

# A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome

## Running Title

Phylogenetic community assembly of microbes

## Authors

John L. Darcy<sup>1</sup>, Alex D. Washburne<sup>2</sup>, Michael S. Robeson<sup>3</sup>, Tiffany Prest<sup>4</sup>, Steven K. Schmidt<sup>4</sup>, Catherine A. Lozupone<sup>1</sup>

## Affiliations

<sup>1</sup> Division of Biomedical Informatics and Personalized Medicine, University of Colorado School of Medicine, Aurora, Colorado, USA.

<sup>2</sup> Department of Microbiology and Immunology, Montana State University. Bozeman, Montana, 59717, USA.

<sup>3</sup> Department of Biomedical Informatics, University of Arkansas for Medical Sciences. Little Rock, Arkansas, 72205, USA.

<sup>4</sup> Department of Ecology and Evolutionary Biology, University of Colorado. Boulder, Colorado, 80309, USA.

## Corresponding Author

J.L. Darcy; darcyj@colorado.edu.

## Conflict of Interest Statement

The authors declare that no conflict of interest exists.

## Support

Funding was provided by an NSF grant for studying microbial community assembly following disturbance (DEB-1258160) and by a NIH NLM Computational Biology training grant (5 T15 LM009451-12). The funding bodies had no role in study design, analysis, interpretation, or in the preparation of this manuscript.

## Abstract

Understanding when and why new species are recruited into microbial communities is a formidable problem with implications for managing microbial systems, for instance by helping us better understand whether a probiotic or pathogen would be expected to colonize a human microbiome. Much theory in microbial temporal dynamics is focused on how phylogenetic relationships between microbes impact the order in which those microbes are recruited; for example species that are closely related may competitively exclude each other. However, several recent human microbiome studies have observed closely-related bacteria being recruited into microbial communities in short succession, suggesting that microbial community assembly is historically contingent, but competitive exclusion of close relatives may not be important. To address this, we developed a mathematical model that describes the order in which new species are detected in microbial communities over time within a phylogenetic framework. We use our model to test three hypothetical assembly modes: underdispersion (species recruitment is more likely if a close relative was previously detected), overdispersion (recruitment is more likely if a close relative has not been previously detected), and the neutral model (recruitment likelihood is not related to phylogenetic relationships among species). We applied our model to longitudinal human microbiome data, and found that for the individuals we analyzed, the human microbiome generally follows the underdispersion (i.e. nepotism) hypothesis. Exceptions were oral communities and the fecal communities of two infants that had undergone heavy antibiotic treatment. None of the datasets we analyzed showed statistically significant phylogenetic overdispersion.

## Introduction

A central question in microbial community assembly, especially in the human microbiome, is when and why microbes are recruited into communities. If there are patterns or general rules for which taxa have higher probabilities of recruitment, these rules can guide habitat restoration projects, help us better understand whether probiotics or pathogens will colonize, and better exploit disturbance as a tool for managing microbial systems related to human health and disease. Recruitment of new species can be studied by evaluating the order in which species are empirically detected in time-series experiments, given data such as which species have already been detected or what changes occur in an environment over time [1; 2]. Although a changing environment clearly selects for new species, it has also been shown that microbial community structure is often historically contingent on previous states of that community [1; 3; 2; 4; 5]. This reflects not only that microbial communities are temporally autocorrelated (gradual change over time), but also that the recruitment of a given species is a function of which species in the community are already present or have modified the local environment. Such historically contingent patterns have mainly been observed and tested within a phylogenetic context, because amplicon data naturally lend themselves to the creation of phylogenies, and because phylogenies have been shown to be predictive of genomic (and perhaps niche) overlap in human associated microbiota [6; 7] and in general [8; 9].

Within this phylogenetic framework, a predominant hypothesis has been that closely related microbes inhibit each other’s successful recruitment [1; 3; 4]. The proposed mechanism for this hypothesis is that closely related microbes likely have similar niches (phylogenetic niche conservatism [10; 8; 9]), and species already established within a community will occupy their niches to the exclusion of ecologically similar organisms. This is also the basis of Darwin’s naturalization hypothesis [11], which proposed that new species are less likely to be recruited if a close relative is present [12]. Indeed, this assembly mode has been found to be the case in artificial nectar microcosms, where phylogenetically similar yeast species had similar nutrient requirements, and inhibited each others’ colonization [13]. If competitive exclusion of closely-related species is the predominant mode of microbial community assembly in the human microbiome, close relatives would be less likely to be recruited into the same community, as compared to more distant relatives. In this paper, we refer to assembly where distant relatives are more likely to be recruited into a community than close relatives as the **overdispersion hypothesis**, since it predicts the preferential addition of novel phylogenetic diversity to a community (*i.e.* phylogenetic overdispersion).

Overdispersion is far from universal, and multiple studies have shown that extremely close relatives can coexist within the human microbiome [14; 15; 16], and may even be preferentially recruited [17]. This is consistent with simulations showing that clusters of closely-related species can persist despite strong within-cluster competition, when immigration rate is high [18]. Indeed, Darwin’s pre-adaptation hypothesis predicts

that species with a close relative present in a community will be preferentially recruited, because they are likely to already be adapted to the new environment [11]. Also, more recent ecological theory posits competition between distantly-related species may result in phylogenetically clustered communities, although it may be more appropriate to include this type of competition within selection [19]. Per the empirical observations and reasoning above, we might expect that new close relatives are more likely to be detected than new distant relatives, so the amount of new phylogenetic diversity added to a community is minimized (phylogenetic underdispersion). Thus, for empirical data our **underdispersion hypothesis** is that species are more likely to be detected when they have a close relative that was previously detected, predicting preferential addition of minimal novel phylogenetic diversity (phylogenetic underdispersion). The over- and underdispersion hypotheses are alternatives to the null hypothesis that recruitment is independent of phylogenetic relatedness among species. Since the null hypothesis is species-neutral (and phylogenetically neutral), we refer to it as the **neutral hypothesis**.

It should be noted that our use of the terms "overdispersion" and "underdispersion" are slightly different in this manuscript compared to use of the same terms elsewhere. In many cases, these words refer to the state of a community at a single timepoint or sample, with overdispersion indicating more diversity in that sample than expected by chance, and underdispersion indicating less [20]. Instead, our use of over- and underdispersion refers to the amount of newly added diversity over time. In our overdispersion hypothesis, phylogenetically novel species are preferentially added to communities, meaning more new diversity is added than expected by chance. Under our underdispersion hypothesis, the reverse is true. Following this, our question concerns the order in which new species are recruited in a time-series, rather than community composition of any given sample. Furthermore, while we are interested in the biological phenomenon of species recruitment, the empirical result of recruitment is detection. Evidence of recruitment is a lack of detection, and then subsequent detection of a species via high-throughput DNA sequencing data. It is possible for a species to have been recruited into a community but not be detected, although this source of experimental error diminishes as sequencing depth increases. Also, the extent to which a species has actually been recruited into a community is questionable, if it is sufficiently rare that it is not detected in an Illumina sequencing run with tens of thousands of reads per sample (e.g. [21]). Still, we have taken care to use the term "recruitment" when discussing the biological phenomenon under investigation, and "detection" when describing empirical data and results.

Here, we use the phylogenetic relationships among species within a time-series to test the extent to which our over- or underdispersion hypotheses hold true. Instead of analyzing broad patterns of community change via beta-diversity statistics (e.g. UniFrac [22]) or analyzing patterns of select clades within the community (e.g. PhyloFactor [23], Edge PCA [24]), we model the probability of new species' recruitment into a community for the first time as a monotonic function of their phylogenetic distances to members of the community that have already been recruited.

The model we present here can be used to estimate the degree to which the recruitment of new species is more or less likely when a close relative has been previously recruited. We fit our model to several time-series human microbiome datasets [25; 26; 21] to compare the strength of under- or overdispersion between subjects, sample sites, or time periods. We found that for the data sets we analyzed (36 individuals across 3 studies), the human microbiome generally follows the underdispersion hypothesis. There were exceptions where this pattern was not significantly different than the neutral model, but none of the longitudinal datasets we analyzed showed statistically significant overdispersion.

## Materials and Methods

### Overview

With our model, our goal is to estimate the extent to which recruitment of new species over time is related to the new species' phylogenetic similarity to (or distance from) species that were already recruited at previous timepoints. Our **Statistical Model** describes the probabilities of detecting new species over time. We use our model with empirical data via **Simulations**, where we re-sample the empirically detected species using our model with known parameter values, to produce surrogate datasets. Specifically, we fix and record the model's dispersion parameter ( $D$ ), which determines the extent to which species with a close relative are preferentially added to the surrogate community (or, conversely, if species without a close relative are

preferred). Our **Parameter Estimation** compares the empirical pattern of species detection to that of the surrogate datasets (which have known  $D$  values), in order to determine which value of  $D$  best describes the empirical data. **Hypothesis Testing** is done by comparing empirical data to repeated simulations under the neutral model, which is  $D = 0$ . We describe the bioinformatic and technical details of this process in our **Analysis** section, and make our code available to others in the **Code and Data** section.

## Statistical Model

At any point in time, a community is composed of many species, and other species are not present but are available to be added ("species pool"). Our model parameterizes the probability of detecting species in a local community for the first time, based on their phylogenetic distances from species that have already been detected. In a species-neutral model of community assembly, each species  $i$  in the species pool has the same probability of detection at time  $t$ , irrespective of how different it is from species that have already been detected. Thus, the neutral model for first-time species detections is a random draw without replacement of species from the species pool. We extend the species-neutral model by modeling the probability  $p_{it}$  of species  $i$  being detected for the first time at time  $t$  as,

$$p_{it} = \frac{d_{it}^D}{\sum_i d_{it}^D} \quad (1)$$

where  $d_{it}$  is the phylogenetic distance from species  $i$  to its closest relative that has already been detected prior to timepoint  $t$ , and  $D$  is a dispersion parameter.

When  $D = 0$ , our model functions as a neutral model; all species have the same probability of being detected for the first time, since  $p_{it}$  is the same for every species. When  $D < 0$ ,  $p_{it}$  decreases with  $d_{it}$  meaning that species from the species pool have higher probabilities of detection when they are more closely related to species that have already been detected in the local community (underdispersion; phylogenetically constrained). When  $D > 0$ , the opposite is true (overdispersion; phylogenetically divergent). Our hypothesis testing and parameter estimation focus on the dispersion parameter,  $D$ .

## Simulations

Our analysis of a dataset relies on re-constructing that dataset via simulation of our statistical model using known values of  $\hat{D}$ , allowing for hypothesis testing and parameter estimation (we refer to the empirical dispersion parameter as  $D$ , and use  $\hat{D}$  to refer to surrogate values used in simulations). Using the empirical data as a starting point, we simulate many surrogate datasets with  $\hat{D}$  values ranging from  $\hat{D} < 0$  (underdispersed) to  $\hat{D} = 0$  (neutral) to  $\hat{D} > 0$  (overdispersed). This is done so that the empirical data can later be compared to the surrogate datasets, to estimate the empirical value of  $D$ .

We start each surrogate dataset with the same species present in the first sample in the time-series of its corresponding empirical dataset. Then, surrogate datasets are constructed forward in time by randomly drawing  $r_t$  new species from the species pool, where the probabilities of detecting those species are given by Equation 1, and  $r_t$  is the number of new species detected in the empirical dataset from times  $t - 1$  to  $t$ . The number of new species detected from the empirical dataset is used so that species richness is kept constant between the empirical dataset and all surrogate datasets. The species pool is updated to exclude those species drawn at previous timepoints, and the newly sampled species are recorded. Surrogate datasets are produced for many different  $\hat{D}$  values, ranging from underdispersed to overdispersed models. We performed 500 simulations (as described above) for each dataset analyzed.

## Parameter Estimation

Our main goal is to estimate the empirical dispersion parameter  $D$  (Equation 1), which quantifies the degree to which first-time species detections are phylogenetically underdispersed ( $D < 0$ ), neutral ( $D = 0$ ), or overdispersed ( $D > 0$ ), corresponding to our hypotheses. To this end, we use Faith's phylogeny [27] to compare each of the 500 surrogate datasets (described above) to the empirical dataset. Phylogeny is the sum of branch-lengths on a phylogenetic tree for a set of species, so phylogeny of a set of highly

related species is low (phylogenetically constrained) because there are no long branch lengths in the tree, but phylodiversity is higher (phylogenetically divergent) for a set of more distantly related species [27]. If  $D \neq 0$ , then species are preferentially added if they have relatively low ( $D < 0$ ) or relatively high ( $D > 0$ ) phylogenetic distance to the resident community ( $d_{it}$ , Equation 1), yielding accumulations of total phylodiversity that are relatively slow ( $D < 0$ ) or relatively fast ( $D > 0$ ) compared to the neutral model (Fig. 1A). In other words, at any timepoint  $t$ , the phylogenetic diversity of species that have already been observed is  $PD_t$ , and the extent to which  $PD_t$  accelerates or decelerates over a sampling effort depends on  $D$ . Because of this, we can estimate  $D$  by comparing the empirical phylodiversity curve to our surrogate phylodiversity curves, which have known  $\hat{D}$  values.

For the comparison of an empirical phylodiversity accumulation curve to curves for corresponding surrogate datasets, we evaluate the amount of phylodiversity  $PD_m$  accumulated at time index  $m$ , midpoint between the first and final samples. Time  $m$  is used because this leaves many species yet to be observed in the species pool, so that there can be variability in surrogate datasets. Multiple time indices are not used to compare surrogate and empirical datasets because each value  $PD_{\hat{t}}$  is a function of all values  $PD_{t < \hat{t}}$ .  $PD_m$  values are calculated for all surrogate datasets, and a  $PD_m$  value is calculated for the empirical dataset. The difference between the empirical  $PD_m$  and  $PD_m$  simulated with  $D = \hat{D}$  is  $\Delta PD_{\hat{D}}$ , which is the error between surrogate and empirical data. We then estimate the empirical value of  $D$  by minimizing  $\Delta PD_{\hat{D}}$  (Fig. 1B). This minimization is performed using a logistic error model,

$$\Delta PD_{\hat{D}} = \frac{a - b}{1 + e^{-r(\hat{D} - i)}} + b \quad (2)$$

where  $a$  and  $b$  are the upper and lower horizontal asymptotes, and  $r$  and  $i$  are rate and inflection parameters for the logistic model.  $\Delta PD_{\hat{D}}$  is modeled with a logistic function because there is a maximum and minimum observable  $\Delta PD_{\hat{D}}$  value as a function of the phylogeny; this is because there are strict minimum and maximum limits to the amount of phylodiversity obtainable by observing  $n$  species where  $n$  is the total species richness accumulated up to time  $m$ . The two horizontal asymptotes of the logistic model are easily fit to these extremes (Fig. 1B). Once fit, the error model is solved for  $\Delta PD = 0$ , giving an estimate for the empirical  $D$ . Confidence intervals for this estimate are obtained via bootstrapping our error model.

## Hypothesis Testing

For this test, our null hypothesis is the neutral model, where  $D = 0$ , since this model represents the absence of the effect we are testing. We test this null hypothesis competitively by simulating 1000 surrogate datasets at  $D = 0$  (Fig. S1A) to generate a null  $PD_m$  distribution. The empirical  $PD_m$  is compared to this distribution (Fig. S1B), and if the empirical  $PD_m$  is below the 2.5% quantile or above the 97.5% quantile, we reject the null (*i.e.* neutral) hypothesis. Evidence of either overdispersion ( $D > 0$ ) or underdispersion ( $D < 0$ ) allows us to reject. A second method of simulating communities under  $D = 0$  was used as well, where simulated communities were drawn without replacement from the pool of all individual observations within a rarefied data set. In this context, an "observation" is a datum within a species-by-sample matrix of count data; a given column contains  $d$  observations where  $d$  is the sequencing depth. This "individual null" accounts for differences in total species relative abundance across a time series, while the simulation above only considers presence-absence of species. Results of this model were computed the same as above.

## Analysis

This section is a summary of our data analysis. Detailed methods for this section are available as supplemental information.

We ran our model on data from 36 individuals from three data sources. Two individuals were from Caporaso *et al.* [25], 33 were from Yassour *et al.* [21], and one was from Koenig *et al.* [26]. In all cases, data were downloaded and processed using the unoise3 pipeline [28], which clusters sequence data into exact sequence variants called zOTUs. The Koenig *et al.* infant gut data set was split into two data sets, one for samples collected before the subject began consuming baby formula, and one after. Our model was run on these data as described above, resulting in  $D$  estimates for the before and after formula data sets.

The "moving pictures" [25] data were split into eight datasets, one for each combination of subject ( $n=2$ ) and body site (feces, right and left palms, tongue), and our model was run on each of these datasets. Analyses

of these data was also done using two approaches that allowed us to test the importance of the set of species that are included in the species pool. One alternate approach analyzed communities in a "meta" context, where the species pool for a given palm was composed of all four palms in the whole dataset. If we were to estimate similar  $D$  values for both the "meta" and "self" analyses, the inclusion of extra species in the species pool would be of little importance to the model. The other alternate approach analyzed data using a sliding-window approach, wherein our model was run separately on multiple overlapping windows of 5 consecutive days within the same dataset, in order to see how  $D$  varied over time.

Finnish infant sequence data from Yassour *et al.* [21] were split into data sets for each of 33 individuals, and our model was run for each. Estimated  $D$  values were compared between subjects that had been treated with oral antibiotics ( $n=18$ ) and subjects that had not ( $n=15$ ) using a Mann-Whitney test. Because this data source had so many subjects, we used these data to test whether the number of zOTUs, total phylodiversity, or number of timepoints had an effect on  $D$  estimates via correlation analysis.

## Code and Data

R code and data to replicate our analysis, or to perform a similar analysis on other data, are available on GitHub, at [https://github.com/darcyj/pd\\_model](https://github.com/darcyj/pd_model).

## Results

By varying  $\hat{D}$ , we successfully changed the rate at which phylodiversity is added to surrogate (*i.e.* re-sampled) microbial communities over time (Fig. 1A). Compared to the neutral model where  $\hat{D} = 0$ , higher  $\hat{D}$  values result in phylodiversity accumulating quickly, since in the overdispersed model, species that contribute more phylodiversity are preferentially sampled. Conversely, lower  $\hat{D}$  values result in phylodiversity accumulating slowly, since in the underdispersed model, species that contribute less phylodiversity (since they are very similar to species that are already present) are preferentially sampled. These results show that the  $D$  parameter in our model successfully corresponds to over- and underdispersion relative to the neutral model. Our error model also fit well to the differences between empirical and surrogate datasets ( $\Delta PD_{\hat{D}}$ , Fig. 1B). Each error model fit was visually inspected for goodness of fit, to be sure that  $D$  estimates were not spurious. All data sets passed this inspection.

## Results from "moving pictures" data

All time-series from adult feces and palm microbiomes [25] showed significant phylogenetic underdispersion of first-time zOTU detections (Fig. 2). This means that when a zOTU was detected for the first time in one of these communities, it was more likely to be phylogenetically similar to a zOTU that had previously been detected in community. For both the male and female subject,  $D$  estimates were lower (more underdispersed) in the feces than in the palms, left and right palm  $D$  estimates were similar to each other, and tongue  $D$  estimates were higher. All sites except the tongue showed statistically significant underdispersion in both subjects, while tongue data were not significantly different than the neutral model. The results for the "individual null" were the same. In the comparison between "meta" and "self" models, "meta" models needed to be much more underdispersed than "self" in order to approximate empirical phylogenetic diversity accumulation (Fig. S2). We also observed a general upward trend in  $D$  in our sliding window analysis of the male right palm dataset (Fig. S3), although this trend was only observed over 19 days.

## Results from infant gut data

Empirical phylodiversity accumulation in the infant gut microbiome [26] showed a sharp increase in phylodiversity after day 161 (Fig. 3), the same date that the subject began consuming baby formula. This suggests that baby formula changed the phylogenetic colonization patterns of the developing infant gut. We analyzed this dataset as two separate time-series, one before formula use and one during, and both had negative  $D$  estimates, with the pre-formula  $D$  estimate being lower (Fig. 4). While the pre-formula dataset was significantly underdispersed ( $P = 0.007$ ), the formula dataset was not significantly different from the neutral model, although this result is marginal ( $P = 0.107$ ). When the "individual null" was used instead,

both were significantly underdispersed. Infant gut data from Finnish infants [21] were sampled at a much lower temporal resolution, and as such were not split between formula use. 31 out of 33 individuals analyzed exhibited significant underdispersion, and the other two were not significantly different from the neutral model. Both nonsignificant individuals were from the group treated with heavy antibiotics, but even so, no significant difference in  $D$  values was detected between antibiotics and control groups (Fig. S4). However, when the "individual null" was used, all subjects exhibited significant underdispersion. Estimates of  $D$  did not significantly correlate with the number of zOTUs in a dataset, the total phylodiversity of the dataset, the initial phylodiversity of the dataset, or the number of samples in a dataset (Fig. S5).

## Discussion

Any organism of interest in a human microbiome dataset, from the pathogenic to the probiotic, will at some point be recruited for the first time, and the order in which these organisms are detected in the community is determined by community assembly processes [1]. Predicting which lineages of organisms can be recruited into a given environment has far-reaching implications for ecosystem remediation and management, especially in microbial communities where the medical and ecological importances of many microbes are still largely unknown [29; 30]. Identifying conditions under which assembly mechanisms change, or under which non-neutral assembly is particularly strong, may facilitate microbial community rehabilitation by understanding when and how microbial communities can be colonized by close/distant relatives. We found that assembly during primary succession of the infant gut (Fig. 4, Fig. S4) and during turnover of the microbial communities on the adult palms and gut (Fig. 2) follows a predictable pattern: new species are more likely to be recruited if a close relative has been recruited previously.

The generally "nepotistic" pattern we observed supports our underdispersion hypothesis, which follows Darwin's pre-adaptation hypothesis [11] and more recent ecological theory as well [18; 19]. Much work in phylogenetic community ecology posits that competition tends to be strongest among closely-related species due to phylogenetic niche conservatism [8], so many closely-related species are able to coexist in a community, competition must not be an important factor structuring that community [20]. However, strong competition between distantly related species may actually cause groups of phylogenetically similar species to coexist, especially when immigration is high [18; 19; 31]. This type of competition is perhaps better conceptualized as selection (*i.e.* environmental filtering) instead [19], especially since studies showing evidence for competitive exclusion in microbial communities focus on competition between closely-related species [13; 2].

our model investigates the extent to which newly recruited species are likely to be similar to previously recruited close relatives, but "previously recruited" may include a significant time span. Thus, the observation of underdispersion may not reflect a lack of importance of competition between close relatives *per se*. However, asking whether new species recruitment is likely after recruitment of a close relative has relevance; for instance in human microbiome systems it may be beneficial to understand if a pathogen's probability of recruitment may be higher if a conspecific strain was previously observed [14; 16]. Approaches that consider only recent community membership may more directly inform hypotheses regarding direct competition between close relatives, or regarding more recent recruitment of close relatives. For this reason, we included a sliding-window analysis of 5-day intervals for a subset of intensively-sampled data, and showed significant underdispersion in a majority of windows analyzed (Fig. S3). This type of analysis can satisfy the issue of recency when using our model, but only when data collection is sufficiently frequent.

Regardless, non-neutral patterns of phylogenetic community structure have been interpreted to mean that traits are under ecological selection [32; 20; 33; 34]. If traits are not driving community assembly [35] or if the traits driving community assembly are largely horizontally transferred between taxa independent of their relatedness (as estimated by a 16S rDNA phylogeny), we would expect no phylogenetic signature, and a  $D$  estimate that is not significantly different from 0 (the neutral model). Instead, we observed a very strong and significant phylogenetic signal in species recruitment order for almost all datasets we analyzed. However, if selection on traits is driving this pattern, selection itself may not occur within the host environment. An alternative explanation for the underdispersion we observed is that selection is external to the host environment (*i.e.* selection occurs within the neighboring species pool from which emigration occurs), causing change in the community entering the host to already be underdispersed. Similarly, phylogenetic dispersion of community structure has been unable to distinguish between selection and differences in migration rates [36], so a pre-underdispersed community entering the host is a plausible mechanism for phylogenetic under-

dispersion of species recruitment. But selection of microbial communities within the host has been shown by multiple studies [37; 38; 39], so it is our opinion that selection within the host is a more likely scenario.

A similar point is that species compositions of these datasets do not reflect the actual species pool from which they originate, which is likely not static over time. Our model draws species from the pool of all species observed over the course of a time-series, and with that model we ask whether the order in which those species were recruited is consistent with under- or overdispersion hypotheses. It is for this reason that we discuss our model in the context of the order of recruitment, regardless of whether selection took place in the host or was external as described above. A null model that accounts for a dynamic species pool that is external to the host could be used with the type of analysis we present here, in order to understand if the processes underlying over-/underdispersion take place in the host or in the host's environment. Our "meta vs. self" comparison (Fig. S2) uses the combined habitat-specific microbiomes of two co-habiting individuals as a broader species pool, and we found even stronger underdispersion when empirical "self" patterns were compared to the broader "meta" species pool, suggesting strong selection within the host environment. Still, this broader species pool is not a substitute for a thorough inventory of the real species pool, which would be required to ascertain whether over-/underdispersion originated in the host or elsewhere. As such, future studies may wish to intensively sample a subject's environment in order to better catalog the species pool, and perhaps use qPCR to obtain environmental concentrations of species within the species pool [40].

As to why no datasets analyzed showed significant phylogenetic overdispersion ( $D > 0$ ), we are not certain. At the beginning of development of this model, we expected microbial communities in the human microbiome to follow the overdispersion hypothesis, partly from microbiome studies suggesting competition among closely-related bacteria is an important factor in human gut microbial community assembly [41; 42; 1], and also because of work in experimental microcosms [13]. However, the human microbiome environments analyzed here are environments that undergo constant physical disturbance, unlike aqueous microcosms. Palm communities are physically disturbed with every use of the hands, and by the sampling procedure itself. Gut (fecal) communities are also disturbed constantly by the movement of feces through the gut. It may be possible that continuous disturbance allows for underdispersion via constant re-assembly of communities. In this case, niches may be filled by random "winners" after each disturbance, as in a competitive lottery scenario [4]. These "winners" would still need to be pre-adapted to their environment, so they would be more likely to be closely related to previous "winners", as in our findings. Similarly, environments with fluctuating resource profiles may result in clusters of organisms occupying the same niche [43]. The datasets we used are also somewhat limited in terms of phylogenetic resolution, as short reads of the 16S marker gene are insufficient to detect strain-level variation [44; 42; 15]. Thus, competitive exclusion could occur at the extreme tips of the bacterial phylogenetic tree, and this would not be detectable using 16S rDNA data. Even so, broader patterns of underdispersion at phylogenetic depths accessible with 16S data could still result in significantly underdispersed model fits.

A strength of our model is that it estimates values of  $D$  that can be compared among datasets (Fig. 2) or potentially across time (Fig. 4, Fig. S3) in order to learn how differences between datasets impact community assembly. We found that gut and palm communities were almost universally underdispersed (Fig. 2, Fig. 4, Fig. S4), and that the  $D$  value for a community appears to be a function of body site (Fig. 2). Although this result is only shown across two subjects, the parallel patterns between the male and female subject are striking, in that fecal communities are the most strongly underdispersed (lowest  $D$ ), palm communities are similar to each other, and tongue communities had the highest  $D$  estimates. Similarly, comparing  $D$  before and after an event can be used within an experimental framework to see how that event may affect community assembly. Our analysis of infant gut microbiome data [26] before and during the use of baby formula (Fig. 4) showed that while the pre-formula community was significantly underdispersed, community assembly during formula consumption was more neutral. While the post-formula trend was not significantly different from the neutral model, this finding was marginal ( $P = 0.107$ ).

In addition to showing that our model can be a useful tool for future studies, our findings also hint that phylogenetic underdispersion may be a common trend for the human gut microbiome, although demonstrating a general trend would require analysis of more than the 36 individuals we analyzed. Indeed, recent research has shown that for fecal transplants, donor strains are able to integrate into the recipient's gut community when a conspecific strain is already present, but novel donor strains are unlikely to successfully integrate into the recipient [14]. Congeneric bacteria have also been shown to be predictors of each others' recruitment in the mouse gut microbiome, both for pathogens and commensals [16]. Different body sites - as we saw with the



skin – may have qualitatively similar patterns of underdispersion, yet quantitatively different magnitudes of this effect. Thus the efficacy of an engineered probiotic based on similarity to organisms already present in the community for which it was engineered may largely depend on the body site for which it’s intended, although again more exhaustive study is needed. To facilitate further discovery both in the human microbiome and in other environments, we have made our R code and a tutorial available on GitHub: [https://github.com/darcyj/pd\\_model](https://github.com/darcyj/pd_model).

## Acknowledgements

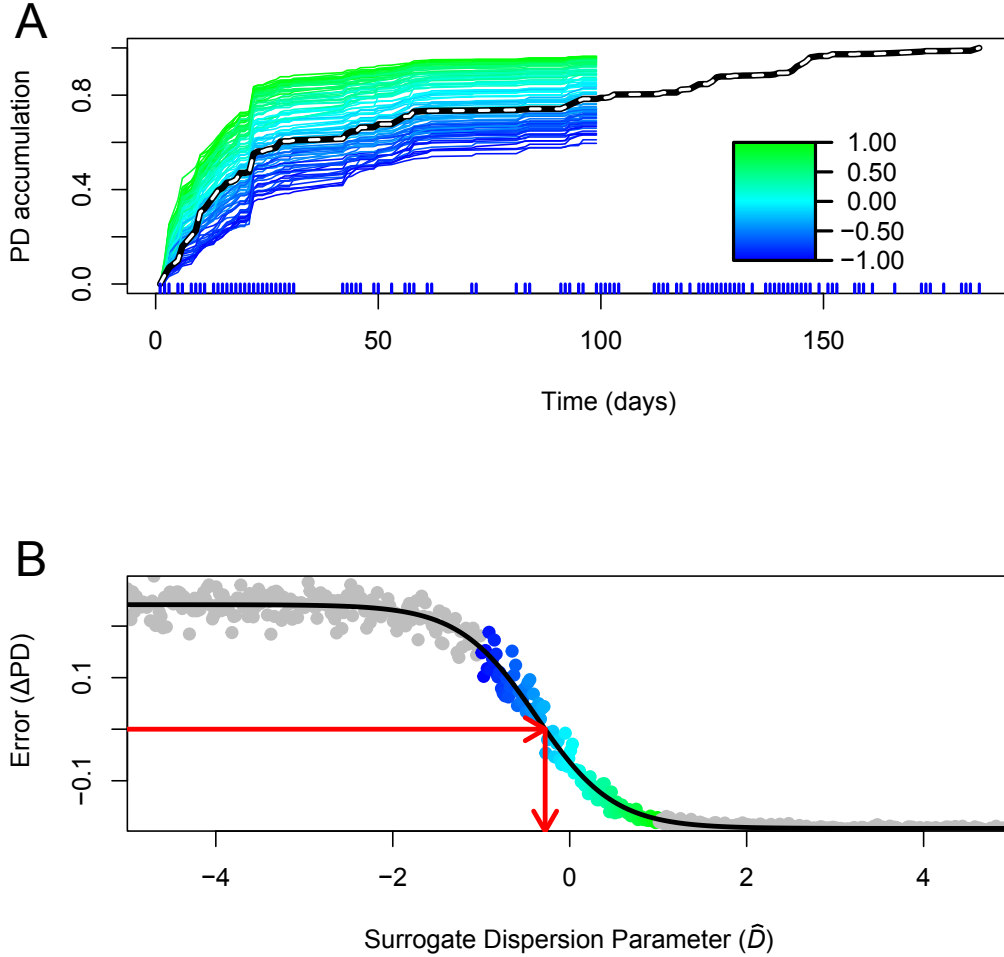
The authors thank D.R. Nemergut for her help and support, and also thank P. Sommers, E. Gendron, A. Solon, E. Pruesse, A. Armstrong, C. Martin, K. Hazleton, and S. Sauce for many helpful discussions. Funding was provided by an NSF grant for studying microbial community assembly following disturbance (DEB-1258160) and by a NIH NLM Computational Biology training grant (5 T15 LM009451-12).

## Conflict of Interest

The authors declare that no conflict of interest exists.

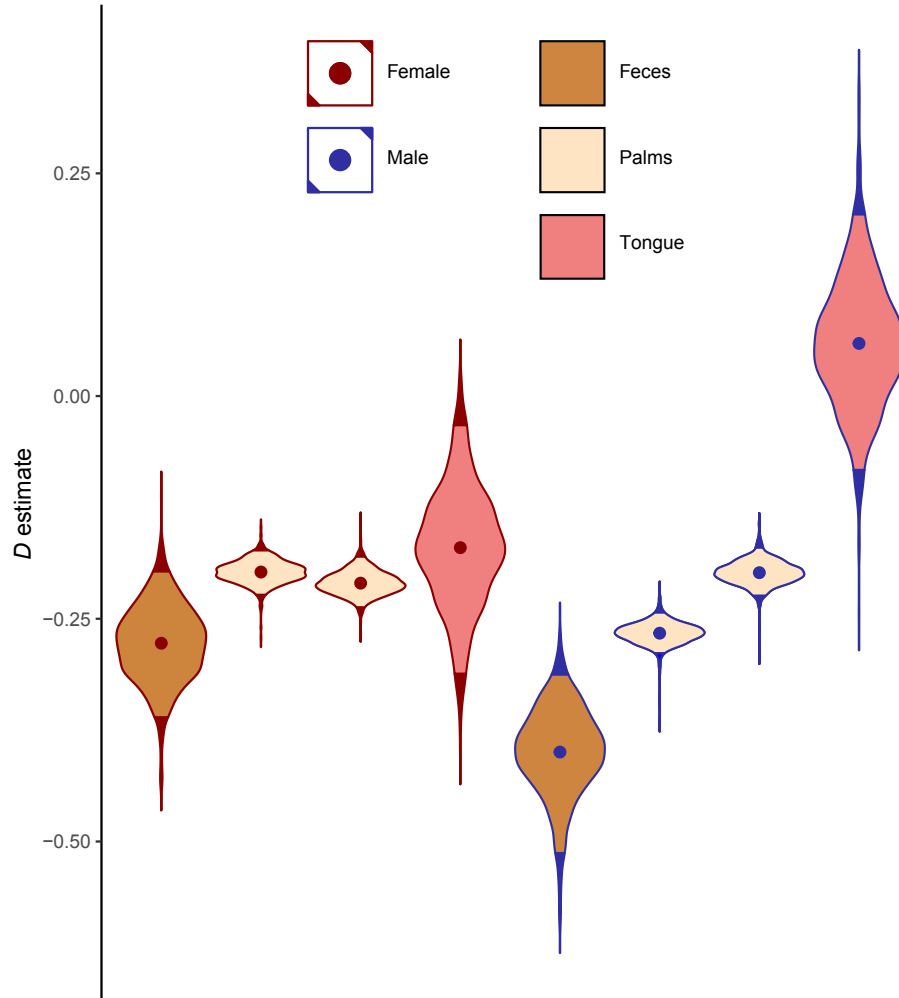
## Figures

Fig. 1



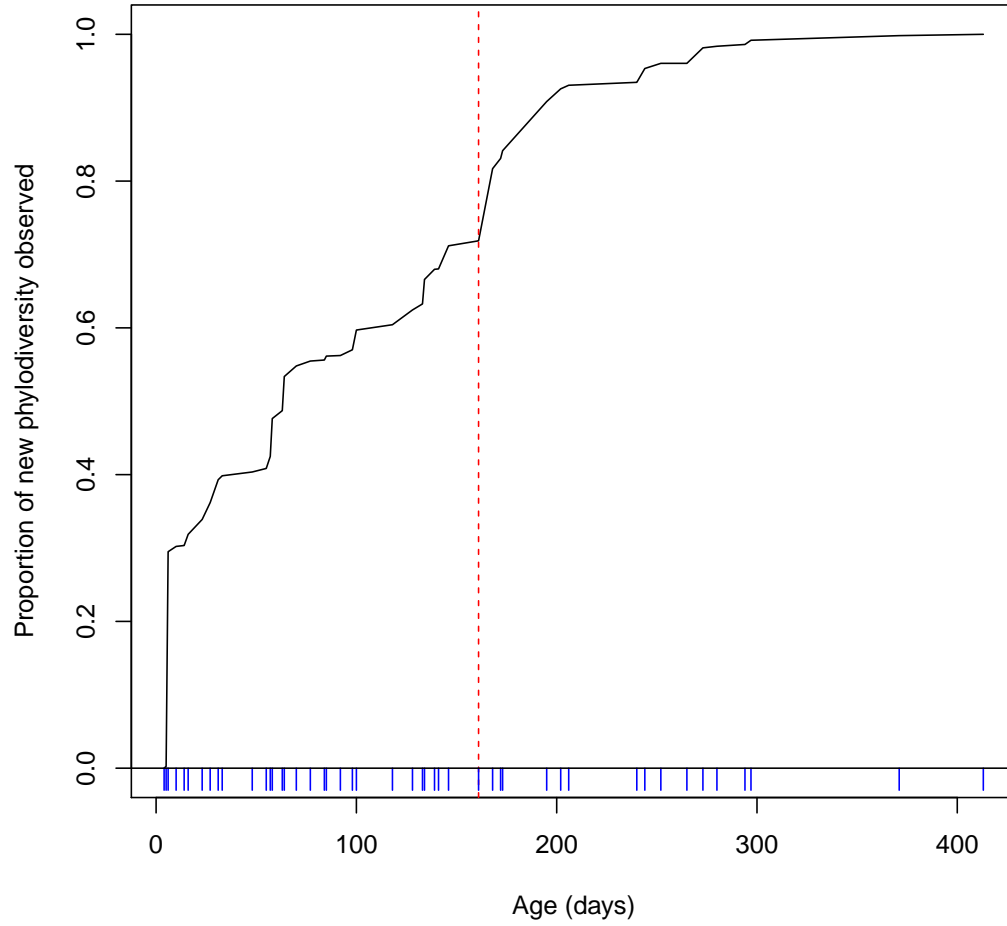
Phylodiversity accumulation and model fitting in the female feces dataset [25]. Plot A shows empirical (dashed) and surrogate phylodiversity accumulation curves. Surrogate curves are colored according to  $\hat{D}$  value (Equation 1). New species that have a previously-detected close relative contribute little phylodiversity and cause slow phylodiversity accumulation (blue). New species that do not have a close relative contribute more phylodiversity and cause faster accumulation (green). The empirical model (dashed) is below the neutral model (teal), signifying underdispersion in the order of first-time species detections. The times of sampling points are shown as vertical blue lines below the X-axis. Curves are rescaled from 0 to 1 in this figure. Plot B shows how empirical and surrogate data are compared to generate an estimate for  $D$ . Differences between empirical and surrogate data at time  $m$  are shown on the Y-axis, and the  $\hat{D}$  values used to generate surrogate datasets are shown on the X-axis. Color-coded points correspond to surrogate datasets shown in plot A. Values shown in gray result from using extreme values of  $\hat{D}$ , which help the logistic error model (black line) fit to the data, and are not shown in plot A. The red arrows show the process of error minimization, yielding a  $D$  estimate. A figure showing significance testing for these data is available as Fig. S1.

408 **Fig. 2**



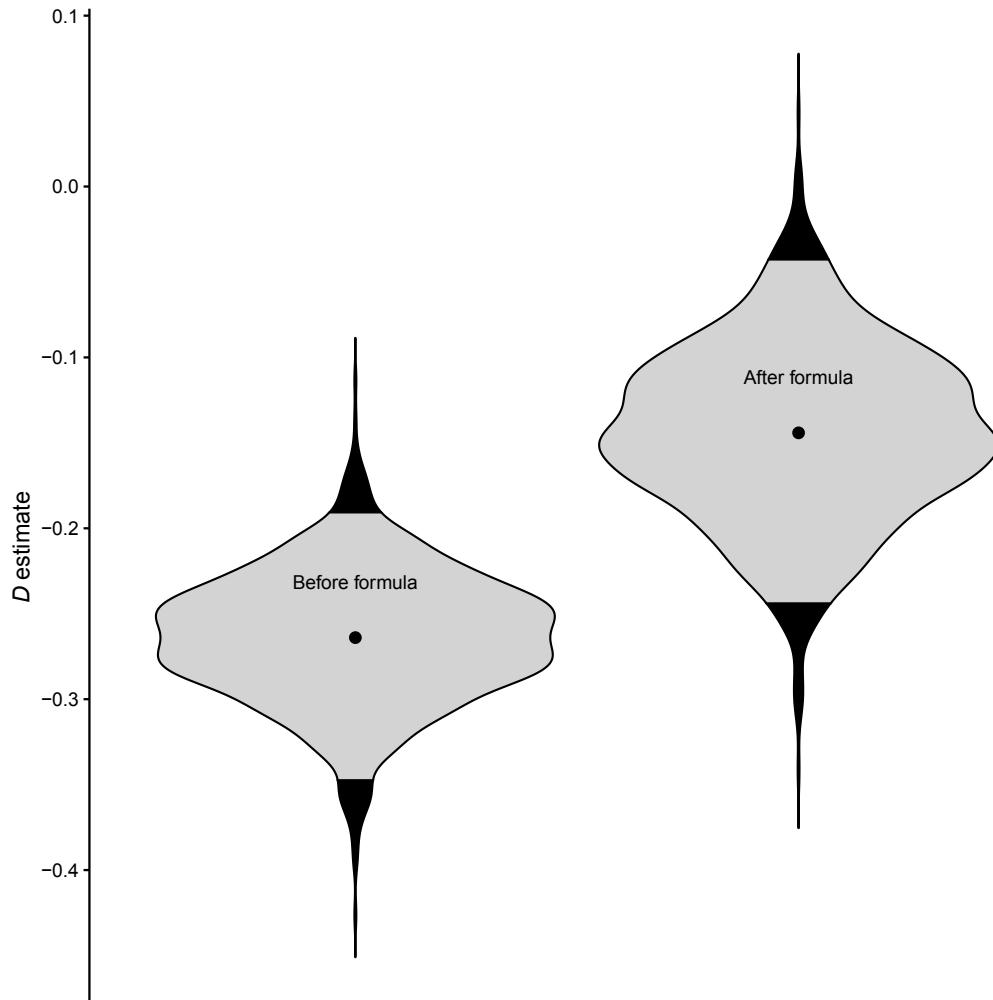
409 Dispersion parameter ( $D$ ) estimates for "moving pictures" [25] datasets. The subject's sex is shown as the  
 410 outline color of each violin, and the body site is shown as fill color. The four body sites for the female  
 411 subject are shown at left, and the four body sites for the male subject are shown at right. Each violin  
 412 shows the distribution of  $D$  estimates given by logistic error model bootstraps, and the dots within violins  
 413 are means. Light-colored portions of violins represent 95% of bootstraps. The two subjects analyzed show  
 414 parallel  $D$  estimates, with feces being the lowest, followed by palms which are all similar, followed by tongue  
 415 communities. For both subjects, tongue patterns were not significantly different than the neutral model.

416 **Fig. 3**



417 Empirical phylodiversity accumulation in the infant gut microbiome [26]. Phylodiversity increases sharply  
418 after day 161 of the infant's life, then plateaus. This timing coincides with the day the subject began  
419 consuming baby formula. The times of sampling points are shown as vertical blue lines below the X-axis.

420 **Fig. 4**



421 Dispersion parameter ( $D$ ) estimates in the infant gut, pre-formula and during formula use. Formula use began  
 422 on day 161, thus the first 160 days of the subject's life were analyzed separately. Community assembly was  
 423 significantly underdispersed in the pre-formula dataset, but was not significantly different from the neutral  
 424 model during formula use ( $P = 0.107$ ).

## References

- [1] Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, et al. Patterns and processes of microbial community assembly. *Microbiology and molecular biology reviews* : MMBR. 2013 sep;77(3):342–56.
- [2] Sprockett D, Fukami T, Relman DA. Role of priority effects in the early-life assembly of the gut microbiota; 2018.
- [3] Fukami T. Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and Priority Effects. *Annual Review of Ecology, Evolution, and Systematics*. 2015;.
- [4] Verster AJ, Borenstein E. Competitive lottery-based assembly of selected clades in the human gut microbiome. *Microbiome*. 2018;.
- [5] Litvak Y, Bäumlér AJ. The founder hypothesis: A basis for microbiota resistance, diversity in taxa carriage, and colonization resistance against pathogens. *PLOS Pathogens*. 2019 feb;15(2):e1007563.
- [6] Zaneveld JR, Lozupone C, Gordon JI, Knight R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Research*. 2010;38(12):3869–3879.
- [7] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*. 2013 aug;31(9):814–821.
- [8] Wiens JJ, Ackerly DD, Allen AP, Anacker BL, Buckley LB, Cornell HV, et al. Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*. 2010 oct;13(10):1310–1324.
- [9] Pyron RA, Costa GC, Patten MA, Burbrink FT. Phylogenetic niche conservatism and the evolutionary basis of ecological speciation. *Biological Reviews*. 2015;.
- [10] Losos JB. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species; 2008.
- [11] Darwin C. *On the Origin of Species*, 1859. London: Murray; 1859.
- [12] Ma C, Li Sp, Pu Z, Tan J, Liu M, Zhou J, et al. Different effects of invader–native phylogenetic relatedness on invasion success and impact: a meta-analysis of Darwin's naturalization hypothesis. *Proceedings of the Royal Society B: Biological Sciences*. 2016 sep;283(1838):20160663.
- [13] Peay KG, Belisle M, Fukami T. Phylogenetic relatedness predicts priority effects in nectar yeast communities. *Proceedings of the Royal Society B: Biological Sciences*. 2012;.
- [14] Li SS, Zhu A, Benes V, Costea PI, Hercog R, Hildebrand F, et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science (New York, NY)*. 2016 apr;352(6285):586–9.
- [15] Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, et al. The *Prevotella copri* complex comprises four distinct clades that are underrepresented in Westernised populations. *bioRxiv*. 2019;.
- [16] Stecher B, Chaffron S, Käppeli R, Hapfelmeier S, Friedrich S, Weber TC, et al. Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS pathogens*. 2010 jan;6(1):e1000711.
- [17] Brown CT, Xiong W, Olm MR, Thomas BC, Baker R, Firek B, et al. Hospitalized Premature Infants Are Colonized by Related Bacterial Strains with Distinct Proteomic Profiles. *mBio*. 2018 may;9(2):e00441–18.
- [18] D'Andrea R, Riolo M, Ostling AM. Generalizing clusters of similar species as a signature of coexistence under competition". *PloS Computational Biology*. 2019 jan;15(1).

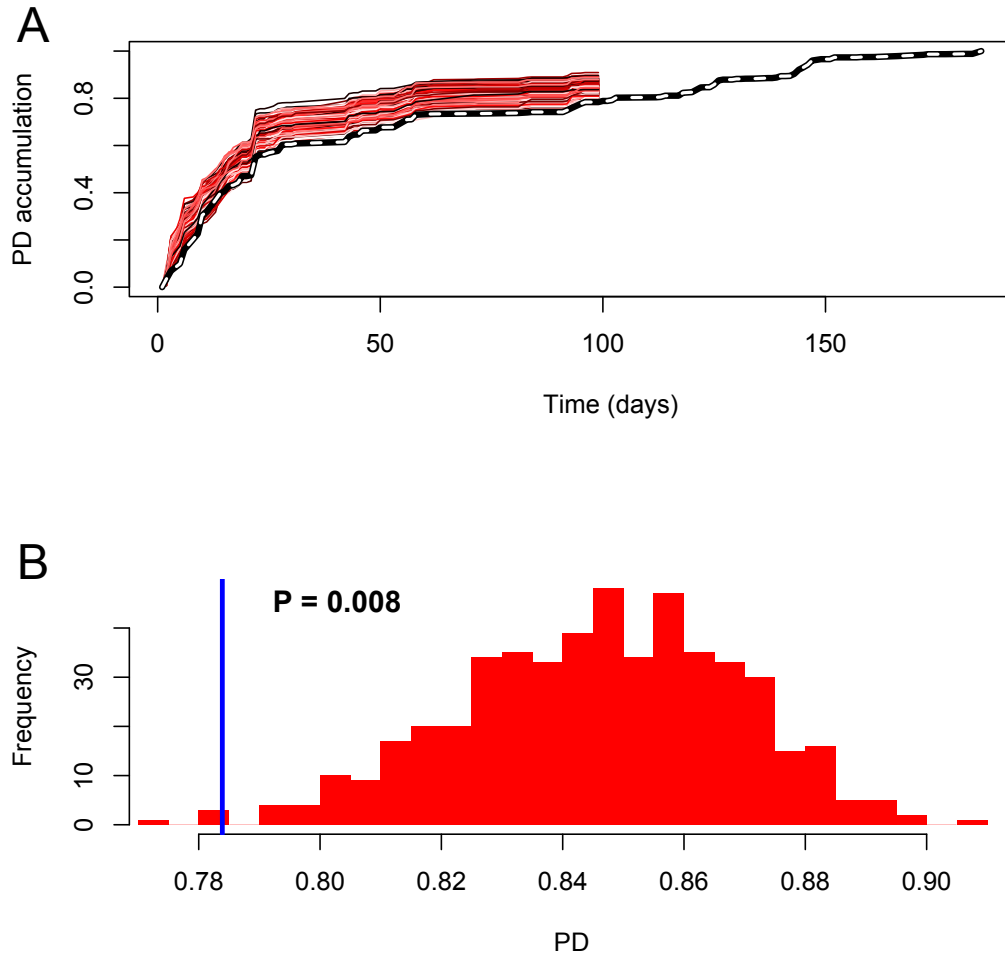
- [19] Mayfield MM, Levine JM. Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecology Letters*. 2010;.
- [20] Webb CO, Ackerly DD, McPeck MA, Donoghue MJ. Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*. 2002 nov;33(1):475–505.
- [21] Yassour M, Vatanen T, Siljander H, Hämäläinen AM, Härkönen T, Ryhänen SJ, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine*. 2016;.
- [22] Lozupone C, Knight R. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and environmental microbiology*. 2005;71(12):8228–8235.
- [23] Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, et al. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*. 2017 feb;5:e2969.
- [24] Matsen FA, Evans SN, Gilks W, Ghodsi M, Kingsford C. Edge Principal Components and Squash Clustering: Using the Special Structure of Phylogenetic Placement Data for Sample Comparison. *PLoS ONE*. 2013 mar;8(3):e56859.
- [25] Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. *Genome biology*. 2011 jan;12(5):R50.
- [26] Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*. 2011 mar;108 Suppl(Supplement\_1):4578–85.
- [27] Faith DP. Conservation evaluation and phylogenetic diversity. *Biological Conservation*. 1992 jan;61(1):1–10.
- [28] Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. 2016; Available from: <http://www.biorxiv.org/content/early/2016/10/15/081257>.
- [29] Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A phylogenetic perspective. *Science*. 2015 nov;350(6261):aac9323–aac9323.
- [30] Vázquez-Baeza Y, Callewaert C, Debelius J, Hyde E, Marotz C, Morton JT, et al. Impacts of the Human Gut Microbiome on Therapeutics. *Annual Review of Pharmacology and Toxicology*. 2018;.
- [31] Scheffer M, van Nes EH. Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences*. 2006;103(16):6230–6235. Available from: <https://www.pnas.org/content/103/16/6230>.
- [32] Webb CO. Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. *The American naturalist*. 2000 aug;156(2):145–155.
- [33] Cavender-Bares J, Ackerly DD, Baum DA, Bazzaz FA. Phylogenetic Overdispersion in Floridian Oak Communities. *The American Naturalist*. 2004;.
- [34] Gerhold P, Cahill JF, Winter M, Bartish IV, Prinzing A. Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Functional Ecology*. 2015 may;29(5):600–614.
- [35] Hubbell SP. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press; 2001.
- [36] Emerson BC, Gillespie RG. Phylogenetic analysis of community assembly and structure over space and time. *Trends in ecology & evolution*. 2008 nov;23(11):619–30.

- [37] Peterfreund GL, Vandivier LE, Sinha R, Marozsan AJ, Olson WC, Zhu J, et al. Succession in the gut microbiome following antibiotic and antibody therapies for *Clostridium difficile*. *PloS one*. 2012 jan;7(10):e46966.
- [38] David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. *Genome biology*. 2014 jan;15(7):R89.
- [39] Kennedy RC, Fling RR, Robeson MS, Saxton AM, Donnell RL, Darcy JL, et al. Temporal Development of Gut Microbiota in Triclocarban Exposed Pregnant and Neonatal Rats. *Scientific reports*. 2016;6:33430.
- [40] Williamson BD, Hughes JP, Willis AD. A multi-view model for relative and absolute microbial abundances. *bioRxiv*. 2019;Available from: <https://www.biorxiv.org/content/early/2019/09/08/761486>.
- [41] Chatzidaki-Livanis M, Geva-Zatorsky N, Comstock LE. *Bacteroides fragilis* type VI secretion systems use novel effector and immunity proteins to antagonize human gut Bacteroidales species. *Proceedings of the National Academy of Sciences*. 2016;113(13):3627–3632.
- [42] Hecht AL, Casterline BW, Earley ZM, Goo YA, Goodlett DR, Bubeck Wardenburg J. Strain competition restricts colonization of an enteric pathogen and prevents colitis. *EMBO reports*. 2016;17(9):1281–1291.
- [43] Sakavara A, Tsirtsis G, Roelke DL, Mancy R, Spatharis S. Lumpy species coexistence arises robustly in fluctuating resource environments. *Proceedings of the National Academy of Sciences*. 2018;115(4):738–743. Available from: <https://www.pnas.org/content/115/4/738>.
- [44] Morowitz MJ, Denef VJ, Costello EK, Thomas BC, Poroyko V, Relman DA, et al. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proceedings of the National Academy of Sciences*. 2011;108(3):1128–1133.



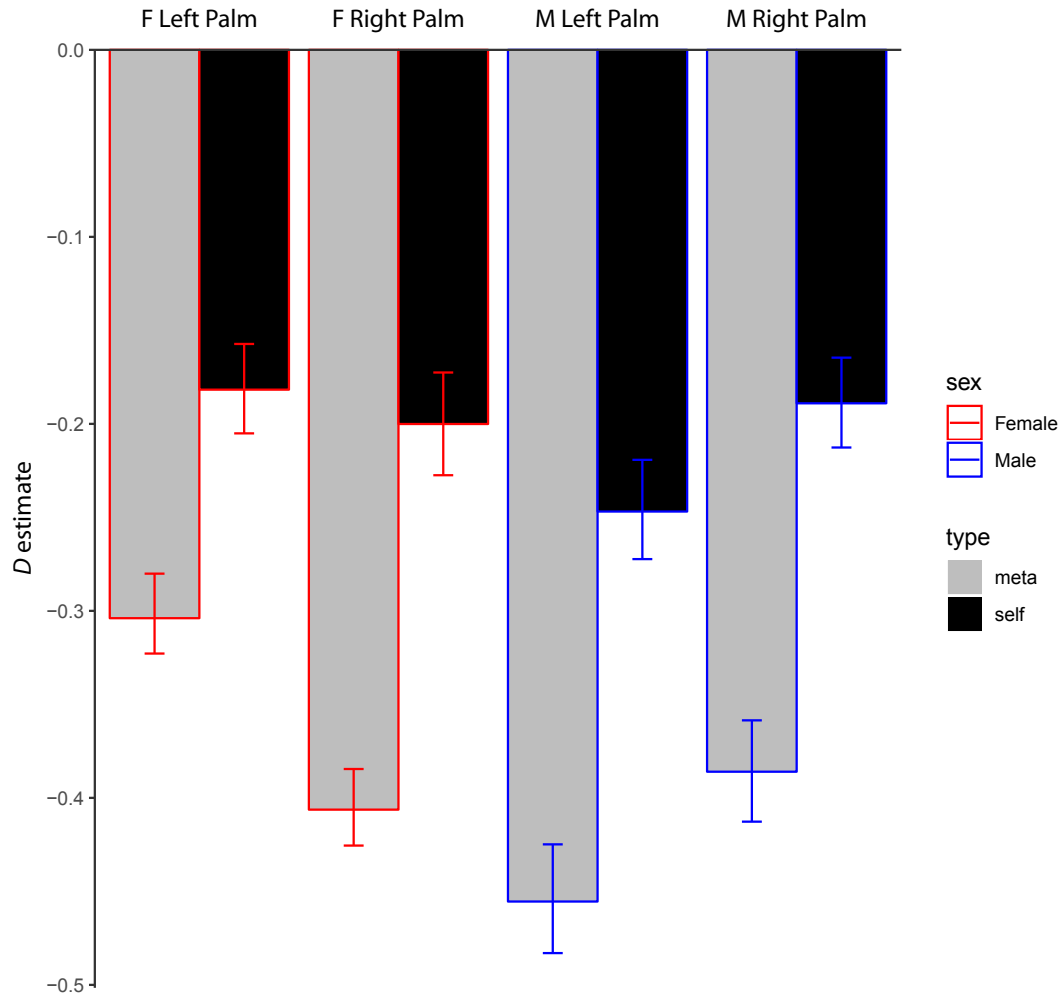
## Supplementary Material

Fig. S1



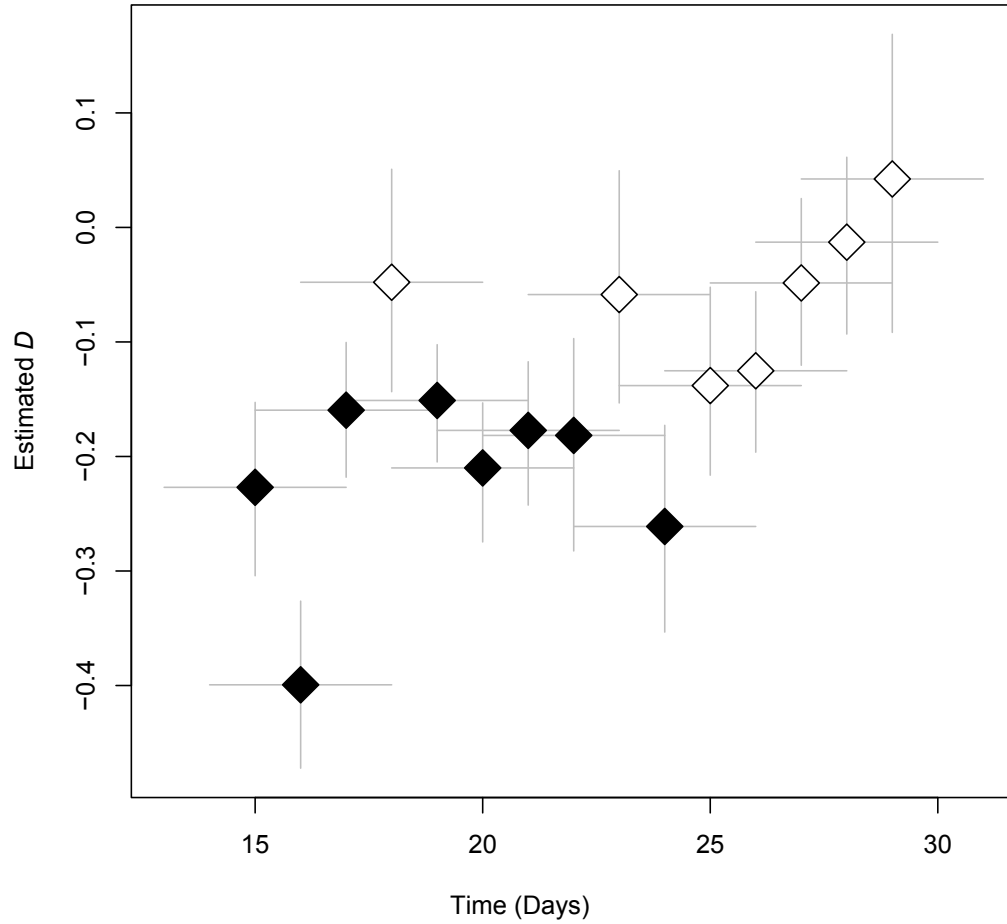
Significance testing for the female feces dataset. Plot A shows the empirical phylogenetic diversity accumulation (dashed; same as Fig. 1A) but with neutral model surrogate datasets shown in different shades of red. These are produced by running the neutral model 500 times, to generate a distribution of phylogenetic diversity values under  $D = 0$  (Plot B). As with all surrogate datasets, these are run until time  $m$  (see Parameter Estimation section of Materials and Methods). Empirical phylogenetic diversity at time  $m$  (blue line) is compared to the distribution of neutral model phylogenetic diversities at time  $m$  (red histogram), and a  $P$ -value is calculated as the proportion of neutral phylogenetic diversities more extreme than the empirical value.

540 **Fig. S2**



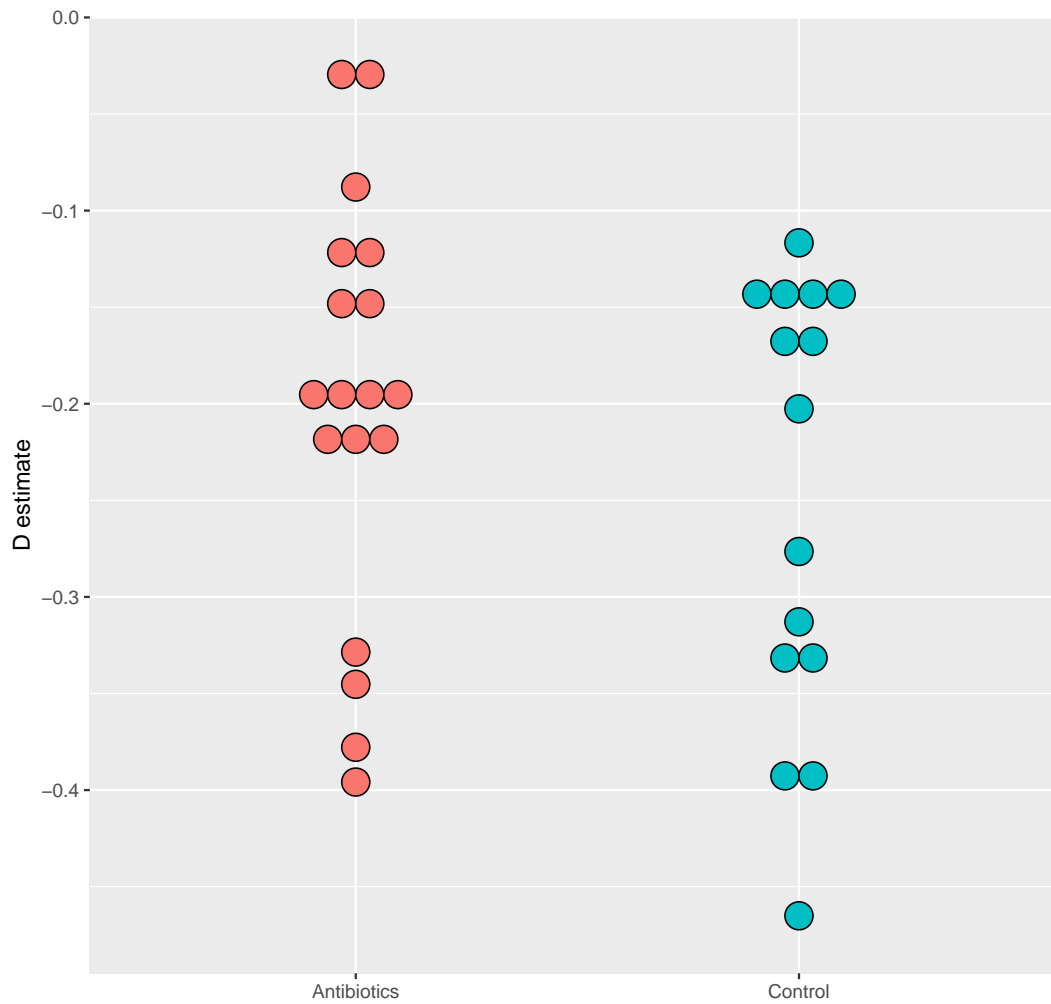
541 Comparison of "self" vs "meta" model results from palm communities. "Self" (black) models were run  
 542 identically to Fig. 2), but "meta" (gray) models were run where the species pool for each palm community  
 543 surrogate dataset was composed of all zOTUs observed across all four palm datasets. The difference between  
 544 the "self"  $D$  estimate (generated above) and the "meta"  $D$  estimate (estimated with a metapopulation of  
 545 zOTUs) is related to the exclusivity of recruitment into the community. In other words, if we were to estimate  
 546 similar  $D$  values for both the "meta" and "self" analyses, the inclusion of extra species in the species pool  
 547 would be of little importance to the model, and we would learn that it would make little difference to  
 548 community assembly patterns if the species pool really was composed of the "meta" set.

549 **Fig. S3**



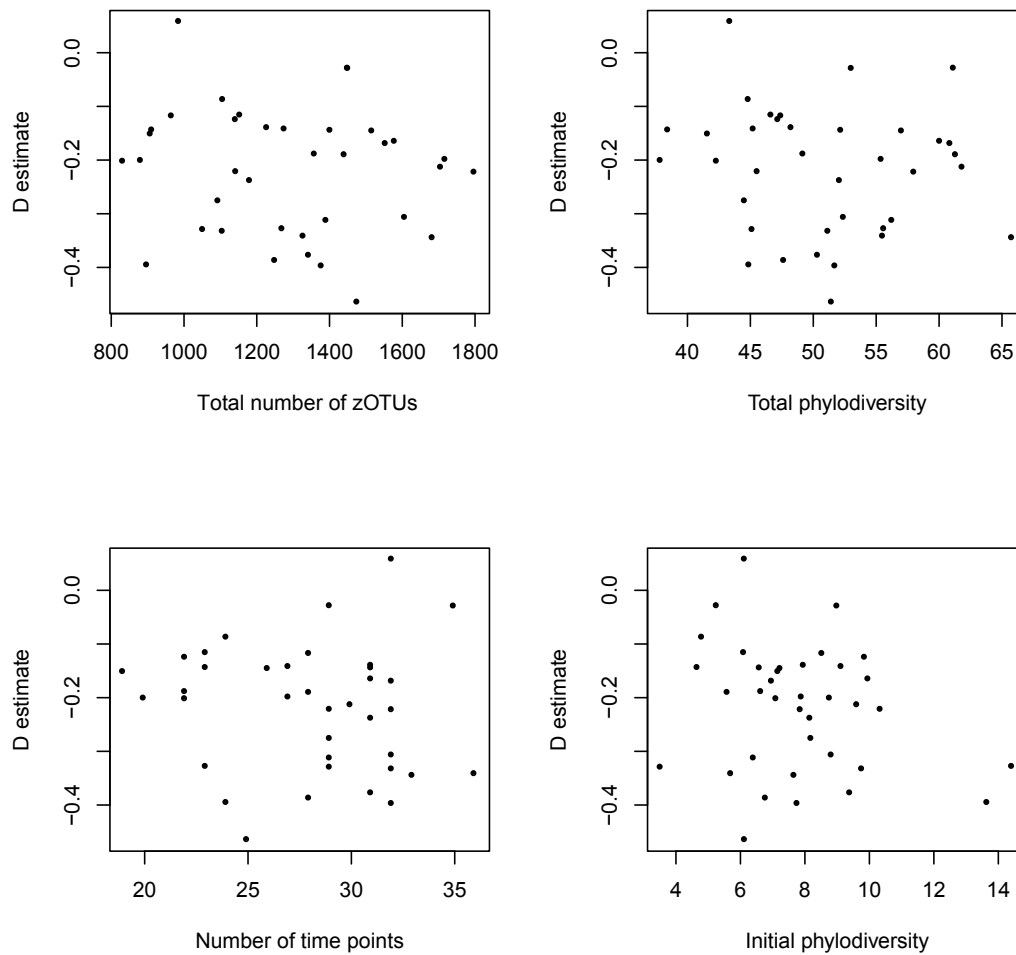
550 Sliding window analysis of male right palm data over 19 consecutive samples. We ran our model on each  
551 window of 5 continuous days (15 windows), in order to see how  $D$  varied over time. We only conducted this  
552 analysis for the section of samples that were sampled every day, so that comparisons between windows would  
553 not be confounded by window size. This analysis was done to demonstrate a potential use case for our model,  
554 and not to test any specific hypothesis. Filled shapes represent windows that were significantly different than  
555 the neutral model. Vertical bars represent 95% confidence intervals for  $D$  estimate, and horizontal bars  
556 represent window size.

557 **Fig. S4**



558 *D* estimates of Finnish infant datasets. All but two subjects exhibited significant phylogenetic underdis-  
559 persion. The two subjects that were not significantly different from the neutral model were both in the  
560 antibiotics cohort, which is comprised of infants that were treated with frequent antibiotics, almost all for  
561 ear infections. There was no significant difference between *D* values for the two groups.

562 **Fig. S5**



563 Relationship of  $D$  estimate to total zOTU richness, total phylodiversity, number of timepoints sampled, and  
 564 initial phylodiversity (of first sample) for Finnish infant data. No statistically significant correlation was  
 565 detected in any of these four analyses.