University of Colorado Anschutz Medical Campus

Division of Biomedical Informatics and Personalized Medicine

Mail Stop 8617
12700 E. 19th Ave.
Aurora CO 80045

ATTN: The Editors of *The ISME Journal*

Dear Editors,

Please consider our manuscript "A phylogenetic model for the arrival of species into microbial communities and application to studies of the human microbiome" for publication in *The ISME Journal*. In our manuscript we present a novel statistical model we developed in order to investigate the extent to which arrival of bacteria into the human microbiome is contingent on closely-related bacteria already living there. This question is borne out of previous findings that bacterial recruitment can be "nepotistic", but also from recent and historical ecological research that suggests the opposite pattern may be true due to factors such as competition for resources. We apply our model to three longitudinal human microbiome high-throughput sequencing datasets. Our results strongly support the "nepotism" hypothesis, but also show that the strength of this pattern (effect size) varies by individual and by body site. Our model and results have broad implications for both human and natural systems, *e.g.* where it is valuable to predict whether a pathogenic bacterium may successfully colonize, or implications for the development of pre- and probiotic therapeutics.

This manuscript is completely re-written from a manuscript previously submitted to *The ISME Journal*. That manuscript received generally minor comments from reviewers, but was nonetheless rejected. The current version is extensively revised, although the core mathematical model (Equation 1) remains the same. Our rationale (introduction), results, methods, and discussion have been re-written with the reviewers' comments in mind; and in our opinion this manuscript is sufficiently distinct to warrant your further consideration. We have appended responses to reviewer comments to the end of this cover letter (next page). A pre-print of this manuscript is also available on BioRxiv (doi: 10.1101/685644), and all the code and data necessary to repeat our analysis or to conduct similar analyses are available on GitHub. No part of our manuscript is currently under consideration for publication elsewhere, and we have no competing interests to declare.

Thank you for your consideration of our manuscript.

Sincerely,

John L. Darcy, PhD

Comments from reviewer(s): Referee #1 (Comments to the Author):

The manuscript seeks to describe the assembly of microbial ecosystems, and particularly to characterise them as appearing to be shaped by overdispersal or underdispersal of individuals.

The topic of study of certainly of reasonably broad interest, but the model and the implied results need more explanation and justification. Additionally, much of the writing is really quite sloppy and below publishable standard. This, though, can easily be fixed up. The discussion is a little repetitive in places and, in general, the paper could be written a little more concisely.

*****We have totally re-written our manuscript with a focus on better explaining the theoretical and mechanistic rationale for our model and subsequent analyses. We took extra care to make sure our writing is as crisp and succinct as possible.**

Major points:

The first lines of the abstract read very strangely. Without getting too philosophical, is *why* really a question which is being addressed here? I would also argue that the wording blurs the line between model mechanisms and actual community mechanisms, which may not be the same. What has been done is that mechanisms have been proposed within the mathematical model and the consequences of these mechanisms compared to actual community observations. This, of course, does not mean that the mechanisms have been described correctly. That is true of almost all modelling, but the wording here is in danger of overselling mechanistic inference.

*****We agree with the reviewer's criticism of our previous manuscript's epistemology. Indeed, "why" a microbe joins a community is of great import to our model's rationale, but this was not adequately articulated in that manuscript. We now discuss historical contingency in microbial community assembly in much more depth (2nd and 3rd paragraphs of our new introduction) and how such patterns are explicitly under question in our manuscript. We also now introduce our hypotheses in a broader ecological context (instead of simply a mathematical/theoretical context) in order to make the motivation for testing such hypotheses more clear and more easily generalized.**

There's no discussion of why the assumption that D should vary monotonically with phylogenetic distance. Has this been shown to be the case in other studies? Has it been previously hypothesised?

*****D is just a parameter in our model, and it varies monotonically with phylogenetic distance as a consequence of our model itself (Equation 1). We now make it clear that this is a feature of our model itself.**

(Page 8) Is the assumption of a constant population size realistic for all these communities? Again, a justification or similar study would help here. How sensitive are the results to relaxing this assumption?

*****We have removed the discussion of island-biogeographic equilibrium because it was only tangentially relevant to our model, and we wanted to make our results and discussion more concise per the reviewer's recommendation. There is no assumption of population sizes in our model, since (as we now clarify in our introduction) we model arrival of species, not communities themselves.**

The discussion is weak. “If there are patterns or general rules”. There clearly are some rules, albeit perhaps ones we don’t fully understand. I don’t see why “The burst of newly observed phylogenetic diversity indicates that this particular disturbance[...]resulted in a re-assembly of a community that was approaching island-biogeographic equilibrium” How does this follow?

*****We agree with the reviewer that our previous statement did not 100% follow, since the influx of diversity would indicate a community shift, but not necessarily that the community had reached island biogeographic equilibrium beforehand. We have removed discussion of equilibria for the sake of brevity. We now discuss more of our findings in context of previous ecological theory, and have removed more speculative commentary from our discussion.**

Minor points:

It is more normal to use the word “dataset” as opposed to “data set”. Similarly “timepoint” instead of “time point” or “time-point”, both of which are used in the manuscript.

*****We agree and have made this change.**

Some variables not in italics, whereas others are. Should be more consistent.

*****We agree and have made this change.**

(Page 6) The idea of metacommunities is introduced without reference or proper definition. This could easily be referenced out to some of the early studies which rely on neutral (and near-neutral) models with such a framework.

*****We agree and have made this change.**

(Page 9) “KS Distance” introduced without full name.

*****We no longer use KS distance to fit our model to empirical data. Our new approach is more intuitive, more robust, and makes prettier figures.**

(Page 11) Probably best to mention PCR fully by name before using abbreviation.

*****We disagree, and think PCR is so commonly used that it does not need to be spelled out.**

(Page 12) “L hand” and “R hand”. Full words please.

*****We agree and have made this change.**

(Page 13) The first clause in the first sentence is missing a verb.

*****The entire paragraph has been re-written.**

Are emojis the best labels for bars on a figure? In particular, you cannot be sure which hand is left and which hand is right – is the palm meant to be facing up or down? A simple word would have helped.

*****We maintain that the pictograms we used made our figure easier to understand at-a-glance. When we completely re-made this figure, we did not use pictograms per the reviewer's suggestion.**

Referee #3 (Comments to the Author):

This manuscript tries to answer to which extent the recruitment of new species into ecological communities is determined by their phylogenetic distance to community members. To do so, the authors develop a mathematical model that describes the probability of a species to enter into the community based on its phylogenetic distance to its closest relative in the community. They use their model with two published empirical datasets, ie. time-series of the human microbiome described through 16S rRNA gene amplicons, which they take as the 'observed communities'. Then, they create 'surrogate communities', taking the first sample of the time series of the observed communities as a starting point and reconstructing the subsequent samples in the time series. To do so, they randomly sample new species from the global pool at different levels of parameter D, which determines the extent to which new species with a close relative in the community are added to the community. The final aim is to estimate parameter D in observed communities by comparing the observed and simulated values at which phylogenetic diversity accumulates through the time series, and finally compare estimated D values to those expected under a null model of immigration in which the likelihood of joining the community is not related to phylogenetic relationships among species. The manuscript tackles a fundamental question in community assembly, describes an original research, is well written and relatively easy to follow, and beyond providing the mathematical model reaches a clear conclusion: the likelihood that a bacterial species has of joining a community in the human microbiome increases if a close relative is present. The conclusion seems consistent across data sets.

I have some questions about the methodology. First, the phylogenetic distance of first-time arriving species i is calculated to its closest relative that has already been observed prior to time point t. There would be other possibilities to compute the phylogenetic distance (e.g. mean pair-wise phylogenetic distance of species i to all community members). Could you justify the biological meaning of your choice to compute the phylogenetic distance? (e.g. are you assuming that competition based on niche similarity is the main mechanism preventing the entrance in the community?).

*****We now clarify why we used distance to the closest relative (instead of aggregate distance to the community). This is presented in our manuscript in context with previous studies that found that close relatives often join communities in short sequence. Thus, our choice of distance metric is a consequence of hypotheses born of previous research.**

The second methodological issue has to do with the uncertainty of phylogenetic reconstruction using short sequences. In my experience, FastTree is a deterministic algorithm, in the sense that it generates exactly the same tree irrespective of how many times you run it. I'd recommend using a different algorithm (e.g. RAxML) to generate replicated trees and checking for the consistency of the results. Please, give the replicated trees in the manuscript.

*****We now use IQtree for phylogeny construction instead of FastTree. Showing phylogenetic trees for high-throughput sequencing data sets is not feasible because any visualization of a tree with over 10,000 tips will simply be black mess. All input files are, however, available from our GitHub repository.**

Third, the Materials & Methods section lacks the description of statistical tests used to compare estimated D values across samples which are discussed in the manuscript (see specific comments below).

*****We agree with the reviewer that our hypothesis testing description was insufficient. That section has been re-written. Statistical comparison of D values across datasets is currently qualitative. Although we could have included a Tukey HSD test or FDR-corrected pair-wise unequal variance t-tests to compare bootstrapped D estimates, we do not think this is appropriate since all of those values are pseudoreplicates in the strict sense. Thus we have decided to err on the side of caution and present those comparisons qualitatively, but with 95% confidence intervals of estimates for easy visual comparison.**

My second general question deals with the discussion and conclusions of the manuscript. Working with experimental bacterial communities, Tan et al. (2012, Ecology 93: 1164-1172) showed that the immigration history determines the outcome of community assembly depending on the phylogenetic relatedness among members in the initial community. This makes me wonder how does the phylogenetic diversity of the initial (observed) community from which surrogate communities are constructed determine the outcome of the model (if it does at all). Tan et al. also showed an influence of species identity. Digging into these questions would help improve the level of the Discussion of this manuscript, which in my opinion does not conform to the general quality of the paper.

*****We added a third data source to our manuscript ("finnish infants"), and since this data source had 33 subjects, it was an ideal opportunity to do the analysis the reviewer suggested. We compared initial alpha-diversity and total alpha-diversity observed for each subject to the subject's estimated D parameter, but found no significant relationship. We do not present this result as testing any specific hypothesis informed by Tan et al. (2012), because we feel it would be misleading to present that finding as an a priori hypothesis. This lack of pattern is not directly comparable to the findings of Tan et al. (2012), since the assembly patterns in question in our studies are not directly comparable either.**

I think that the manuscript would gain if the authors would discuss deeper the biological (community assembly) mechanisms that can underlie the phylogenetic underdispersion detected (e.g. is this due to a strong abiotic filter?). And further try to unravel where do differences in estimated D across samples arise from (e.g. abiotic conditions, phylogenetic diversity in the species pool, community composition, abundance of a specific taxon?).

*****We have re-written our discussion to be more focused on the potential mechanistic (i.e. ecological) underpinnings of the patterns we observed. We also use the newly-added "Finnish infants" data source to dig deeper into potential differences between individuals that may impact D (diversity, richness, antibiotic use, sampling intensity) but found no significant patterns.**

Specific comments

The abstract needs, in my view, a more general conclusion. This is related to my general comments on the need of a discussion of the biological mechanisms underlying the results.

*****We have added "the human microbiome generally follows an assembly pattern characterized by phylogenetic underdispersion, i.e. nepotism" as a general conclusion within our abstract.**

P7. Please change “first sample of its corresponding empirical data set” to “first sample in the time series of its corresponding empirical data set” for clarity.

*****Done.**

P9. Are PDo and PDmax “phylodiversities” or “accumulated phylodiversities”?

*****They were accumulated, but we no longer discuss the scaling of PD values in the manuscript both for the sake of brevity and also because our new model fitting approach no longer requires that operation. Curves shown in figures are still scaled (since pd is arbitrary), but this scaling is now mentioned in the figure caption.**

P11. Please be more specific in defining the set of samples that were used, so that the work is fully reproducible

*****We have made our entire analysis is available on GitHub, and readers who want to know exactly which samples were excluded can easily find that information. We also clarify for the moving pictures data set exactly which logical conditions were used to retain or exclude samples.**

P13. Section “Results from ‘moving pictures’ data”
- Please, rephrase “moving pictures”, which is not evident here.
- Typo in second line
- Describe in Material & Methods sections, the statistical tests to compare D estimates described here

P13. Section “Results from infant gut data”
- Describe in Material & Methods sections, the statistical tests to compare D estimates described here (e.g. sharp increase around day 160). Through visual inspection maybe the increase around day 60 is as sharp as that on day 160

*****We have made the above changes, except in regard to comparison between D estimates which we respond to above.**

P14. The first two lines of the Discussion are redundant with the general ideas posed in the Introduction

*****This is intentional, we think it is good writing practice to remind the reader of “what the paper is about” when they reach the discussion. Context is easily lost after reading through statistically heavy methods and results.**

P15. “This “nepotistic” pattern in arrivals suggests that traits are driving community assembly in these human environments”. Traits could be driving other types of patterns indeed, depending on trait type and the extent to which it is phylogenetically conserved.

*****We agree with the reviewer. We have re-written our discussion of phylogenetic niche conservatism to add emphasis to traits, and how traits may or may not drive the patterns we model and observe.**