<i>Phocaeicola vulgatus</i>	1
<i>Eubacterium rectale</i>	2

Table S6: Number of strains inferred for each species analyzed in host *ae*.

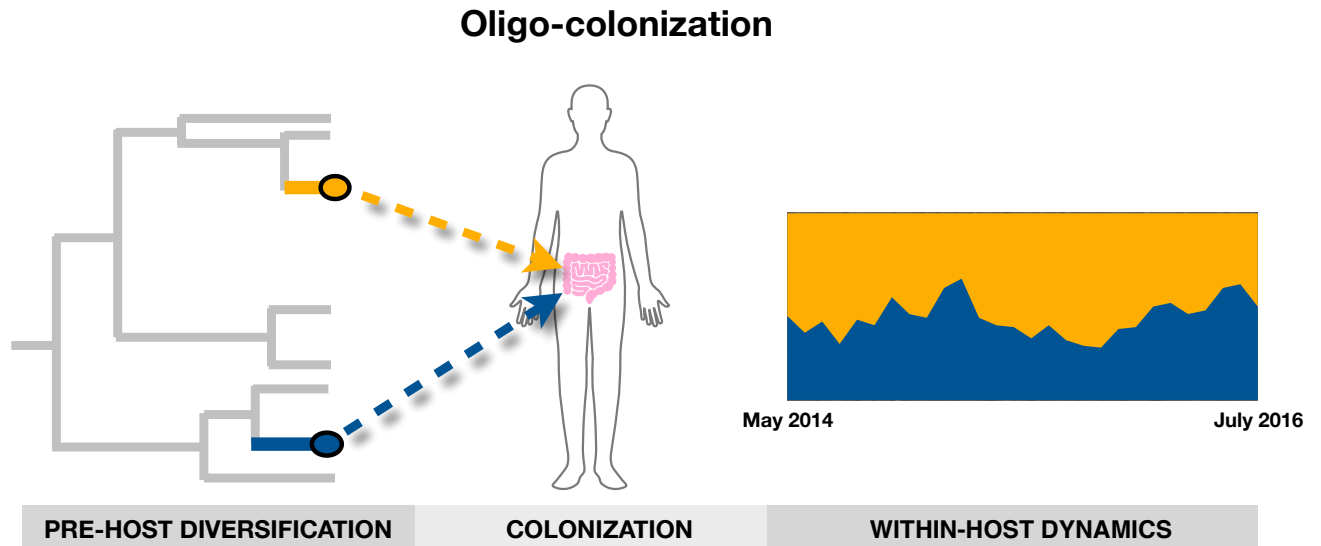


Figure S1: A range of human gut microbiome studies have shown that only a small handful of strains belonging to a species are present within a host at any one time, a phenomenon known as oligo-colonization [5, 8, 9, 10, 11, 14]. Strains diversify before colonizing the host, typically accumulating $\mathcal{O}(10^3) - \mathcal{O}(10^4)$ mutations in their shared core genome. In this study, we explore the within-host dynamics of oligo-colonization both in cases where a single strain belonging to a species is present and when multiple strains are present. We find that a large majority of strains stably colonize hosts for periods of months to years, while a smaller percentage show highly non-stationary dynamics. For further discussion of oligo-colonization, Supplementary section 3.1.

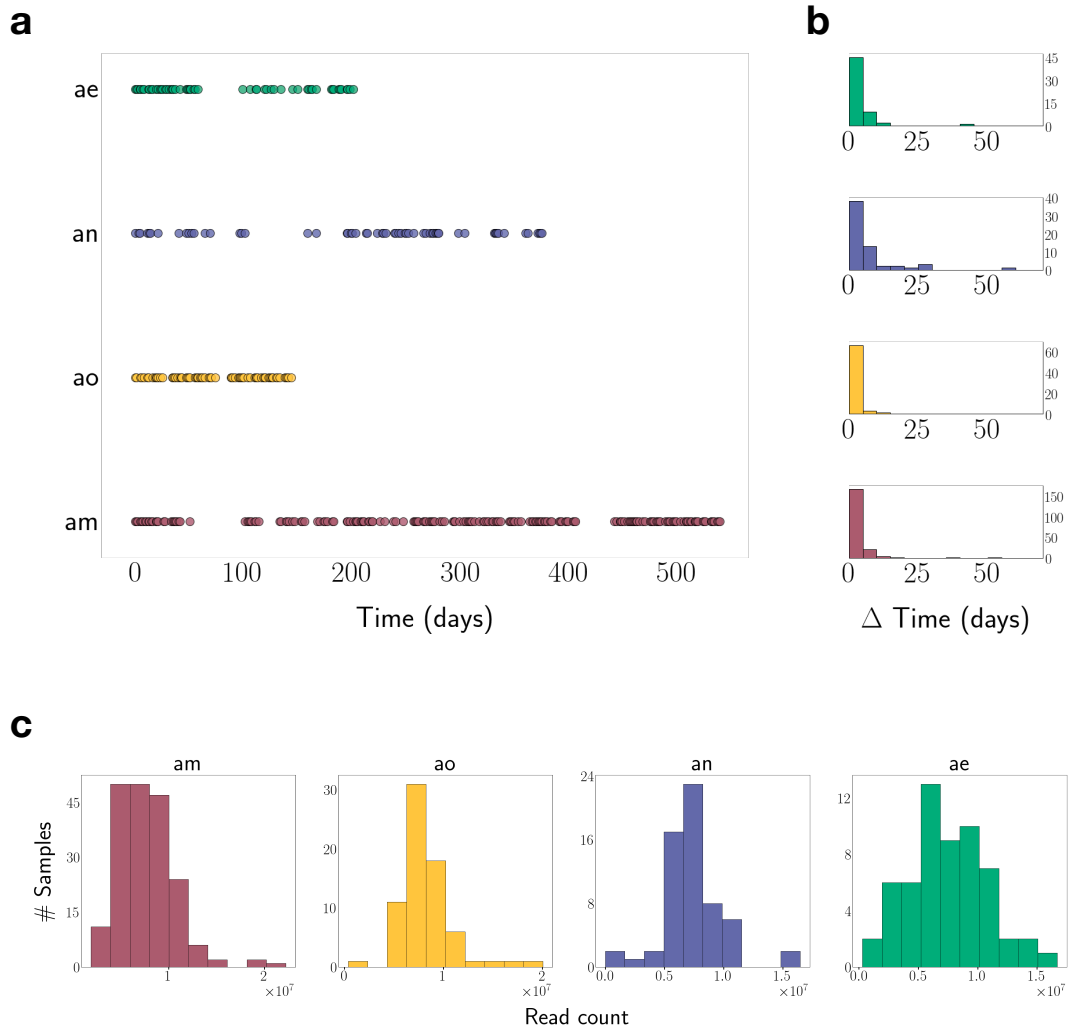


Figure S2: **a.** Four BIO-ML stool donors (aliases: *am*, *ao*, *an*, and *ao*) were longitudinally sampled more than 50 times. Circles indicate that shotgun metagenomic sequencing data was collected at that timepoint. **b.** For each host, we plot the distribution of the number of days between consecutive samples. While gaps between successive successive samples were as long as 58 days, the median time between samples was either one or two days for each host (Supplementary Table 1). **b.** Distribution of the number of shotgun reads per sample, by host. The mean number of reads per sample in each host were *am*: 7.5m, *ao*: 7.6m, *an*: 7.3m, and *ae*: 8.3m.

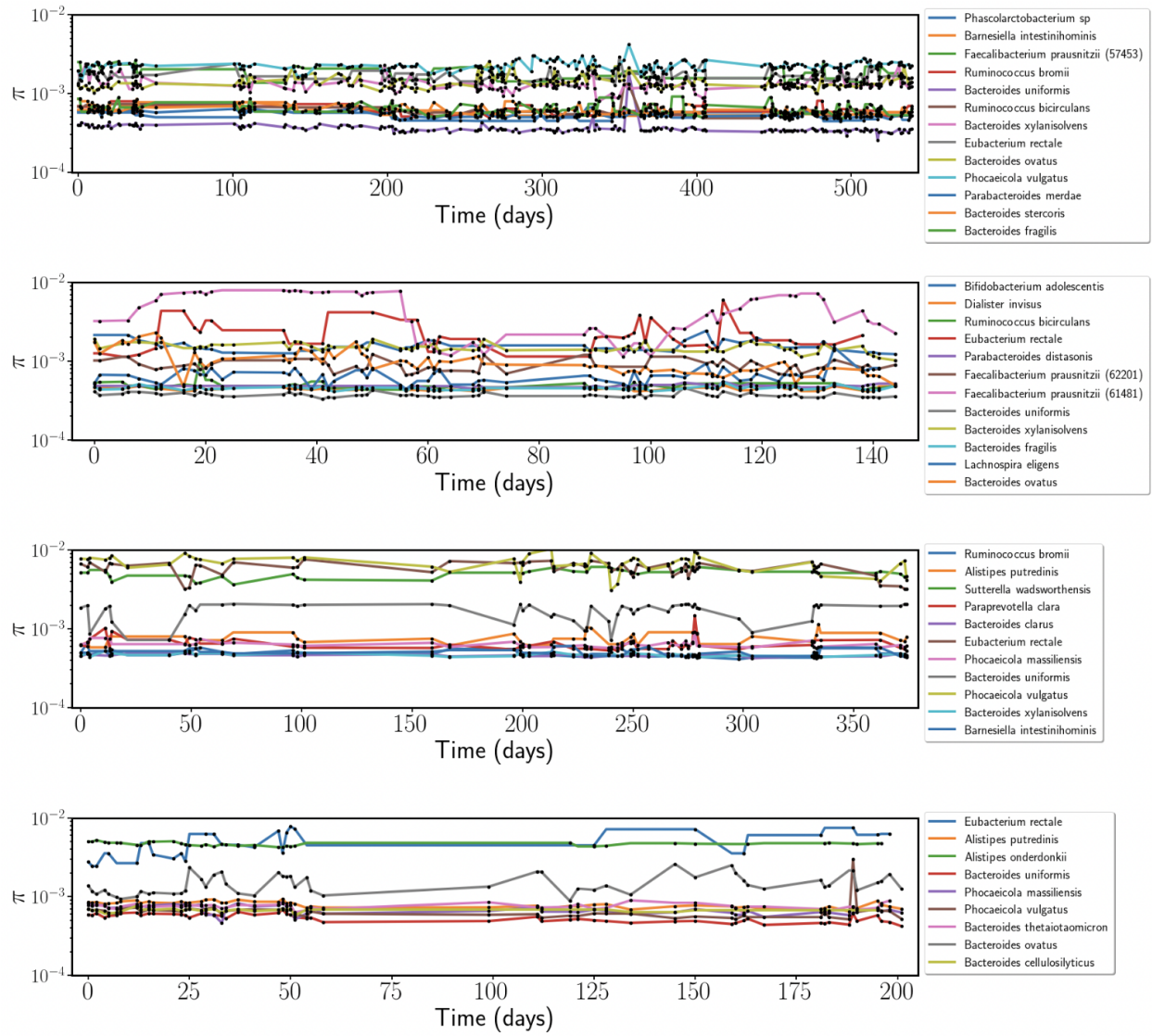


Figure S3: Nucleotide diversity π for each species/host combination.

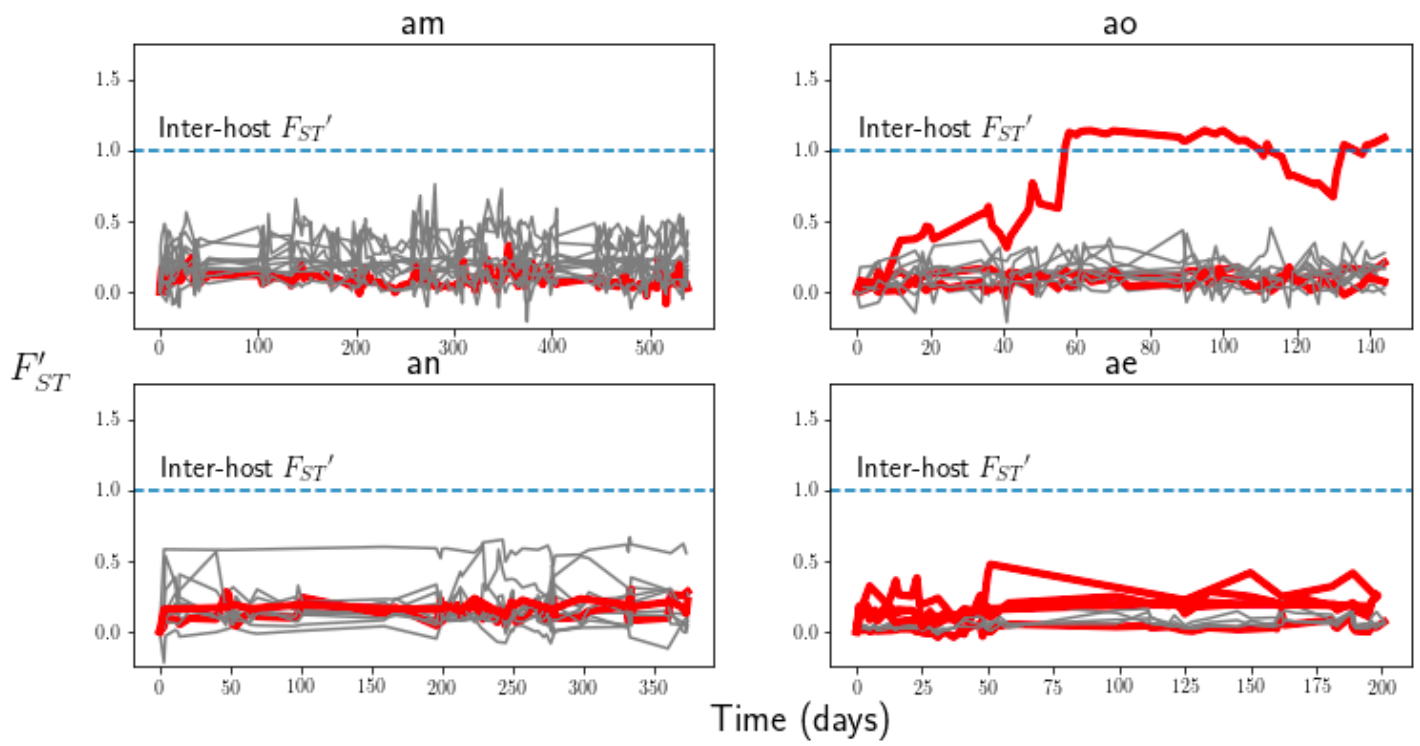


Figure S4: F'_{ST} values plotted for each species in each host, with species failing the ADF test of stationarity highlighted in red.

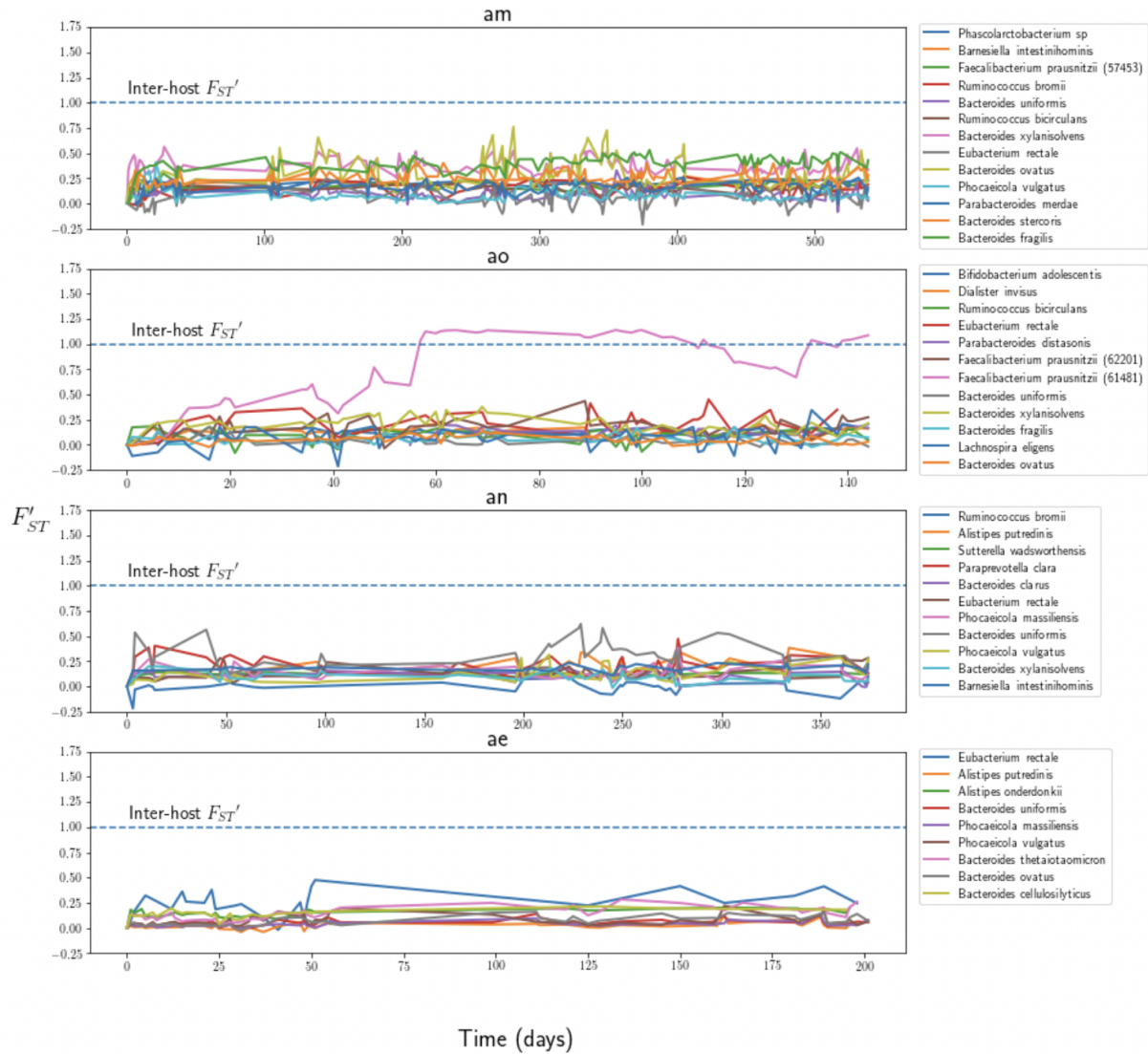


Figure S5: F_{ST}' values plotted for each species in each host. For details on how F_{ST}' was calculated, see Materials and Methods of Main Text.

Supplementary Text

1 Data

Publicly available shotgun metagenomes from the Broad Institute-OpenBiome Microbiome Library (BIO-ML) were downloaded from the Sequence Read Archive (BioProject PRJNA544527). A total of 402 samples drawn from four individuals (aliases *am*, *ao*, *an*, and *ae*) were downloaded, with 206 samples coming from *am*, 74 from *ao*, 63 from *an*, and 59 from *ae* (Table S1). All four hosts were healthy adults between the ages of 21 and 37, of whom three were male and one female. The full metadata concerning these hosts reported by BIO-ML is available as Supplementary File S1.

The four hosts analyzed in this paper were sampled near daily for between 144 days (*ao*) and 539 days (*am*). While gaps between successive successive samples were as long as 58 days (*an*), the median time between samples was either one or two days (Table S1).

When sample metadata indicated that multiple samples had been collected on the same day, or were technical replicates of one another, samples were merged by concatenating fastq files before running subsequent metagenomic processing scripts (Supplementary section 2). Thus, all samples taken from a single host used in this study had a unique date, and were temporally ordered.

Additionally, to calculate between-host F_{ST} , we used data from a previously collated panel [8] of healthy human subjects of 250 North American subjects sequenced by the Human Microbiome Project [19, 20]. Using this much broader database of samples, we estimated levels of inter-host genetic diversity.

2 Metagenomic pipeline

We used the MIDAS software package to align metagenomic reads to reference genomes, estimate species abundances, and call copy number variation (CNV) and single nucleotide variant (SNV) content [3]. MIDAS's reference database is made up of 31,007 high quality bacterial genomes, clustered into 5,952 species groups based on a 96.5% sequence identity across 30 universal marker genes (MIDAS reference database version 1.2). This marker gene average nucleotide identity (ANI) threshold was chosen by the original authors of MIDAS to ensure a 95% genome-wide ANI species definition. Throughout this manuscript, strains belonging to a species are much more closely related than this 95% threshold—typically, between 99% and 99.9% ANI.

All three basic modules in the MIDAS pipeline (estimating species abundance, calling CNV content, and calling SNV content) consist of two steps. First, each sample is processed individually to ascertain the relative abundance of all species, as well as the CNV and SNV content for those species passing user-specified thresholds in that sample (for details on threshold choices, see Supplementary section 2.2). And second, individual samples belonging to the same host are integrated to identify patterns of longitudinal variation within hosts.

2.1 Species abundance

MIDAS first estimates species abundances by aligning reads to 15 universal, single-copy marker gene sequences, which are known for each species in the MIDAS database. The average read coverage across these single-copy marker genes is then taken to be the abundance of the species in the sample. In our subsequent analyses, we excluded species which were considered present in fewer than 25 samples (see Supplementary section 5).

2.2 Building reference panels

When calling CNV and SNV content, MIDAS relies on a sample-specific reference panel to ensure that reads are only mapped to species which are truly present. We considered a species present if it had an average marker gene coverage ≥ 3 in that sample. While many more species are present at lower abundances, and are detected as such by MIDAS, SNV and CNV content cannot be confidently determined in these cases due to their relatively poor coverage [8].

To conservatively guard against spurious detection of intra-species genetic diversity, a "black-list" of genes defined in [8] which are known to be commonly shared between species were excluded from further analysis were excluded from all downstream analyses.

Lastly, to ameliorate read stealing and donating, we included all species which were present at any timepoint within a host in the reference panel for all samples in that host. By doing so, we keep the set of pan-genomes to which reads are aligned consistent across samples. By taking this precaution, we ensure that the species a specific read is mapped to will remain the same regardless of the sample in which this read originates.

2.3 CNV content

CNV content is determined by aligning reads to pan-genomes present in the MIDAS database for each species present in our data. MIDAS defines a pan-genome as the set of unique genes identified across all genomes available in the MIDAS reference database belonging to a species.

Reads were aligned to the host-specific reference panel using default MIDAS settings (local alignment, MAPID $\geq 94.0\%$, READQ ≥ 20 , and ALN_COV ≥ 0.75), and the average coverage was estimated by dividing the total number of reads mapping to the gene cluster by its total target size.

MIDAS determines the copy number c of a gene by dividing the average coverage of the gene by the average coverage of the species at its universal, single-copy markers (i.e. the species abundance). Using these copy number estimates, we then defined the core genome for each species in each host as the set of genes with copy number $c \geq 0.3$ in 90% of samples belonging to that host. Because we defined the core genome in a host-specific manner, this core represents the set of genes which are shared between all strains *within a host*. Species detected in multiple hosts, therefore, may have somewhat different gene content in the "core" genome.

2.4 SNV content

MIDAS calls SNVs with reference to a single representative genome per species group (specified in the MIDAS database, and chosen so as to have minimal marker gene distance to all other isolate genomes associated with the same species). The total length, in base pairs, of the representative genomes used by MIDAS for SNV calling purposes for species analyzed in this work is shown in (Table S2).

To ascertain SNVs, we aligned reads to the species included for a given sample's reference panel using default MIDAS mapping parameters: global alignment, MAPID $\geq 94.0\%$, READQ ≥ 20 , ALN_COV ≥ 0.75 , and MAPQ ≥ 20 . After alignment, we further excluded from analysis species that did not have a mean genome-wide coverage ≥ 5 and non-zero coverage at $\geq 40\%$ of reference sites within a sample.

3 Detecting strains

3.1 *Oligo-colonization*

A wide range of human microbiome studies have established that gut microbial species harbor genetic diversity within individual hosts [6, 2, 20]. Species are in fact frequently composed of multiple strains, and which are much more distantly genetically related than lineages which diverged subsequent to colonization [10, 14, 23]. When levels of recombination between such strains are sufficiently low, the clonal descendants of the initial colonizers may persist within hosts as genetically distinguishable subpopulations [24]. Interestingly, a number of studies have demonstrated that this kind of multiple colonization is subject to some degree of ecological constraint: only a small handful of strains (typically between one and four) are ever observed within a host at any one time, a phenomenon dubbed “oligo-colonization” (see Figure S1) [5, 8, 11, 9]. The general ecological and evolutionary mechanisms enabling a small number of strains to colonize a host and rise to high frequency, but preventing a large number of exogenous strains from doing the same, are not known as yet. Indeed, the factors governing strain coexistence may vary greatly between species. For instance, only a single strain of the common gut commensal *Bacteroides fragilis* is ever found within a host, potentially due to unusually potent inter-strain competition in this species mediated by the Type-VI secretion system [9]. However, other species, like *Phocaeicola vulgatus*, frequently harbor several strains [5, 23]. Whether such fine-scale diversity is maintained due to the partitioning of metabolic niches, by spatial segregation, or due to chaotic or neutral dynamics is an active area of research [17, 18]. In this work, we examine the dynamics of oligo-colonization through time.

Genetically diverged strains of a species residing in different hosts will typically differ from one another at $\mathcal{O}(10^3) - \mathcal{O}(10^4)$ sites in their shared core genome [2, 4, 8, 23]. Likewise, when multiple strains of the same species inhabit the same host, they also tend to differ from one another at a similar number of sites—put differently, each strain will share a more recent common ancestor with a strain found in a different host than with its other co-colonizing strains [23]. On the other hand, even under conservative estimates of the per basepair mutation rate and host lifetime, lineages sharing a common ancestor within the host are expected to harbor orders of magnitude fewer segregating mutations [8]. We will leverage this separation of scales in the number of sites expected to segregate between lineages sharing a common ancestor within the host, and those belonging to separately colonizing strains, to infer the presence of strains.

Resolving the number, relative abundances, and genotypes of co-colonizing strains from a single shotgun metagenomic sample is a difficult inverse problem. Because short read shotgun sequencing destroys evidence of physical linkage information over , methods of resolving strains within a sample rely on statistically deconvolving the distribution of allele frequencies to find areas of high density [15, 16]. However, due to finite and variable metagenomic read coverage across the genome, even perfectly linked alleles residing on a single strain’s genetic background may have quite different observed allele frequencies due solely to sampling noise. However, several recent studies [5, 13, 14, 24, 26] have shown that strain inference can be significantly improved when many samples have been collected longitudinally from a host, as alleles linked on a single strain’s genetic background will have clustered allele frequency trajectories. Strain inference algorithms designed specifically for such longitudinal data can leverage these temporal correlations between trajectories as a potent additional source of information when inferring true patterns of linkage within species.

3.2 Strain inference

In our study, we modified the allele frequency trajectory clustering algorithm developed by Roodgar *et al.* [5] to specifically detect large clusters of SNVs (> 1000 SNVs) in a species' core genome which strongly support the presence of multiple strains. While the approach of Roodgar *et al.* was appropriate for their purposes—namely, detecting selective sweeps of linked variants that deviated substantially from the overall background—the original clustering UPGMA-based clustering procedure was not sensitive enough to reliably detect low frequency strains in our dataset and thus required minor modifications. Our choice of 1000 SNVs as a cutoff was informed by the previously mentioned (Supplementary section 3.1) lower bound on the typical number of SNVs segregating between strains found in different hosts ($\mathcal{O}(10^3)$ sites). While lineages can and do diverge due to evolution within hosts [6, 7], by imposing the minimum cluster size of 1000 core genome SNVs, we expect largely to exclude cases of within-host diversification.

To cluster SNV trajectories into strains, we first calculated the distance metric defined in Roodgar *et al.* (which we review here) between every pair of polymorphic SNVs with sufficient coverage. To be considered, we required a site to have depth of coverage $D \geq 10$ and some level of polymorphism ($0 < f < 1$, where f is the frequency of the reference allele) in at least 25% of timepoints. By imposing these filters, we excluded sites that were not truly polymorphic, and/or for which true allele frequencies could not be accurately estimated.

This metric weights differences in allele frequency trajectories between sites by their joint sequencing depth, such that small differences in frequency at high coverage sites contribute proportionally more than equally small differences at low coverage sites. In practice, this ameliorate the of sampling noise. For a pair of SNVs i and j , the distance between their observed allele frequency trajectories \hat{f}_i and \hat{f}_j is:

$$d(\hat{f}_i, \hat{f}_j) = \frac{1}{T} \sum_{t=1}^T \frac{2(D_{it} + D_{jt})(\hat{f}_{it} - \hat{f}_{jt})^2}{(\hat{f}_{it} + \hat{f}_{jt})(1 - \hat{f}_{it} + 1 - \hat{f}_{jt})} \quad (1)$$

where T is the number of sampling timepoints, \hat{f}_{it} and \hat{f}_{jt} are the allele frequencies of i and j at a timepoint t , and D_{it} and D_{jt} are the depths of sequencing coverage at timepoint t .

At any biallelic site, there are two possible choices of allelic reference state. While the choice is *a priori* arbitrary, it will impact our ability to infer strain structure, as the metric (Equation 1) relies on both sites having the same relative polarization. To account for differences in relative polarization, we define a pair of metrics:

$$\begin{aligned} d_{ij}^+ &= d(\hat{f}_i, \hat{f}_j) \\ d_{ij}^- &= d(\hat{f}_i, 1 - \hat{f}_j) \end{aligned} \quad (2)$$

and take the distance between i and j to be the minimum of the two polarizations:

$$d_{ij} = \min(d_{ij}^+, d_{ij}^-). \quad (3)$$

Still following [5], we calculated d_{ij} between every pair of SNVs, forming a distance matrix. However, having calculated this SNV distance matrix, we took a different approach to clustering the SNVs into strains. Specifically, while [5] applied UPGMA clustering to the matrix of SNV distances, we implemented a greedy, network-based clustering algorithm. Our algorithm allows us to extract only large clusters of SNVs which are likely in perfect linkage with one another.

We considered each SNV a node in a network, and formed an edge between any pair of nodes (i, j) if $d(\hat{f}_i, \hat{f}_j) < d^*$, where d^* is a critical distance indicating likely true linkage. We used the empirically derived threshold maximum distance criterion of $d^* = 3.5$, established in [5], to identify pairs of SNVs that were likely linked to one another on a haplotype background. In essence, this network connects SNVs whose allele frequency trajectories indicate that they likely rely on the same haplotype background.

Next, we calculated the degree $|D_i|$ (i.e. number of edges) for all nodes in the network. We identified the node k with the highest degree $|D_k|$ in the network, and if $|D_k| > 1000$, we assigned k and all nodes connected to it to be a strain. To ensure that no outlier SNVs not truly linked to this cluster were assigned to it, we removed all nodes from the cluster which were connected to less than 25% of other nodes. Lastly, we removed all nodes assigned to this cluster from the network. We then repeated this procedure until no clusters of greater than 1000 nodes remained. If there were no clusters of more than 1000 SNVs initially, we inferred that only a single strain of the species was present.

When two strains are present, all SNVs which segregating between the strains will ultimately be polarized in the same way, and so only a single cluster of SNVs will be detected (for ease of visualization, we have shown both the detected cluster and in the Figure 1 of the Main text, as well as in the Supplemental strain figure files). However, when three strains are present, each pair of strains will have some number of sites segregating between them. Consequently, our strain deconvolution algorithm will detect three clusters of SNVs when three strains are present. In the two cases in which we detected three strains, we determined the correct relative polarizations of these clusters with respect to one another by summing all different possible polarizations, and looked for the choice of polarization that yielded a sum which was close to one at all timepoints, as the sum of the relative strain frequencies should be precisely one at each timepoint. In both cases, there was a single choice of polarization that resulted in the sum of strain trajectories remaining near one at all timepoints (green trajectories), and these polarizations were therefore chosen when determining the true frequencies of the strains (Figure S6).

Recently, Zheng *et al* [25] employed a high-throughput single-cell barcoding scheme to analyze strain dynamics in samples collected from host *am* of the BIO-ML project. As in our work, Zheng *et al* uncovered three dominant strains of *P. vulgatus* (a fourth strain was recovered from their co-assembly, but was detected only at a single timepoint, had very low genome completeness (6%), and was of "low quality"). Though collected at a far coarser sampling interval (seven timepoints, versus the 193 timepoints available using shotgun metagenomic samples), there is a striking visual concordance in the dynamics of the three dominant strains of *P. vulgatus* between our inference and that of Zheng *et al*. Interestingly, in the original BIO-ML publication [6], isolates of *P. vulgatus* collected from *am* indicated the presence of only two strains. That our inference, which relies on relatively easily obtainable shotgun metagenomic sequencing data processed through publicly available software in a matter of days, reveals similar strain dynamics at far greater temporal resolution as the state-of-the-art sequencing approach of Zheng *et al* in this case underscores the power of densely temporally sampled, high coverage shotgun sequencing data to recover patterns of intra-specific genetic variation in the microbiome.

4 SLM

To simulate the SLM, we used the Euler-Maruyama method:

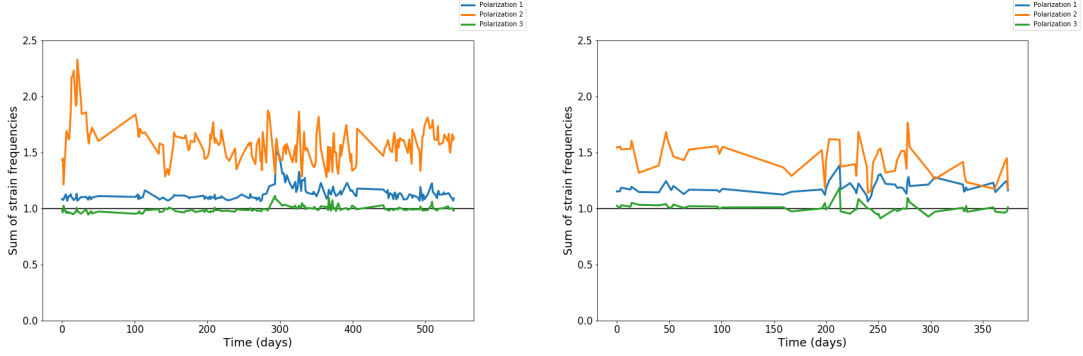


Figure S6: In two samples, *am* (left) and *an* (right), three strains of *P. vulgatus* were detected. In order to assess the correct relative polarizations of the three strains, we summed all different possible polarizations of the detected clusters. In both cases, the green polarization shown in the figure above was approximately equal to one at all timepoints, and so this polarization was chosen when determining the true frequencies of the strains.

$$X(t + \delta t) = X(t) + \frac{x(t)}{\tau_i} \left(1 - \frac{x(t)}{K}\right) \delta t + \sqrt{\frac{\sigma}{\tau}} x(t) Z_t \sqrt{\delta t} \quad (4)$$

where Z_t is a standard normal random variable. In simulations, we set $\delta t = \frac{1}{1000}$.

The SLM associated with population i depends on three parameters: K_i , σ_i , and τ_i . K_i and σ_i are not fit, but rather are determined directly from the mean and variance of the actual time series using the formulae:

$$\sigma_i = \frac{2}{\frac{\langle x_i \rangle^2}{\sigma_{x_i}^2} - 1}, \quad K_i = \frac{\langle x_i \rangle}{1 - \frac{\sigma_i}{2}} \quad (5)$$

where $\langle x_i \rangle$ is the mean abundance of the population and $\sigma_{x_i}^2$ is its variance. The parameter τ_i was held constant ($\tau_i = 1$) for all strains to avoid overfitting.

For each time series of strain abundances, we fit the K_i and σ_i parameters using the first third of timepoints.

5 Goodness of fit test

To assess the predictive capacity of the SLM, we devised a goodness of fit test. The test seeks to determine not only whether the true dynamics of a strain's abundance match the behavior expected under the SLM, but also whether these dynamics can be predicted from a subset of initial points. Heuristically, the test determines whether the full sequence of transitions between timepoints statistically resembles the sequence of transitions expected were the data truly generated by the SLM.

First, we calculated the total abundance of the strain in question, observed at timepoints $t_0, t_1, t_2, \dots, t_T$, to be $x(t_0), x(t_1), x(t_2), \dots, x(t_T)$. The total abundance of a strain is its relative abundance within the

community at large—that is, its relative frequency within the species (determined using the strain inference algorithm described in the preceding section), multiplied by the relative frequency of the species within the community.

For each strain, the SLM is “trained” on the first 33% of data points (i.e. $x(t_0), x(t_1), \dots, x(t_{T/3})$), with the model parameters K_i and σ_i determined as described in the main text directly from the mean and variance in abundance across these timepoints.

The fundamental approach employed to test the fit of the SLM is to estimate an empirical distribution under the model of possible abundances at t_{i+1} , given the known initial abundance $x(t_i)$, and then observing where the true $x(t_{i+1})$ lies. To do so, we simulated the model 1000 times between each pair of subsequent timepoints t_i and t_{i+1} ($i \in [(T/3 + 1), (T/3 + 2), \dots, T]$), starting with initial abundance $x(t_i)$. Next, to assess where the true $x(t_{i+1})$ lay relative to our simulations, we subdivided this distribution into M equally likely quantiles (that is, M intervals of the real line each having equal probability under the empirical distribution of simulations), determined which quantile $x(t_{i+1})$ fell in, and denoted this value q_i .

Under the null hypothesis, $x(t_{i+1})$ should have equal probability of lying in each quantile, precisely because the quantiles are non-overlapping areas of equal probability under the SLM. Therefore, $\{q_i\}$, the set of quantiles determined in this manner between all pairs of consecutive timepoints, is expected to follow a uniform distribution on $[1, 2, \dots, M]$ if the true trajectory was generated by the SLM. To assess whether this was in fact the case, we performed a χ^2 test with M degrees of freedom. We rejected the null at a significance level of 5%. We do not perform multiple hypothesis testing here, as we only make a single comparison (pass/fail of SLM).

In determining M , we followed the convention for χ^2 tests that the expected number of observations in each bin (here, $[1, M]$) should exceed 5. As there are a total of T data points, M was chosen so that $\frac{T}{M} > 5$. Additionally, to ensure that our χ^2 test had sufficient power, we only considered trajectories which had $T > 25$ timepoints [1].

6 Taylor’s Law

Taylor’s Law, a power law scaling between the mean and variance of the observed abundances of populations within a community across either space or time, is an extremely general phenomena observed in a wide variety of ecosystems [28, 27, 30, 31, 32]. The specific scaling exponent observed is affected both by true biological variability in the underlying populations’ abundances and by noise introduced during the sampling process. While many different ecological models can produce Taylor’s Law with a wide variety of scaling exponents [29, 33], sampling noise, by itself, is expected to induce a Taylor’s Law α exponent of one. As an example, consider a community where population abundances are governed by Poisson fluctuations introduced solely by sampling noise:

$$P(x_i = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (6)$$

where λ is a composite rate parameter, the product of the total strength of observation (e.g. total number of individuals censused in a community) and the focal population’s relative abundance. Because the mean of the Poisson distribution equals its variance (both are equal to λ , in this case), we should observe a Taylor’s Law with $\alpha = 1$ (i.e. $\sigma_{x_i}^2 \propto \langle x_i \rangle$, trivially).

Real metagenomic data, however, is compositional—when comparing samples, we do not assess abundance directly, but rather observe the relative abundances of populations. If we assume that there are a finite number reads, we can model the situation in which the variation in observed

Multinomial compositional sampling

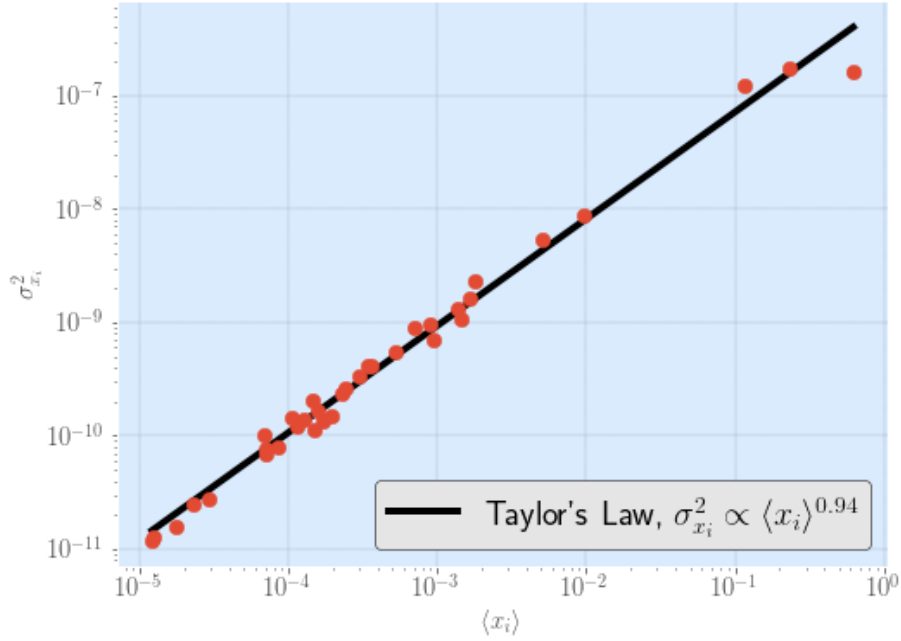


Figure S7: α is approximately equal to 1 under multinomial sampling

frequencies is determined solely by sampling noise by, first, repeatedly sampling total abundances from a multinomial distribution:

$$P(x_1, x_2, \dots, x_N; N, f_1, f_2, \dots, f_N) = \frac{S!}{x_1! x_2! \dots x_N!} f_1^{x_1} f_2^{x_2} \dots f_N^{x_N} \quad (7)$$

where S is the total number of reads, N total number of populations present, and $\bar{\mathbf{f}}$ is vector of underlying true population abundances (here, drawn from a lognormal distribution), and T is the number of timepoints. Following this sampling step, we can then characterize fluctuations in relative abundance by dividing all observed abundances by the total read depth, S . Doing so, we observe a Taylor's Law scaling with $\alpha \approx 1$, as expected (Figure S7).

Moreover, even if the total read count S is allowed to vary between samples, α still remains approximately equal to 1. Specifically, if S is itself drawn from a Poisson distribution—as would be expected if sampling intensity was roughly equal between samples—and observed population frequencies are affected only by sampling, we continue to observe an $\alpha \approx 1$ (Figure S8).

Though the relative abundances of all strains we observe in this study are certainly affected by sampling noise, and (as we have seen) a Taylor's Law scaling can emerge due solely to compositional sampling, compositionality by itself cannot explain the scaling observed in the data. Specifically, the specific exponent we observe here ($\alpha = 1.8$, Main text Figure 3) is highly unlikely to be primarily driven by sampling effects, whether compositional or not, but rather reflects true underlying biological variability in the relative abundance of the populations.

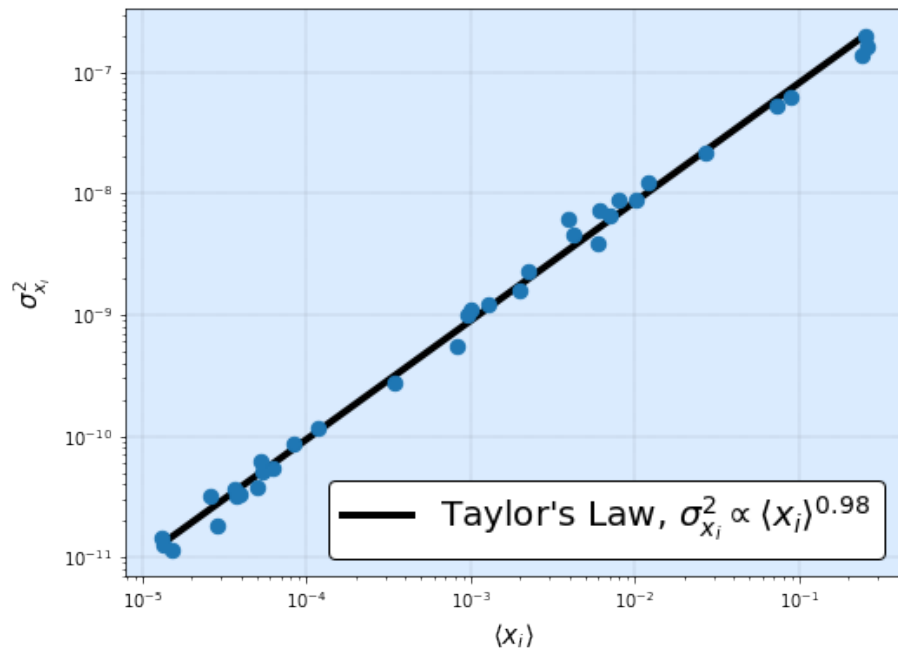


Figure S8: α is approximately equal to 1, even when coverage varies between samples.

References

- [1] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine. Series 5.* 50 (1900), pp. 157-175
- [2] Schloissnig, Siegfried, et al. Genomic variation landscape of the human gut microbiome. *Nature* 493.7430 (2013): 45-50.
- [3] Nayfach, S., Rodriguez-Mueller, B., Garud, N., & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome research*, 26(11), 1612-1625. (2016).
- [4] Van Rossum, Thea, et al. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology* 18.9 (2020): 491-506.
- [5] Roodgar, M., Good, B. H., Garud, N. R., Martis, S., Avula, M., Zhou, W., Snyder, M. P. et al. Longitudinal linked read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. *Genome Research*. (2021).
- [6] Poyet, M., Groussin, M., Gibbons, S.M. et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* 25, 1442–1452 (2019).
- [7] Zhao, S., et al. Adaptive evolution within gut microbiomes of healthy people. *Cell host and microbe*, 25(5), 656-667. (2019).
- [8] Garud N; Good B; Hallatschek O; Pollard K. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biology* (2019).
- [9] Verster, Adrian J., et al. The landscape of type VI secretion across human gut microbiomes reveals its role in community composition. *Cell host and microbe* 22.3 (2017): 411-419.
- [10] Olm, Matthew R., et al. Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different *in situ* growth rates. *Genome research* 27.4 (2017): 601-612.
- [11] Truong, Duy Tin, et al. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research* 27.4 (2017): 626-638.
- [12] Russell, Shelbi L., and Colleen M. Cavanaugh. Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes. *Molecular Biology and Evolution* 34.11 (2017): 2747-2761.
- [13] Goyal, Akshit, et al. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *Elife* 11 (2022): e74987.
- [14] Kang, Jonathan TL, et al. Long-term ecological and evolutionary dynamics in the gut microbiomes of carbapenemase-producing Enterobacteriaceae colonized subjects. *bioRxiv* (2022).

- [15] Smillie, Christopher S., et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell host and microbe* 23.2 (2018): 229-240.
- [16] Smith, Byron J., et al. Scalable microbial strain inference in metagenomic data using StrainFacts. *bioRxiv* (2022).
- [17] Wang, Zihan, et al. Fine-scale diversity of microbial communities due to satellite niches in boom-and-bust environments. *bioRxiv* (2022).
- [18] Mahadevan, Aditya, Michael T. Pearce, and Daniel S. Fisher. Spatiotemporal Ecological Chaos Enables Gradual Evolutionary Diversification Without Niches or Tradeoffs. *bioRxiv* (2022).
- [19] Consortium HMP. A framework for human microbiome research. *Nature*. 2012;486:215– 221.
- [20] 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;550(7674):61.
- [21] Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, et al. Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst*. 2016;3(6):572–584 e3. doi:10.1016/j.cels.2016.10.004.
- [22] Qin J, Li Y, Cai Z, Li S, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490:55–60.
- [23] Yassour, Moran, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science translational medicine* 8.343 (2016): 343ra81-343ra81.
- [24] Quince, Christopher, et al. STRONG: metagenomics strain resolution on assembly graphs. *Genome biology* 22.1 (2021): 1-34.
- [25] Zheng, Wenshan, et al. High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* 376.6597 (2022): eabm1483.
- [26] Dmitrijeva, Marija, et al. Strain-resolved dynamics of the lung microbiome in patients with cystic fibrosis. *MBio* 12.2 (2021): e02863-20.
- [27] Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nat Commun* 11,4743 (2020).
- [28] Ji, Brian W., et al. Macroecological dynamics of gut microbiota. *Nature microbiology* 5.5 (2020): 768-775.
- [29] Kilpatrick, A. M. & Ives, A. R. Species interactions can explain Taylor’s power law for ecological time series. *Nature* 422, 65–68 (2003)
- [30] Taylor, L. R. & Taylor, R. A. Aggregation, migration and population mechanics. *Nature* 265, 415–421 (1977).
- [31] Kendal, W. Taylor’s ecological power law as a consequence of scale invariant exponential dispersion models. *Ecol. Complex.* 1, 193–209 (2004).

- [32] Giometto, A., Formentin, M., Rinaldo, A., Cohen, J. E. & Maritan, A. Sample and population exponents of generalized Taylor's law. *Proc. Natl Acad. Sci. USA* 112, 7755–7760 (2015).
- [33] Zhao, Lei, et al. Proximate determinants of Taylor's law slopes. *Journal of Animal Ecology* 88.3 (2019): 484-494.