A phylogenetic model for the arrival of species into microbial communities and 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] Department 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 (numbers go here, ask Elizabeth or Larry?). The funding bodies had no role in study design, analysis, interpretation, or in the preparation of this manuscript.

Abstract

Understanding when and why species join 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 join communities; 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 often join microbial communities in short succession, suggesting factors such as shared adaptation to similar environments plays a stronger role than competition. To address this, we developed a mathematical model that describes the probabilities of different species joining a community over time, within a phylogenetic framework. We use our model to test three hypothetical assembly modes: underdispersion (species are more likely to join a community if a close relative has been previously observed), overdispersion (likelihood of joining is higher if a close relative has not been previously observed), and the neutral model (likelihood of joining 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 the human microbiome generally follows an assembly pattern characterized by phylogenetic underdispersion. 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 data sets 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 join communities, which can be studied in part by evaluating the order in which particular microbes join communities given data such as which microbes are already present or what changes occur in an environment over time [14; 16]. Although the local environment clearly selects for microbial 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 microbe that arrives first 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 species are less likely to join a community 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 join 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], and may even be preferentially recruited [27]. 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 close relatives are more likely to join a community than 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 species arrival into communities 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**.

Here, we use the phylogenetic relationships among species within a time-series to understand the community's change over time in a new way. Instead of analyzing broad patterns of community change via beta-diversity statistics (*e.g.* UniFrac [28]) or analyzing patterns of select clades within the community (*e.g.* PhyloFactor [29], Edge PCA [30]), we model the probability of each new species' first arrival into the community as a monotonic function of its phylogenetic distance to members of the community that have already arrived.

The model we present here can be used to estimate the degree to which newly arriving species are phylogenetically over- or underdispersed during a time-series microbiome dataset. Said another way, we estimate the extent to which these arrival events are "nepotistic", meaning species are more likely to arrive when a close relative is already present (phylogenetic underdispersion). We fit our model (described in our methods section, below) to several time-series human microbiome data sets [13; 7; 31], to compare patterns of microbial community assembly between subjects, sample sites, or time periods. We found that in general, the human microbiome follows the underdispersion hypothesis – microbes with low phylogenetic distance to the existing community had a higher probability of recruitment than microbes with a high phylogenetic distance to the existing community. There were exceptions where this pattern was not significantly different than the neutral model, but none of the longitudinal data sets we analyzed showed statistically significant overdispersion. Furthermore, the predominance of non-neutral assembly in the microbial communities analyzed here suggests that phylogeny carries relevant and potentially predictive information for microbial community assembly, with implications for microbiome perturbation and rehabilitation.

Materials and Methods

Overview

Here, we describe a statistical model of phylogenetic microbial community assembly, and how we apply that model to time-series microbiome data. Our goal is to estimate the degree to which species arriving into a community for the first time are phylogenetically over- or underdispersed, competitively testing the overdispersion or underdispersion hypotheses against the neutral hypothesis. Said another way, we want to estimate the extent to which recruitment of species into the community is related to arriving species' phylogenetic similarity to (or distance from) species that already arrived. Our **Statistical Model** describes the probabilities of those species arriving into a community over time. We use our model with empirical data via **Simulations**, where we re-sample the empirically observed species using our model with known parameter values, to produce surrogate data sets. 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 arrival to that of the surrogate data sets (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 not yet added ("species pool") represent organisms present within the metacommunity but not the local community. Our model parameterizes the probability of species arriving in a local community for the first time, based on their phylogenetic distances from species that have already arrived. In a species-neutral model of community assembly, each species i in the species pool has the same probability of arrival at time t, irrespective of how different it is from species that are already present at time t. Thus, the neutral model for first-time arrivals is a random draw without replacement of species from the species pool. We extend the species-neutral model by modeling the probability $p_{i,t}$ of species i being observed for the first time at time t as,

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

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

When $D = 0$, our model functions as a neutral model; all species have the same probability of arriving in the community for the first time, since pi,t is the same for every species. When $D < 0$, $p_{i,t}$ decreases with $d_{i,t}$ meaning that species from the species pool have higher probabilities of arriving when they are more closely related to species that have already been observed 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 data set relies on re-constructing that data set 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 data sets with $\hat{D}$ values ranging from $\hat{D} < 0$ (underdispersed) to $\hat{D} = 0$ (neutral) to $D > 0$ (overdispersed). This is done so that the empirical data can later be compared to the surrogate data sets, to estimate the empirical value of D.

We start each surrogate data set with the same species present in the first sample of its corresponding empirical data set. Then, surrogate data sets are constructed forward in time by randomly drawing r_t new arrivals from the species pool (all species observed in the empirical data set that have not yet been sampled by this process), where the probabilities of species arriving at any given time are given by Equation 1, and r_t is the number of new arrivals in the empirical data set from times $t-1$ to t. The number of arrivals from the empirical data set is used so that species richness is kept constant between the empirical data set and all surrogate data sets. The species pool is updated to exclude those species drawn at previous time points, and the newly sampled species are recorded. Surrogate data sets are produced for many different $\hat{D}$ values, ranging from underdispersed to overdispersed models. In the analyses we present here, we performed 500 simulations (as described above) for each data set analyzed.

Parameter Estimation

Our main goal is to estimate the empirical dispersion parameter D (Equation 1), which quantifies the degree to which first-time arrivals are phylogenetically underdispersed ($D < 0$), neutral ($D = 0$), or overdispersed ($D > 0$). As previously stated, these assembly modes correspond to our hypotheses (see Introduction section). To this end, we use Faith's phylodiversity [32] to compare each of the 500 surrogate data sets (described above) to the empirical data set. 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 [32]. 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 ($d_{i,t}$, Equation 1), yielding accumulations of total phylodiversity that are relatively slow ($D < 0$) or relatively fast ($D > 0$) compared to the neutral model (Figure 1A). In other words, at any time point 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 data sets, 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 data sets. Multiple time indices are not used to compare surrogate and empirical data sets because each value $PD_{\hat{t}}$ is a function of all values $PD_{t<\hat{t}}$. PD_m values are calculated for all surrogate data sets, and a PD_m value is calculated for the empirical data set. The differences between the empirical PD_m and each PD_m simulated with $D = \hat{D}$ are given by the vector $\Delta PD_{\hat{D}}$, which is the error between surrogate and empirical data. We then estimate the empirical value of D by minimizing $\Delta PD_{\hat{D}}$ (Figure 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 (Figure 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 (over- or underdispersion, a 2-tailed test). We test this null hypothesis competitively by simulating 1000 surrogate data sets at $D = 0$ (Figure 2A) to generate a null PD_m distribution. The empirical PD_m is compared to this distribution (Figure 2B), 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

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 [33] was used to trim primer regions from these data. Clustering was performed using the unoise3 pipeline [34]; sequences were de-replicated at 100% identity using vsearch [35], zOTU centroid sequences were picked and chimeric sequences were removed using unoise3 [34], then all sequences were mapped onto zOTU seeds to create a zOTU table using vsearch. zOTU stands for "zero-radius operational taxonomic unit" [34]. Unlike traditional de novo clustered OTUs, zOTUs are exact sequence variants (ESVs) which are consistent and easily comparable across data. The SINA aligner [36] 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 [37] 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 time points were excluded as well because they were sampled at a much lower temporal resolution. This left 52 time points spread over the first 469 days of the infant subject's life. The zOTU table was then split into two zOTU tables, one for time points 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 data sets were compared to the empirical data set as described above, using difference between phylodiversity values at the middle time point. 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). Time points were excluded which did not have sequence data for each of the 8 environments (L hand, R hand, mouth, and feces of the male and female subjects), and rarefied to 5000 sequences per sample. This left 107 time-points, ranging from day 1 to day 185. Analysis for each data set (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 were intended to showcase potential functionality of our model, albeit without directly testing any hypotheses. We analyzed palm communities in a "meta" context, where surrogate data sets were generated assuming the species pool for a given palm was composed of all four palms in the data set. 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 arrivals 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. 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.* [31] 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 [31]. Roughly half of these infants were repeatedly treated with oral antibiotics, almost universally for ear infections. Metadata for this data set 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 data sets 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 data set had so many subjects, we used this analysis as an opportunity to analyze whether the number of zOTUs, total phylodiversity, or number of time-points 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 at `https://figshare.com/s/922b268891f1945c1944` (temporary private link, please do not share until publication). R functions to use our model are also available on GitHub, along with a tutorial vignette: (link will be added upon acceptance).

Results

By varying $\hat{D}$, we successfully changed the rate at which phylodiversity is added to surrogate (*i.e.* resampled) microbial communities over time (Figure 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 nicely to the differences between empirical and surrogate data sets ($\Delta PD_{\hat{D}}$, Figure 1B). Each error model fit was manually inspected to be sure that D estimates were not spurious, and all data analyzed produced nice looking fits, including data sets where D was not significantly different from 0.

Results from "moving pictures" data

All time-series from adult feces and palm microbiomes [13] showed significant phylogenetic underdispersion of first-time arrivals (Figure 2). This means that when a zOTU was observed for the first time in one of these communities, it was more likely to be phylogenetically similar to a zOTU that had previously arrived in that 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 (Supplemental Figure 2). We also observed a general upward trend in D in our sliding window analysis of the male right palm data set (Supplemental Figure 3), 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 (Figure 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 data set 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 (Figure 4). While the pre-formula data set was significantly underdispersed ($P = 0.007$), the formula data set was not significantly different from the neutral model, although this result is marginal ($P = 0.107$). Infant gut data from Finnish infants [31] 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 (Supplemental Figure 4). Estimates of D did not significantly correlate with the number of zOTUs in a data set, the total phylodiversity of the data set, or the number of samples in a data set (Supplemental Figure 5).

Discussion

Any organism of interest in a human microbiome data set, from the pathogenic to the probiotic, will at some point arrive for the first time, and the order in which these organisms arrive in the community is determined by community assembly processed [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 [38; 39]. Identifying conditions under which assembly mechanisms change, or under which non-neutral assembly is particular 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 arriving, 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 (Figure 4, Supplemental Figure 4) and during turnover of the microbial communities on the adult

palms and gut (Figure 2) follows a predictable pattern: species are more likely to arrive if a close relative has already arrived.

This generally "nepotistic" pattern in arrivals strongly supports our underdispersion hypothesis, which is similar to Darwin's pre-adaptation hypothesis [23]. Under that hypothesis, species are better able to join a community where a close relative is present because they likely already have the "right stuff" to live there. This is because closely related species are likely to be ecologically similar due to phylogenetic niche conservatism [40]. Although the phylogenetic trees we use here are constructed with 16S rDNA sequences, such phylogenies have been shown to track genomic differences in bacteria [20; 21]. Indeed, non-neutral patterns of phylogenetic community structure have been interpreted to mean that traits are under ecological selection [41; 42; 43; 44]. If traits are not driving community assembly [45] 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 very a strong and significant phylogenetic signal in arrival order for almost all data sets we analyzed.

However, even if microbial community assembly follows the pre-adaptation hypothesis, 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 [46], so a pre-underdispersed community entering the host is a plausible mechanism for phylogenetic underdispersion of arrivals. 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 data sets 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 our own intuition, and also because of work in experimental microcosms supporting Darwin's naturalization hypothesis [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. Analysis using our model captures pattern, but does not interrogate the process by which underdispersion occurs, and future experiments will need to be designed with such considerations in mind.

Nonetheless, a strength of our model is that it estimates values of D that can be compared among data sets (Figure 2) or potentially across time (Figure 4, Supplemental Figure 3) in order to learn how differences between data sets impact community assembly. We found that gut and palm communities were almost universally underdispersed (Figure 2, Figure 4, Supplemental Figure 4), and that the D value for a community appears to be a function of body site (Figure 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 (Figure 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 general trend for the human gut microbiome, and potentially for skin microbiomes as well. If true, our findings have implications for the management and restoration of human-associated microbial systems [47], in particular for probiotic development and remediation of pathological microbial communities. Without any other information, our finding of consistent phylogenetic underdisper-

sion in arrivals suggests that probiotics for sustained colonization of the human gut should be close relatives to the microbes already present. 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]. 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.

Microbial communities provide a unique opportunity to study community assembly in primary and secondary succession. In addition to standard cross-sectional studies of communities of different ages or successional stages, the short timescales of microbial community dynamics allows longitudinal studies of community assembly over manageable time frames [9; 11]. Microbial communities allow large sample sizes, longitudinal studies, and experimental manipulations that enable us to identify general rules and statistical patterns of community assembly. The model we present here makes use of such data, and to facilitate further discovery both in the human microbiome and in other environments, we have made our R code available at FigShare: `https://figshare.com/s/922b268891f1945c1944` (temporary private link, please do not share until publication) and GitHub: (link will be added upon acceptance).

Acknowledgements

The authors thank D.R. Nemergut for her help and support, and also thank J.P. O'Dwyer, P. Sommers, E.M. Gendron, A. Solon, E. Preusse, K. Hazleton, and S. Sauce for many helpful discussions.

Figures

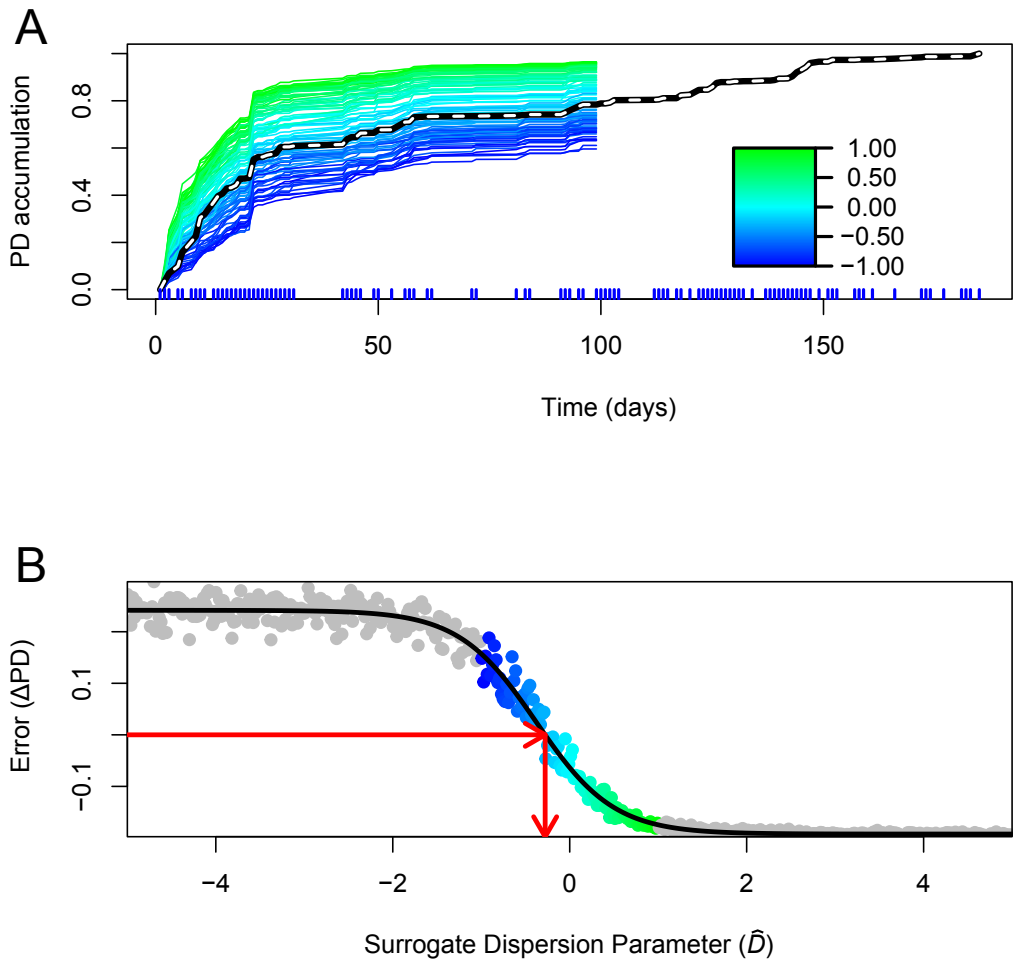


Figure 1: Phylodiversity accumulation and model fitting in the female feces data set [13]. Plot A shows empirical (dashed) and surrogate phylodiversity accumulation curves. Each time point's phylodiversity value is the cumulative sum of all branch lengths observed up to that time point [32]. Curves are 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 time point 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 arrivals. 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 data sets are shown on the X-axis. Color-coded points correspond to surrogate data sets 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 solving the model for 0 error, yielding a D estimate. A figure showing significance testing for these data is available as Supplemental Figure 1.

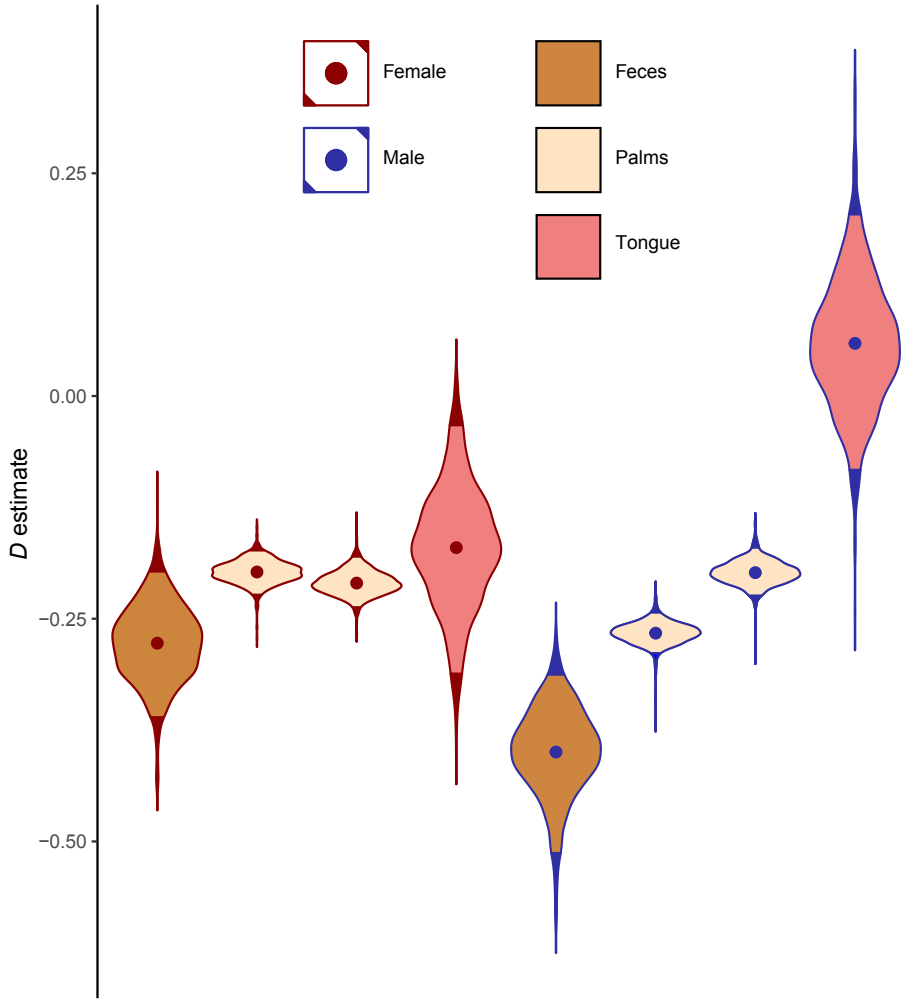


Figure 2: Dispersion parameter (D) estimates for "moving pictures" [13] data sets. The subject's sex is shown as the outline color of each violin, and the body site is shown as fill color. The four body sites for the female subject are shown at left, and the four body sites for the male subject are shown 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.

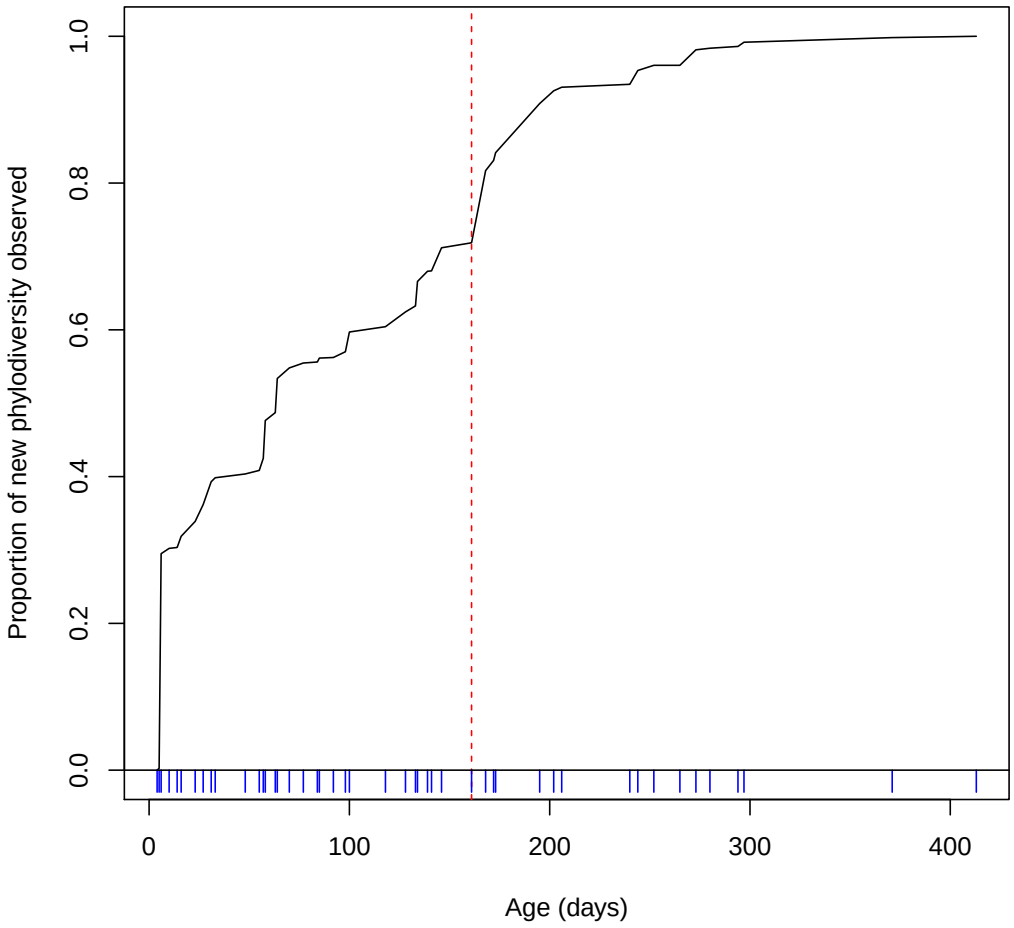


Figure 3: Empirical phylodiversity accumulation in the infant gut microbiome [7]. Phylodiversity increases sharply after day 161 of the infant's life, then plateaus. This timing coincides with the day the subject began consuming baby formula. The times of sampling points are shown as vertical blue lines below the X-axis.

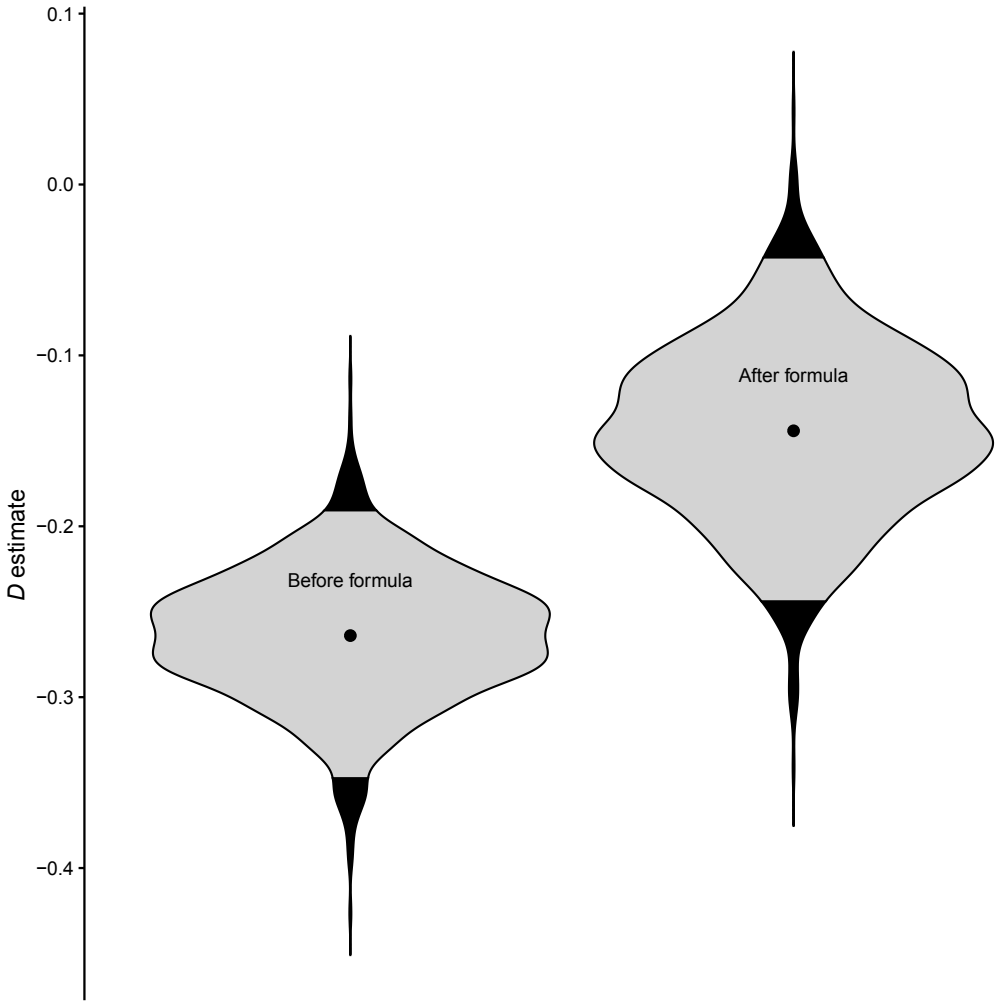


Figure 4: Dispersion parameter (D) estimates in the infant gut, pre-formula and during formula use. Formula use began on day 161, thus the first 160 days of the subject's life were analyzed separately. Community assembly was significantly underdispersed in the pre-formula data set, but was not significantly 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] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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;.

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

[33] 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.

[34] 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`.

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

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

[37] 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;.

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

[39] 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;.

[40] 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.

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

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

[43] Cavender-Bares J, Ackerly D Baum D Bazzaz F Phylogenetic Overdispersion in Floridian Oak Communities. The American Naturalist. 2004;.

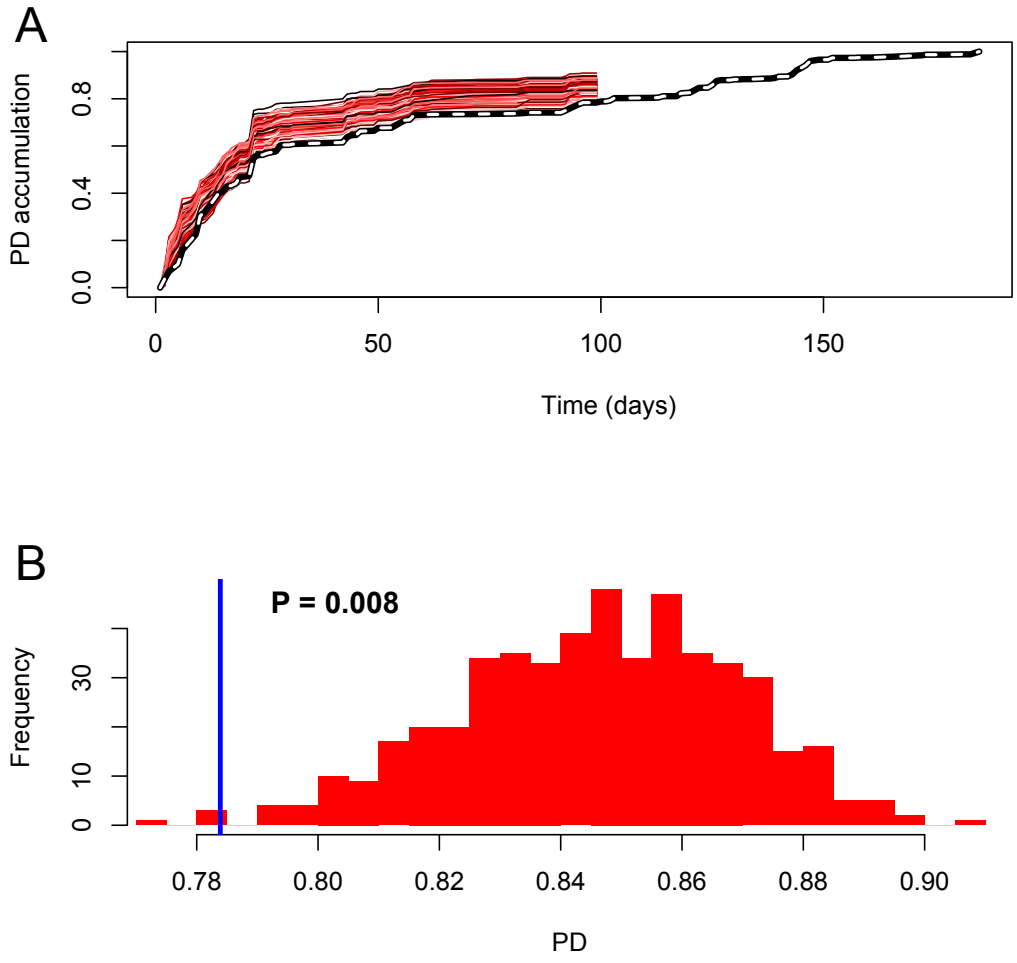
[44] Gerhold P, Cahill JF, Winter M, Bartish IV, Prinzing A. Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). Functional Ecology. 2015 may;29(5):600–614.

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

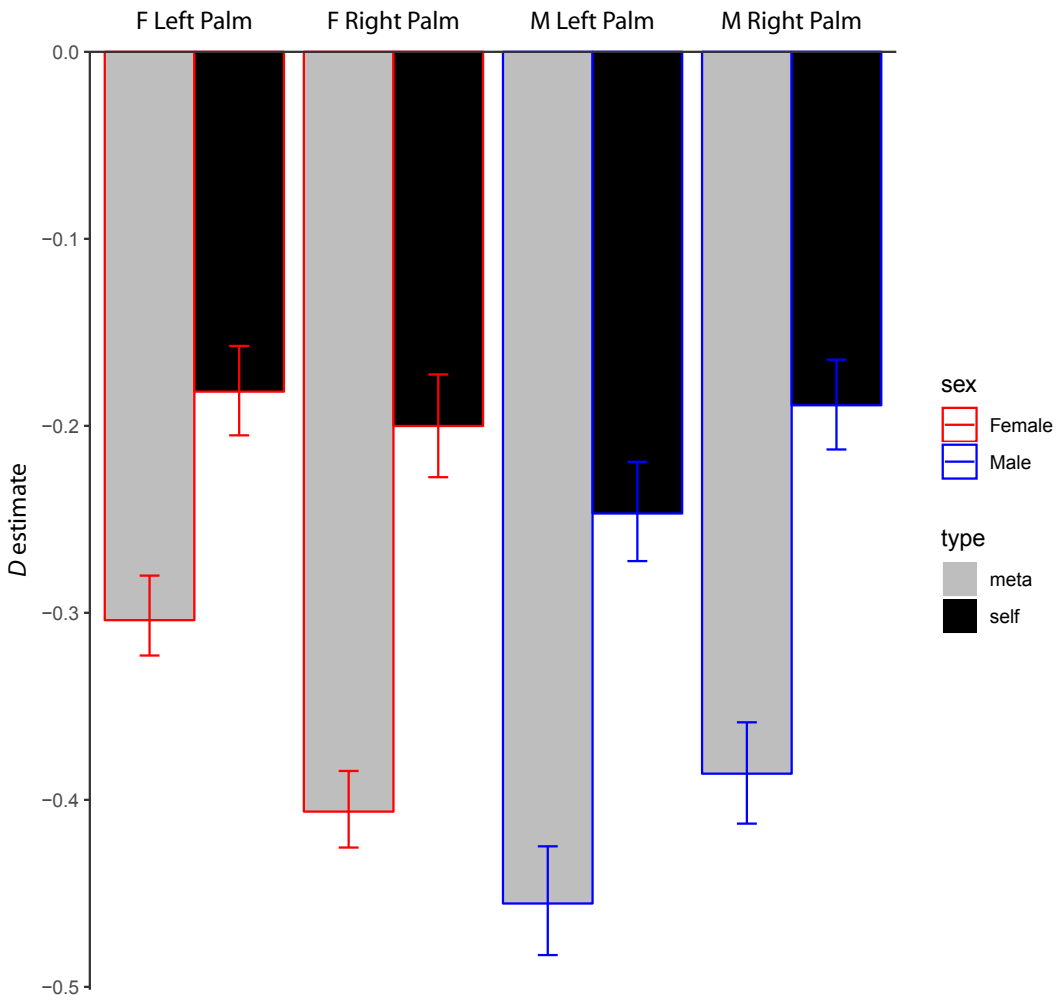
[46] Emerson BC, Gillespie RG. Phylogenetic analysis of community assembly and structure over space and time. Trends in ecology & evolution. 2008 nov;23(11):619–30.

[47] Shooner S, Chisholm C, Davies TJ. The phylogenetics of succession can guide restoration: an example from abandoned mine sites in the subarctic. Journal of Applied Ecology. 2015 dec;52(6):1509–1517.

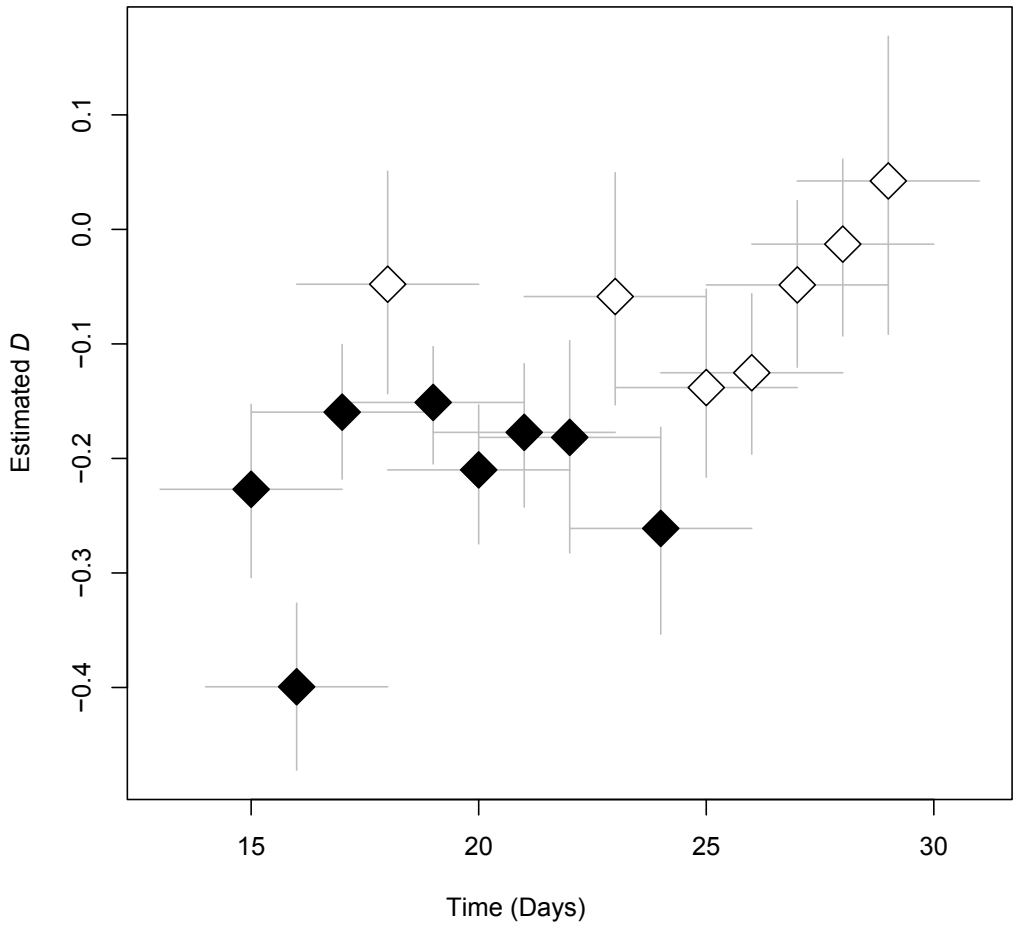
Supplemental Material



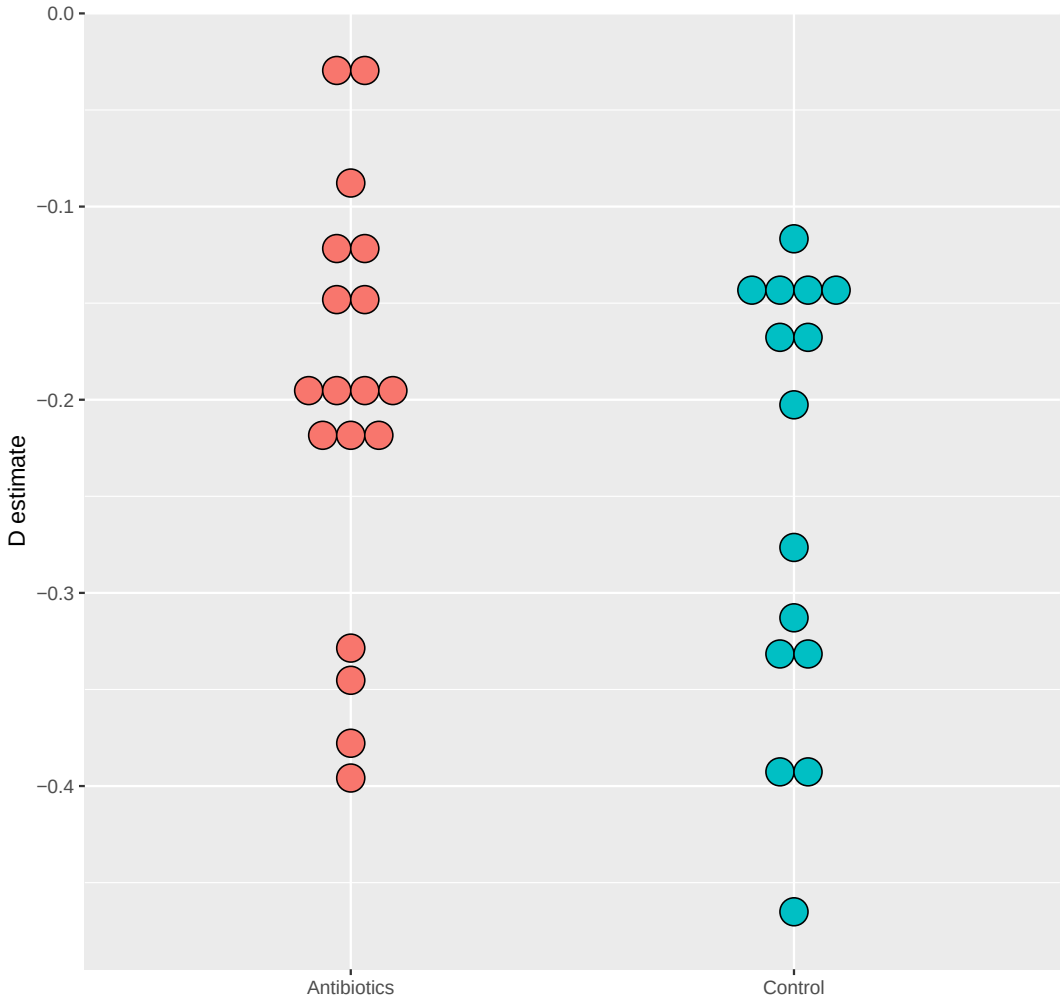
Supplemental Figure 1: Significance testing for the female feces data set. Plot A shows the empirical phylodiversity accumulation (dashed; same as Figure 1A) but with neutral model surrogate data sets shown in different shades of red. These are produced by running the neutral model 500 times, to generate a distribution of phylodiversity values under $D = 0$ (Plot B). As with all surrogate data sets, 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.



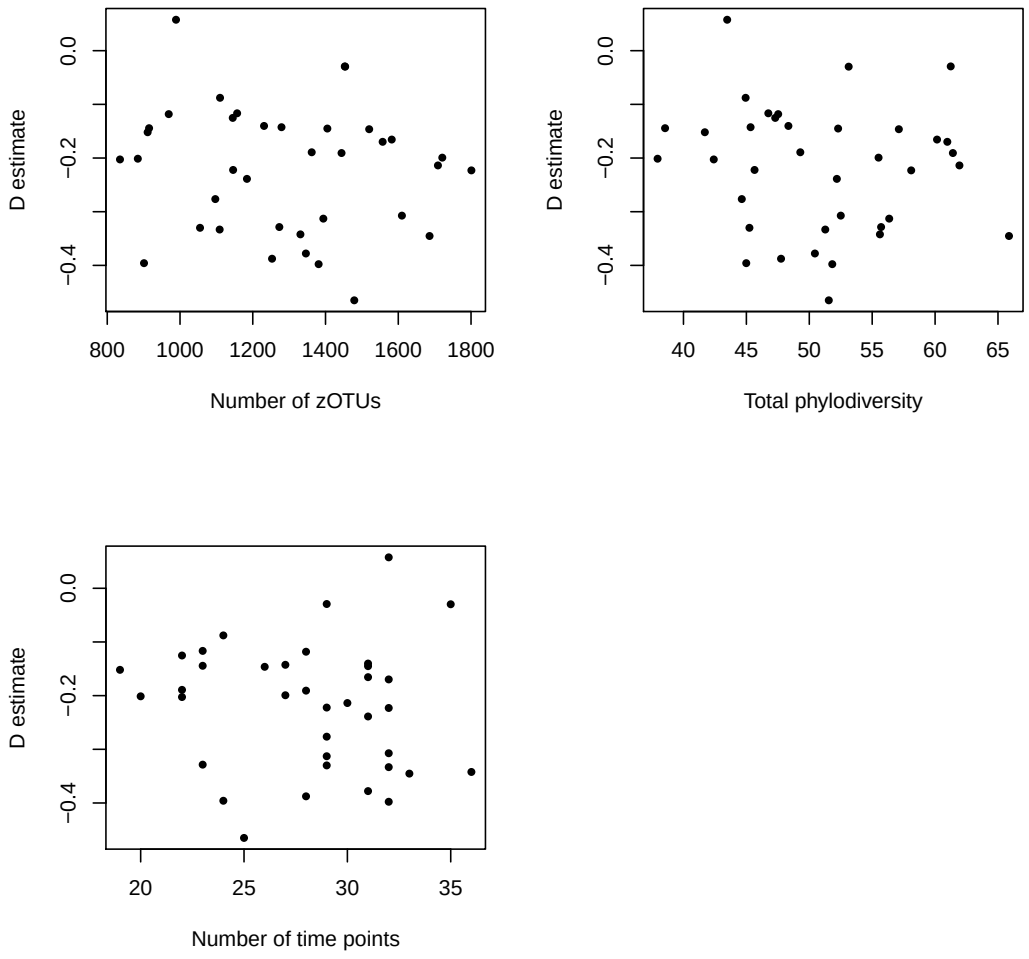
Supplemental Figure 2: Comparison of "self" vs "meta" model results from palm communities. "Self" (black) models were run identically to Figure 2), but "meta" (gray) models were run where the species pool for each palm community surrogate data set was composed of all zOTUs observed across all four palm data sets. 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 arrivals 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.



Supplemental Figure 3: 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.



Supplemental Figure 4: D estimates of Finnish infant data sets. 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.



Supplemental Figure 5: Relationship of D estimate to total phylodiversity, zOTU richness, and number of time-points sampled for Finnish infant data. No statistically significant correlation was detected in any of these three analyses.