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[1], Alex D. Washburne[2], Michael S. Robeson[3], Tiffany Prest[4], Steven K. Schmidt[4], Catherine A. Lozupone[1]

Affiliations

[1] Division of Biomedical Informatics and Personalized Medicine, University of Colorado School of Medicine, Aurora, Colorado, USA.

[2] Department of Microbiology and Immunology, Montana State University. Bozeman, Montana, 59717, USA.

[3] Department of Biomedical Informatics, University of Arkansas for Medical Sciences. Little Rock, Arkansas, 72205, USA.

[4] 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. 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 exclude each other due to high niche overlap. However, several recent human microbiome studies have instead found that close phylogenetic relatives are often detected in microbial communities in short succession, suggesting factors such as shared adaptation to similar environments play a stronger role than competition. To address t[illegible]e developed a mathematical model that describes the probabilities of different species being detecte[illegible] time-series microbiome data, within a phylogenetic framework. We use our model to test three hypothetical assembly modes: underdispersion (species are more likely to be detected if a close relative was previously detected), overdispersion (likelihood of detection is higher if a close relative has not been previously detected), and the neutral model (likelihood of detection is not related to phylogenetic relationships among species). We applied our model to longitudinal high-throughput sequencing data from the human microbiome, and found that for individuals we analyzed, the human microbiome generally follows an assembly pattern characterized by phylogenetic underdispersion (*i.e.* nepotism). Exceptions were oral communities, which were not significantly different from the neutral model in either of two individuals analyzed, 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

Every non-sterile surface in the world is in some stage of community assembly, from a forest of tropical trees to the microbes in a mammalian gut. The communities of organisms inhabiting these environments are dynamic through time, and studying patterns of assembly may shine light on general rules that govern their change. Understanding these community assembly rules may aid habitat restoration [1; 2], the management of ecosystems that have undergone disturbances [3; 4], and ecological theory of phylogenetic signatures in community assembly [5; 6].Patterns and rules of community assembly are particularly important in human systems, including the primary succession of microbes on a human host following birth [7], secondary successions following disease [8; 9], disturbances caused by host lifestyle or antibiotic use [10; 11; 12], and the natural turnover of microbial communities over time [13]. Insights into these difficult-to-observe community assembly processes can be gained via the comparison of microbial communities using high-throughput DNA sequencing [13; 14; 15], especially in longitudinal (time-series) studies [13; 7; 11].

A central question in microbial community assembly is when and why microbes are recruited into communities. The actual process of a species recruitment is not directly observed in studies using high-throughput DNA sequencing; instead the lack of detection of a species, followed by subsequent detection of that species, is used as a proxy. Recruitment results in the empirical detection of new species, and can be studied by evaluating the order in which species are detected in time-series experiments, given data such as which species have already been detected or what changes occur in an environment over time [14; 16]. 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 [14; 17; 16; 18; 19]. 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 [20; 21].

Within this phylogenetic framework, a predominant hypothesis has been that closely related microbes inhibit each other's successful recruitment [14; 17; 18]. The proposed mechanism for this hypothesis is that closely related microbes likely have similar niches (phylogenetic niche conservatism [22]), and the first microbe to establish into a community will occupy its niche to the exclusion of ecologically similar strains. This is also the basis of Darwin's naturalization hypothesis [23], which proposed that new species are less likely to be recruited if a close relative is present [24]. 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 [25]. In this paper, we refer to the assembly mode 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).

However, overdispersion is far from universal, and multiple studies have shown that extremely close relatives can coexist within the human microbiome [26; 27; 28], and may even be preferentially recruited [29]. This is consistent with simulations showing that clusters of closely-related species can persist despite strong within-cluster competition, when immigration rate is high [30]. 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 [23]. This hypothesis predicts 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). For this reason, we refer to this hypothesis as the **underdispersion hypothesis**. Both this and the overdispersion hypotheses are alternatives to the null hypothesis that recruitment is indipendent 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 [31]. Instead of referring to diversity observed in a single sample, our use of over- and underdispersion refers to the amount of new diversity added 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 detected in a time-series, rather than community composition of any given sample.

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 [32]) or analyzing patterns of select clades within the community (*e.g.* PhyloFactor [33], Edge PCA [34]), we model the probability of detecting new species in a community for the first time as a monotonic function of their phylogenetic distances to members of the community that have already been detected.

The model we present here can be used to estimate the degree to which the detection of new species is more or less likely when a close relative is already present, using empirical data. We fit our model to several time-series human microbiome datasets [13; 7; 35], 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 detection of new species over time is related to the new species' phylogenetic similarity to (or distance from) species that were already detected 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,

$$p_{it} = \frac{d_{it}^{D}}{\sum_{\hat{i}} d_{\hat{i}t}^{D}} \tag{1}$$

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 phylodiversity [36] to compare each of the 500 surrogate datasets (described above) to the empirical dataset. Phylodiversity is the sum of branch-lengths on a phylogenetic tree for a set of species, so phylodiversity 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 [36]. 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,

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

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.

Analysis

This section is a summary of our data analysis. Full detailed methods are available as supplemental information: SUPPLEMENT NAME HERE.

We ran our model on data from 36 individuals from three data sources. Two individuals were from Caporaso *et al.* [13], 33 were from Yassour *et al.* [35], and one was from Koenig *et al.* [7]. In all cases, data were downloaded and processed using the unoise3 pipeline [37], 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" [13] 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.* [35] 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.

Supplemental Analysis

Note: this section will be moved to a supplement soon.

Infant gut 16S rDNA sequencing data from Koenig *et al.* [7] were downloaded from the NCBI Short Read Archive (SRA) website (`http://www.ncbi.nlm.nih.gov/sra`) along with their metadata. These data are a time-series of fecal bacterial communities from an infant subject, over the first 500 days of life. QIIME [38] was used to trim primer regions from these data. Clustering was performed using the unoise3 pipeline [37]; sequences were de-replicated at 100% identity using VSEARCH [39], zOTU centroid sequences were picked and chimeric sequences were removed using unoise3 [37], then all sequences were mapped onto zOTU seeds to create a zOTU table using VSEARCH. zOTU stands for "zero-radius operational taxonomic unit" [37]. Unlike traditional de novo clustered OTUs, zOTUs are exact sequence variants (ESVs) which are consistent and easily comparable across data. The SINA aligner [40] was used to align zOTU centroid sequences to the SILVA SSU Ref 128 database (available from `https://www.arb-silva.de/download/arb-files/`). We then used IQ-TREE [41] to build a phylogenetic tree from the aligned sequences.

The resulting zOTU table was rarefied to 1000 sequences per sample, and samples with fewer sequences were excluded. The last five timepoints were excluded as well because they were sampled at a much lower temporal resolution. This left 52 timepoints spread over the first 469 days of the infant subject's life. The zOTU table was then split into two zOTU tables, one for timepoints before the infant started consuming baby formula, and one for those after. The "pre-formula" zOTU table contained ages 4 days through 146, and the "post-formula" zOTU table contained ages 161 days through 469 days. Each zOTU table was used to run our model as described above using 500 D values (Equation 1), ranging from underdispersed ($D = -5$) to overdispersed ($D = 5$), using zOTUs in lieu of species. zOTUs with zero phylogenetic distance between them were combined, because these zOTUs were uninformative for our statistical model (Equation 1; zero raised to a negative exponent is undefined). The 500 resulting surrogate datasets were compared to the empirical dataset as described above, using difference between phylodiversity values at the middle timepoint. The logistic error model (Equation 2) was fit and bootstrapped, yielding an estimate for D and 95% confidence intervals for that estimate.

"Moving pictures" sequence data from Caporaso *et al.* [13] were downloaded from the MG-RAST database (`http://metagenomics.anl.gov/`). These are longitudinal data from one adult male subject and one adult female subject, over a period of several hundred days, across multiple sample sites (feces, both palms, tongue). timepoints were excluded which did not have sequence data for each of the 8 environments (left palm, right palm, mouth, and feces of the male and female subjects), and rarefied to 5000 sequences per sample. This left 107 timepoints, ranging from day 1 to day 185. Analysis for each dataset (e.g. female right palm) was carried out as described above, except raw sequences were trimmed to a length of 91 bp after the end of the forward PCR primer site in order to ensure that all raw sequences spanned the same region of the 16S rRNA gene. 91 bp was chosen as a length cutoff in order to keep 95% of the sequence data (5% of sequences were discarded because they were shorter).

Analysis of the "moving pictures" 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. In principal, the model may perform differently if a broader representation of what is in the environment is in the species pool compared to what eventually colonizes the individual over time, as the latter may result in a species pool that is overall constrained by factors such as competition for niche space. Thus in an alternate approach we included sequences in the species pool from the other individual living in the same household, as these would be in the environment but not competing for the same niche. We analyzed palm communities in a "meta" context, where surrogate datasets were generated assuming the species pool for a given palm was composed of all four palms in the dataset. In this case, the difference between the "self" D estimate (generated above) and the "meta" D estimate (estimated with a metapopulation of zOTUs) is related to the exclusivity of species detected in the community. In other words, 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, and we would learn that it would make little difference to community assembly patterns if the species pool really was composed of the "meta" set. We also analyzed a section of samples from the male right palm data that were collected every day over a period of 19 days, using a sliding window approach. We ran our model as described above on each window of 5 continuous days (15 windows), in order to see how D varied over time. We only conducted this analysis for the section of samples that were sampled every day, so that comparisons between windows would not be confounded by window size.

Finnish infant sequence data from Yassour *et al.* [35] and associated metadata were downloaded from the DIABIMMUNE Microbiome Project website (`https://pubs.broadinstitute.org/diabimmune`). These are longitudinal gut microbiome data from Finnish infants, collected over the first 36 months of life [35]. Roughly half of these infants were repeatedly treated with oral antibiotics, almost universally for ear infections. Metadata for this dataset were compiled in a different re-analysis of these data [12] and were downloaded from the authors' GitHub page (`https://github.com/ShadeLab/microbiome_trait_succession`). Subject datasets belonging to the groups "Antibiotic" (n=18) or "Control" (n=15) were each analyzed using our model, similar to above. These subjects had between 19 and 36 samples collected over 36 months, with a mean of 28 samples. Sequence data were rarefied to 5000 sequences, and our model was run per above. We compared the estimated D values between antibiotic and control groups using a Mann-Whitney test. Because this dataset had so many subjects, we used this analysis as an opportunity to analyze whether the number of zOTUs, total phylodiversity, or number of timepoints had an effect on estimated D values. This was done via correlation analysis of D estimates with the aforementioned potential covariates.

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`.

Results

By varying $\hat{D}$, we successfully changed the rate at which phylodiversity is added to surrogate (*i.e.* resampled) 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 [13] 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. 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 [7] 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 ($P = 0.107$). Infant gut data from Finnish infants [35] were sampled at a much lower temporal resolution, and as such were not split between formula use. 31 out of 33 individuals analyzed exhibited significantly 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). 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 detected for the first time, and the order in which these organisms are detected in the community is determined by community assembly processes [14]. 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 [42; 43]. 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. 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 design probiotics for colonization, and better exploit disturbance as a tool for managing microbial systems related to human health and disease. 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 detected if a close relative has been detected previously.

We describe new species appearing as ”detections” because of the difference between empirical data and actual phenomena. Species recruitment into communities is a phenomenon under investigation in our model, but evidence for recruitment is a lack of detection, and then subsequent detection of a species using high-throughput DNA sequencing data. With such 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. Furthermore, 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.* [35]). Future work may use techniques such as qPCR to quantify abundances of individual species or strains, and exclude those that do not meet an *a priori* abundance threshold for detection. Nevertheless, in order to be conservative in our language and our approach, we have described our model and our hypotheses in terms of modeling the detection of new species, rather than modeling their recruitment.

The generally “nepotistic” pattern we observed in new species detection supports our underdispersion hypothesis, which follows Darwin’s pre-adaptation hypothesis [23] and more recent ecological theory as well. Briefly, the logic of much work in phylogenetic community ecology was that competition would be strongest among closely-related species due to phylogenetic niche conservatism [44], so communities harboring many closely-related species would be less competitive [31]. However, strong competition between distantly related species may actually cause groups of phylogenetically similar species to coexist, especially when immigration is high [30; 45]. This type of competition is perhaps better conceptualized as environmental filtering instead [45], especially since studies showing evidence for competitive exclusion in microbial communities focus on competition between closely-related species [25; 16].

But competition is not necessarily relevant to our model, especially since our original question was if closely-related species predict each others’ recruitment. Under our implementation, our model is used to investigate the extent to which newly detected species are likely to be similar to previously detected close relatives, but indeed the previously detected species may no longer be present in the community. Thus, competition may not apply to this comparison *per se.* The utility of our model is in testing whether newly detected species are more likely to appear after a close relative (underdispersion hypothesis), or less likely (overdispersion hypothesis). This question has relevance to human microbiome systems especially, where it is beneficial to understand if a pahogen’s probability of detection may be higher if a conspecific strain was previously observed [26; 28]. In our implementation, ”previously observed” may include a significant time span, and although we showed nepotism was strong and statistically significant in these anlayses, such long-term patterns may not be of interest to future studies using our model. For this reason, we included a sliding-window analysis of 5-day intervals for a subset of intensively-sampled data, which 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 [46; 31; 47; 48]. If traits are not driving community assembly [49] 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 detection 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 [50], so a pre-underdispersed community entering the host is a plausible mechanism for phylogenetic underdispersion of species detection. But selection of microbial communities within the host has been shown by multiple studies [10; 9; 11], so it is our opinion that selection within the host is a more likely scenario.

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 [51; 52], and also because of work in experimental microcosms [25]. 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 or Darwin’s pre-adaptation hypothesis 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 [18]. These “winners” would still need to be adapted to their environment, so they would be more likely to be closely related to previous “winners”, as in our findings. 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 [53; 52; 27]. 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 [7] 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 ($P = 0$[illegible]

In addition to showing that our model can be a [illegible] 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 [26]. Congeneric bacteria have also been shown to be predictors of each others' recruitment in the mouse gut microbiome, both for pathogens and commensals [28]. 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`.

Acknowledgements

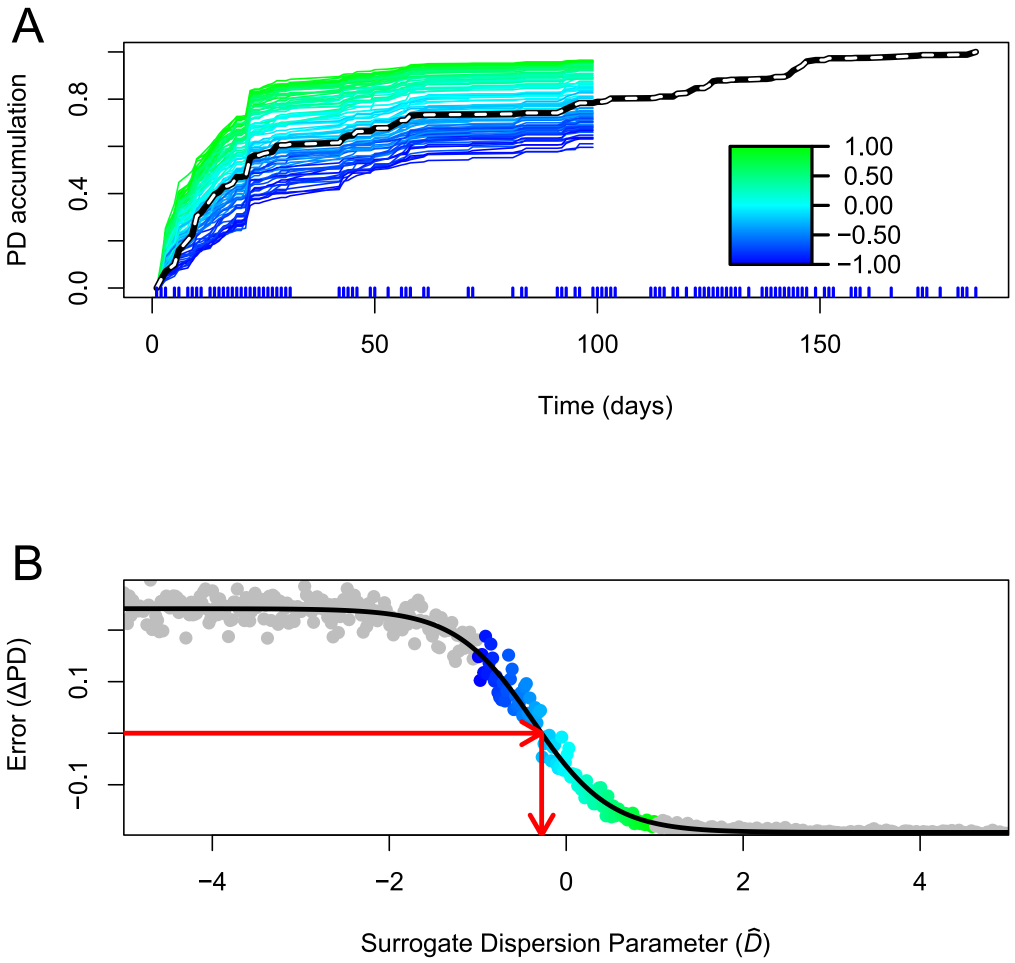
The authors thank D.R. Nemergut for her help and support, and also thank P. Sommers, E. Gendron, A. Solon, E. Pruesse, A. Armstrong, C. Martin, K. Hazleton, and S. Sauce for many helpful discussions. Funding was provided by an NSF grant for studying microbial community assembly following disturbance (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.

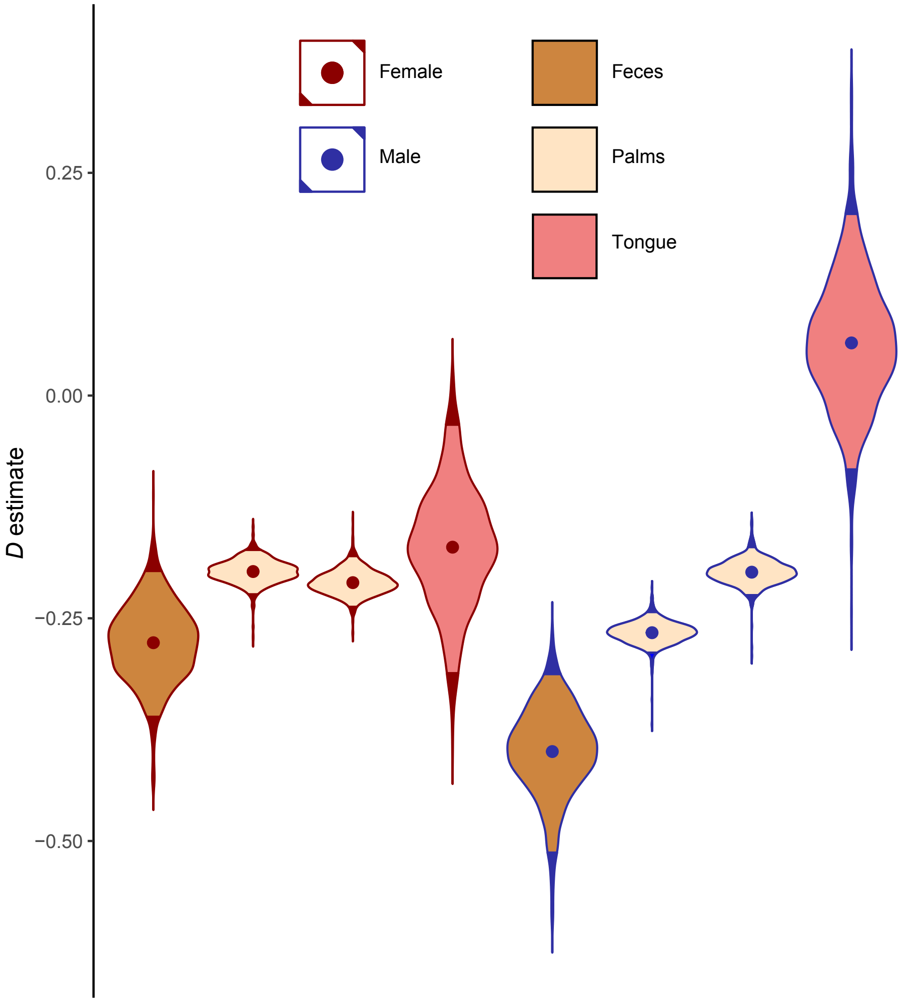
Figures

Fig. 1



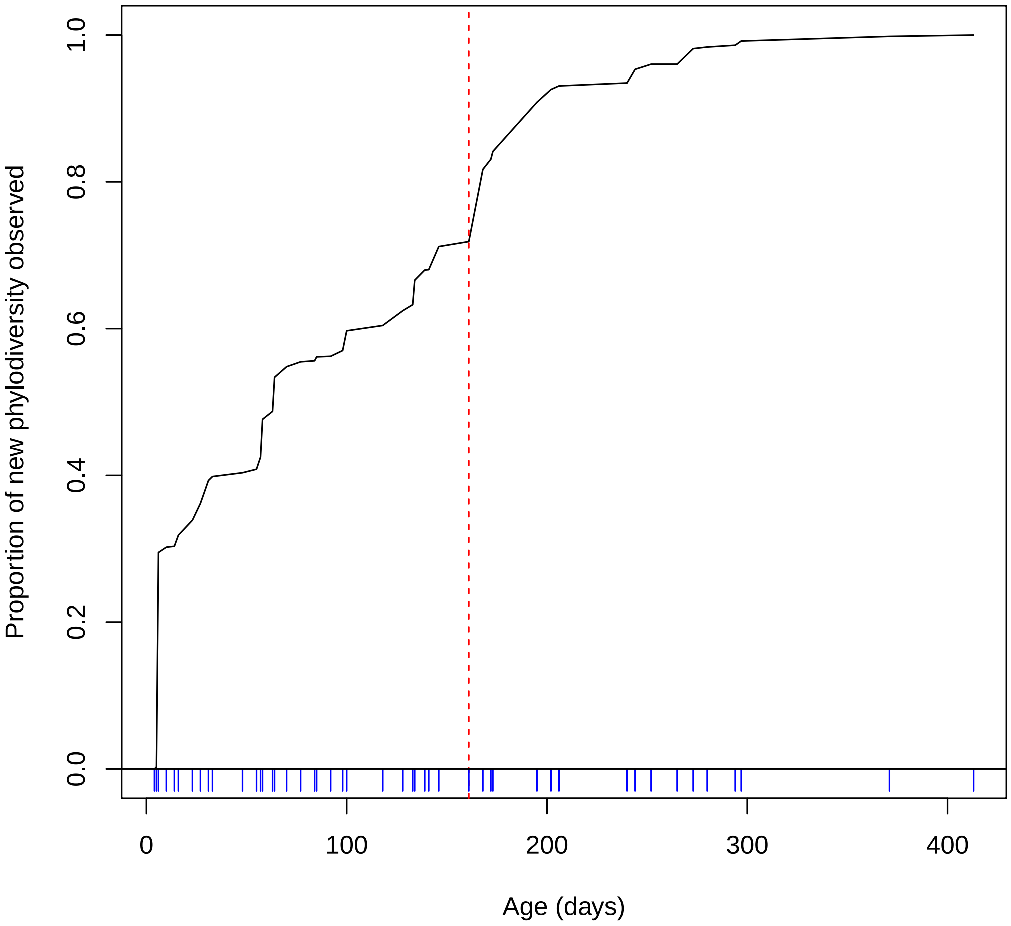
Phylodiversity accumulation and model fitting in the female feces dataset [13]. Plot A shows empirical (dashed) and surrogate phylodiversity accumulation curves. Each timepoint's phylodiversity value is the cumulative sum of all branch lengths observed up to that timepoint [36]. Curves are rescaled from 0 to 1 in this figure. The colored lines are 500 surrogate (*i.e.* resampled) phylodiversity curves with different $\hat{D}$ values (Equation 1), and are only calculated up to timepoint m, which is used to compare empirical and surrogate values. These lines are color-coded by their $\hat{D}$ value (see key at right). 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. 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.

Fig. 2



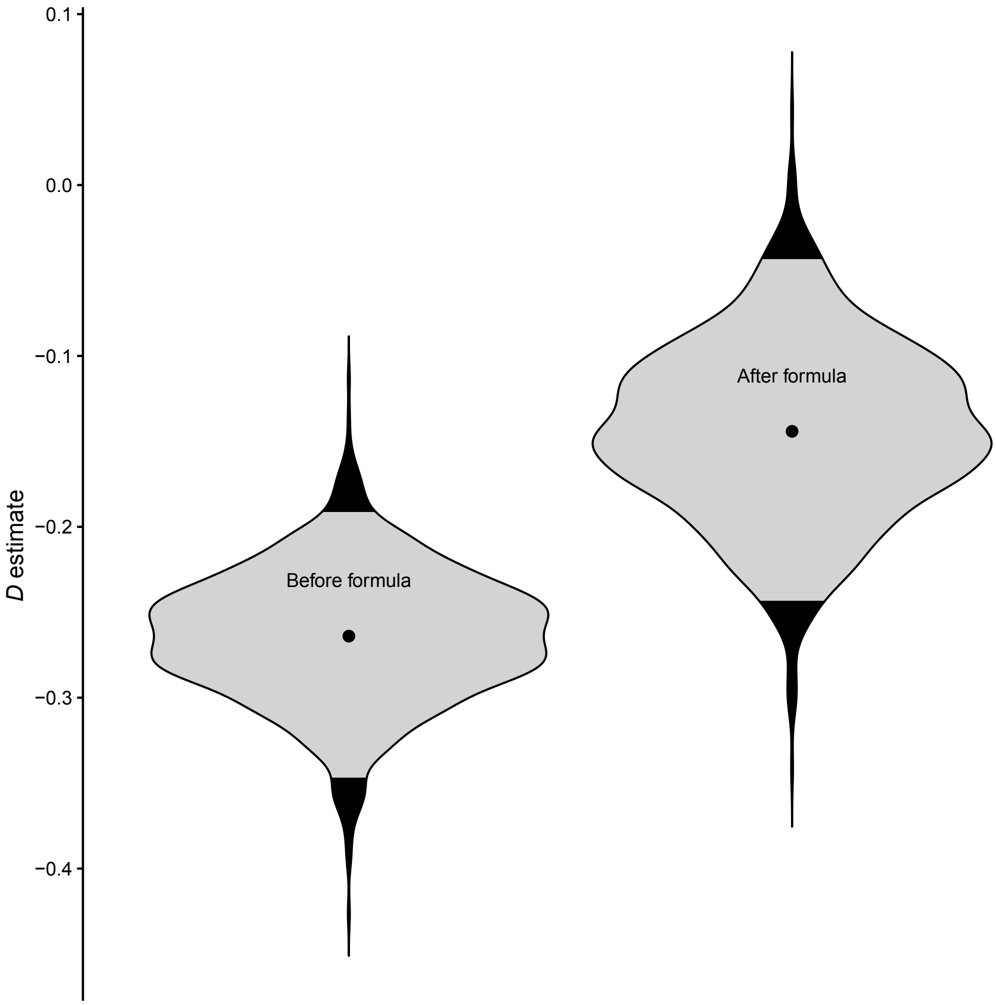
Dispersion parameter (D) estimates for "moving pictures" [13] datasets. The subject's sex is shown as the outline color of each violin, and the body site is shown as fill color. The four body sites for the female subject are shown at left, and the four body sites for the male subject are shown at right. Each viollin shows the distribution of D estimates given by logistic error model bootstraps, and the dots within violins are means. Colored portions of violins represent 95% of bootstraps. The two subjects analyzed show parallel D estimates, with feces being the lowest, followed by palms which are all similar, followed by tongue communities. For both subjects, tongue patterns were not significantly different than the neutral model.

Fig. 3



Empirical phylodiversity accumulation in the infant gut microbiome [7]. Phylodiversity increases sharply after day 161 of the infant's life, then plateaus. This timing coincides with the day the subject began consuming baby formula. The times of sampling points are shown as vertical blue lines below the X-axis.

Fig. 4



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

References

[1] Palmer MA, Ambrose RF, Poff NL. Ecological Theory and Community Restoration Ecology. Restoration Ecology. 1997 dec;5(4):291–300.

[2] Temperton VM. Assembly Rules and Restoration Ecology: Bridging the Gap Between Theory and Practice. Island Press; 2004.

[3] Richards SA, Possingham HP, Tizard J. Optimal fire management for maintaining community diversity. Ecological Applications. 1999 aug;9(3):880–892.

[4] Bengtsson J, Nilsson SG, Franc A, Menozzi P. Biodiversity, disturbances, ecosystem function and management of European forests. Forest Ecology and Management. 2000 jun;132(1):39–50.

[5] O'Dwyer JP, Kembel SW, Green JL. Phylogenetic diversity theory sheds light on the structure of microbial communities. PLoS computational biology. 2012 jan;8(12):e1002832.

[6] Goberna M, Navarro-Cano JA, Valiente-Banuet A, García C, Verdú M. Abiotic stress tolerance and competition-related traits underlie phylogenetic clustering in soil bacterial communities. Ecology letters. 2014 oct;17(10):1191–201.

[7] 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.

[8] Frank DN, Harpaz N, St Amand AL, Pace NR, Feldman RA, Boedeker EC. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences. 2007;.

[9] 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.

[10] 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.

[11] 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.

[12] Guittar J, Shade A, Litchman E. Trait-based community assembly and succession of the infant gut microbiome. Nature Communications. 2019;.

[13] 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.

[14] 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.

[15] Nemergut DR, Knelman JE, Ferrenberg S, Bilinski T, Melbourne B, Jiang L, et al. Decreases in average bacterial community rRNA operon copy number during succession. The ISME Journal. 2016 may;10(5):1147–1156.

[16] Sprockett D, Fukami T, Relman DA. Role of priority effects in the early-life assembly of the gut microbiota; 2018.

[17] Fukami T. Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and Priority Effects. Annual Review of Ecology, Evolution, and Systematics. 2015;.

[18] Verster AJ, Borenstein E. Competitive lottery-based assembly of selected clades in the human gut microbiome. Microbiome. 2018;.

[19] Litvak Y, Bäumler 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.

[20] 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.

[21] 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.

[22] Losos JB. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species; 2008.

[23] Darwin C. On the Origin of Species, 1859. London: Murray; 1859.

[24] 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.

[25] Peay KG, Belisle M, Fukami T. Phylogenetic relatedness predicts priority effects in nectar yeast communities. Proceedings of the Royal Society B: Biological Sciences. 2012;.

[26] 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.

[27] 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;.

[28] Stecher B, Chaffron S, Käppeli R, Hapfelmeier S, Freedrich 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.

[29] 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.

[30] 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).

[31] Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. Phylogenies and Community Ecology. Annual Review of Ecology and Systematics. 2002 nov;33(1):475–505.

[32] Lozupone C, Knight R. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. Applied and environmental microbiology. 2005;71(12):8228–8235.

[33] 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.

[34] 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.

[35] 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;.

[36] Faith DP. Conservation evaluation and phylogenetic diversity. Biological Conservation. 1992 jan;61(1):1–10.

[37] 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`.

[38] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010 may;7(5):335–6.

[39] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2409v1.

[40] Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28(14):1823–1829.

[41] Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution. 2015;.

[42] Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A phylogenetic perspective. Science. 2015 nov;350(6261):aac9323–aac9323.

[43] 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;.

[44] 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.

[45] Mayfield MM, Levine JM. Opposing effects of competitive exclusion on the phylogenetic structure of communities. Ecology Letters. 2010;.

[46] Webb CO. Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. The American naturalist. 2000 aug;156(2):145–155.

[47] Cavender-Bares J, Ackerly DD, Baum DA, Bazzaz FA. Phylogenetic Overdispersion in Floridian Oak Communities. The American Naturalist. 2004;.

[48] 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.

[49] Hubbell SP. The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32). Princeton University Press; 2001.

[50] 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.

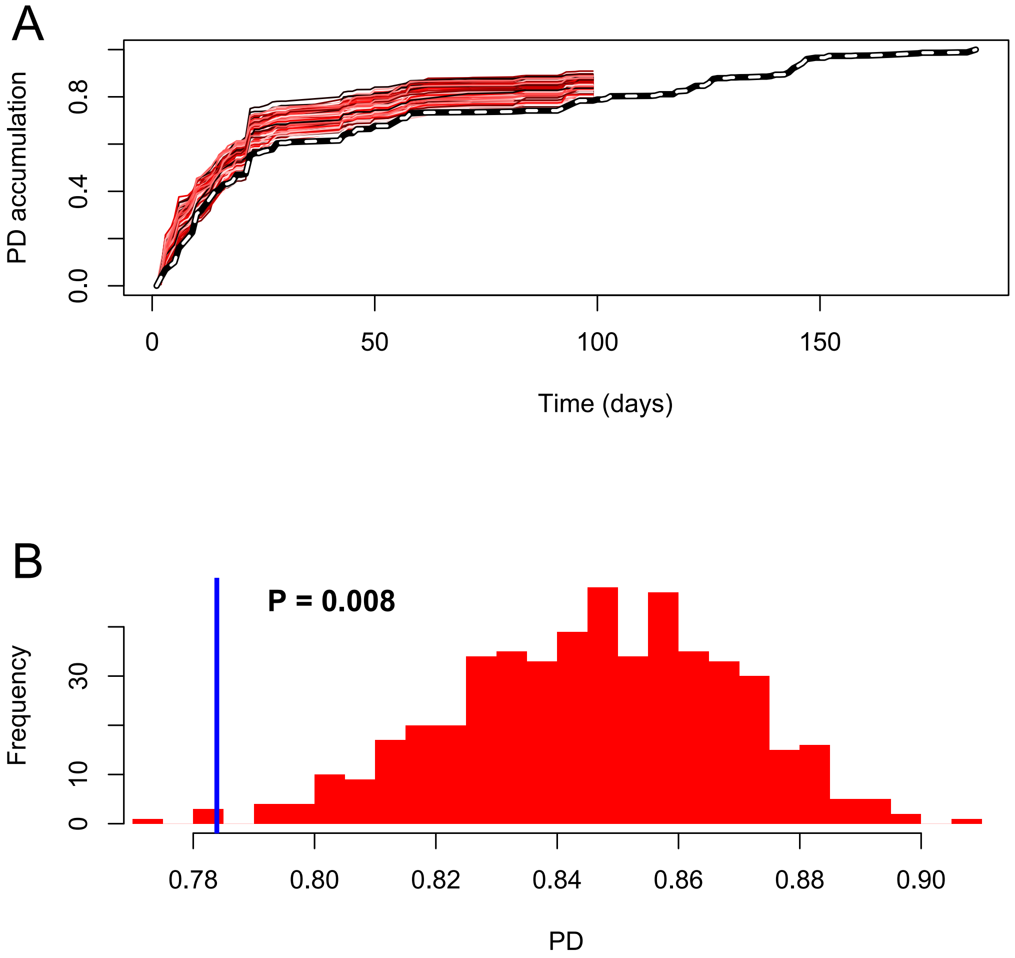
[51] Chatzidaki-Livanis M, Geva-Zatorsky N, Comstock LE. Bacteroides fragilis type VI secretion systems use novel effector and immunity proteins to antagonize human gut Bacteroidales species. Proceedings of the National Academy of Sciences. 2016;113(13):3627–3632.

[52] 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.

[53] 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.

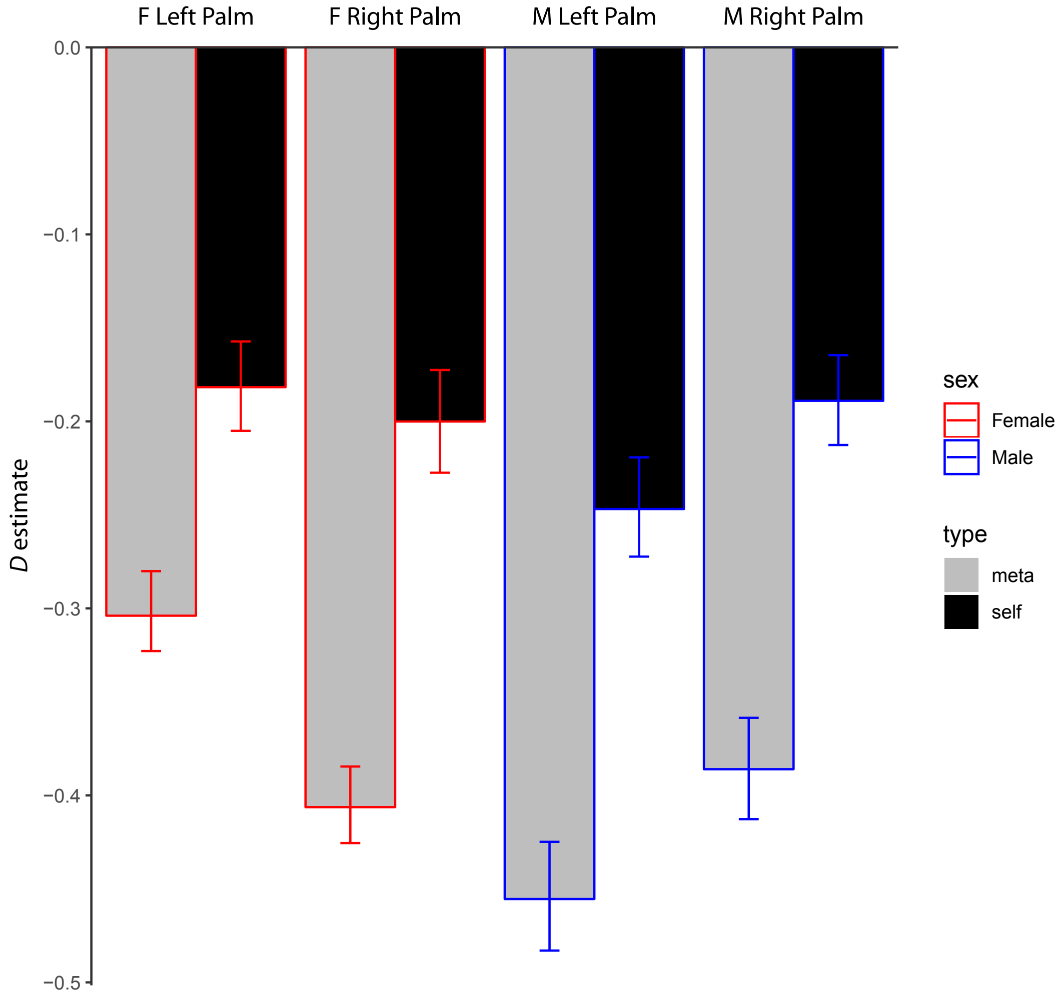
Supplementary Material

Fig. S1



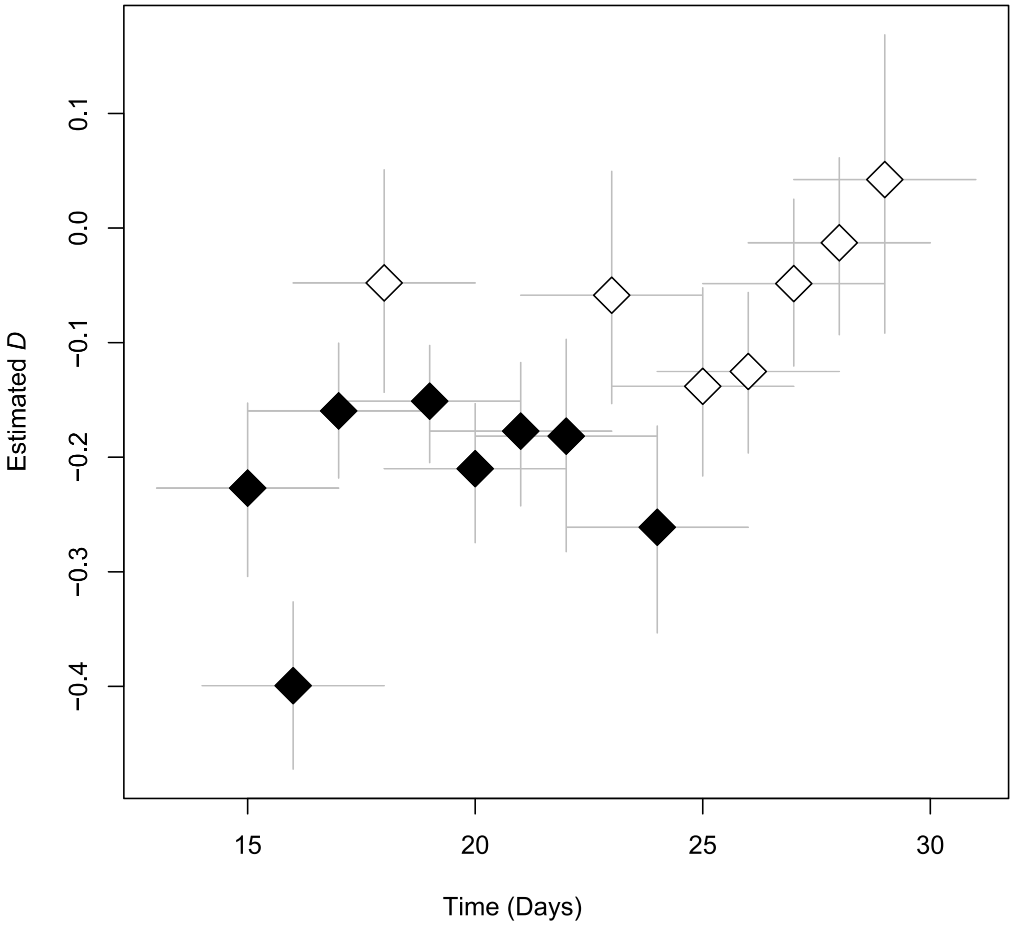
Significance testing for the female feces dataset. Plot A shows the empirical phylodiversity accumulation (dashed; same as Fig. 1A) but with neutral model surrogate datasets shown in different shades of red. These are produced by running the neutral model 500 times, to generate a distribution of phylodiversity 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 phylodiversity at time m (blue line) is compared to the distribution of neutral model phylodiversities at time m (red histogram), and a P-value is calculated as the proportion of neutral phylodiversities more extreme than the empirical value.

Fig. S2



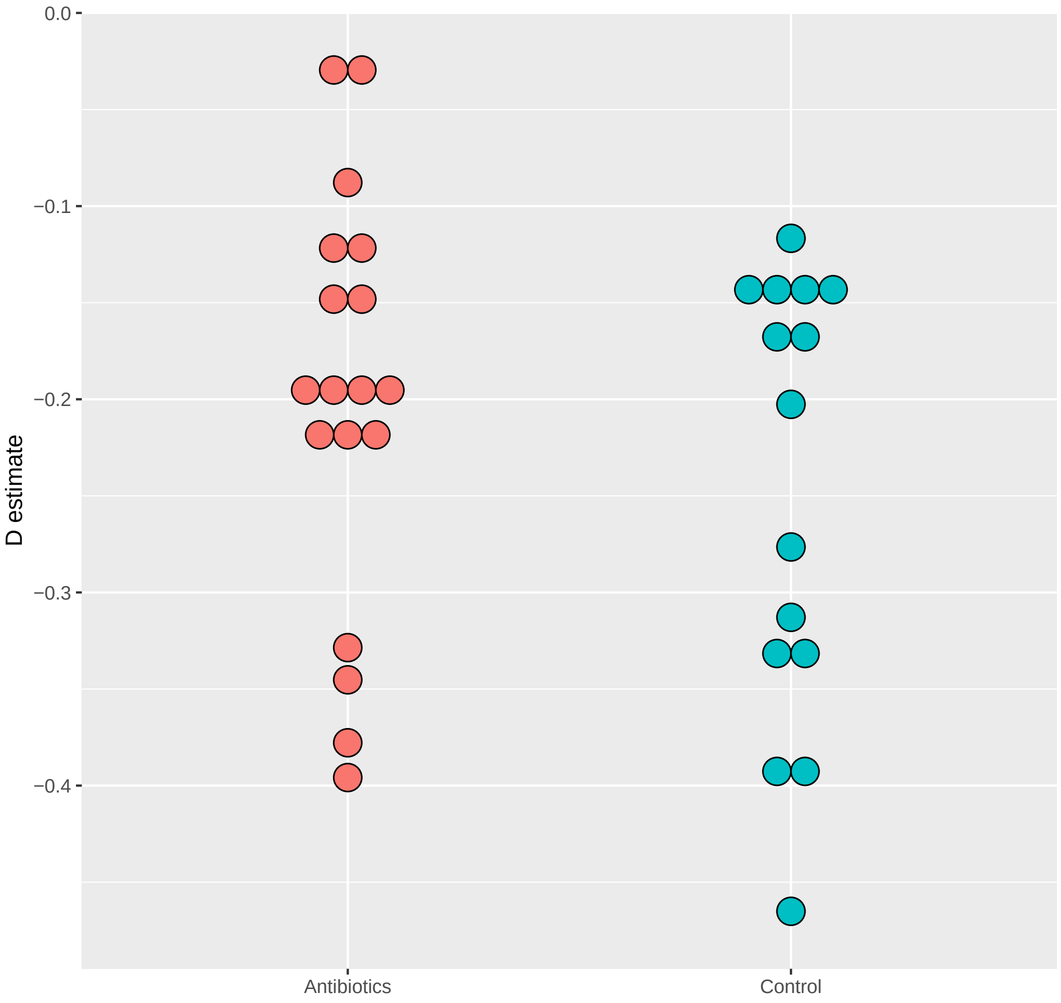
Comparison of “self” vs “meta” model results from palm communities. “Self” (black) models were run identically to Fig. 2), but “meta” (gray) models were run where the species pool for each palm community surrogate dataset was composed of all zOTUs observed across all four palm datasets. The difference between the “self” D estimate (generated above) and the “meta” D estimate (estimated with a metapopulation of zOTUs) is related to the exclusivity of recruitment into the community. In other words, 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, and we would learn that it would make little difference to community assembly patterns if the species pool really was composed of the “meta” set.

Fig. S3



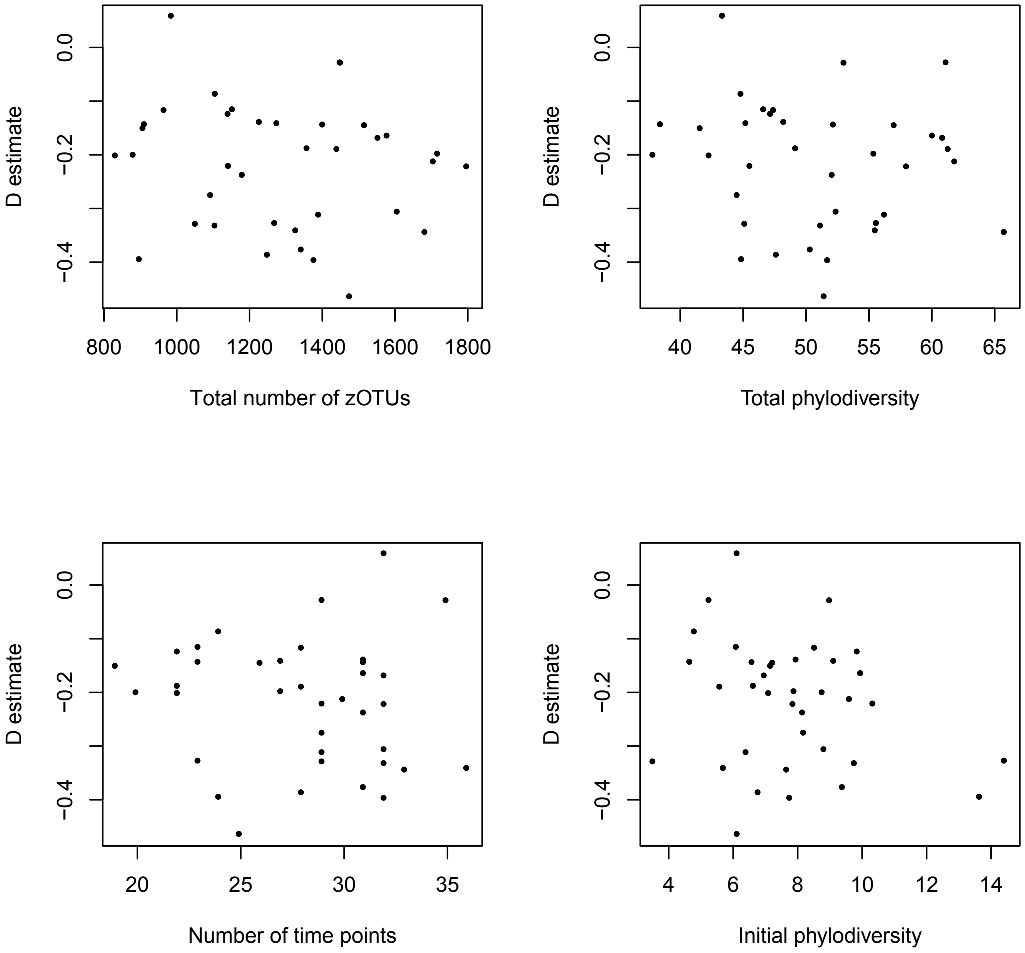
Sliding window analysis of male right palm data over 19 consecutive samples. We ran our model on each window of 5 continuous days (15 windows), in order to see how D varied over time. We only conducted this analysis for the section of samples that were sampled every day, so that comparisons between windows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

Fig. S4



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

Fig. S5



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