nature portfolio

Peer Review File

Differential roles of deterministic and stochastic processes in structuring soil bacterial ecotypes across terrestrial ecosystems

Corresponding Author: Dr Jingqiu Liao

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This is a large-scale study on the relationships between bacterial diversity and soil attributes, including soil chemistry and ecosystem type. Further to this the study provides evidence for how community assembly processes differ in these soils collectively, and among soil ecosystems, and bacteria defined based on ecotype (generalists, specialist, rare or abundant). The study will make a valuable contribution in terms of our understanding of bacterial diversity in soil.
The paper is for the most part clearly written, and was enjoyable to read. Multiple analytical approaches were used to draw consistent and robust conclusions with respect to environmental drivers.

Limitations include an uneven distribution of samples per ecosystem type, and use of a molecular data type that provides limited taxonomic resolution (16S rRNA gene amplicons) and limits the study to comparison based on phylogeny. Is also not novel as a stand-alone fact that soil diversity is influenced by environmental factors such as pH. However, these limitations do not majorly undermine the value of the studies findings, which provide a detailed analysis of bacterial diversity, and insights into community assembly processes, in a wide range of US soils.

Other comments:

Sampling across ecosystems is a key part of this study. As such, the ecosystems would benefit from a description. How were ecosystems defined in terms of plant species presence?

It should be stated upfront, how many sites there were per ecosystem type (it's not entirely clear from figure 1e), and how many samples there were per ecosystem type and site. This is particularly important to note upfront given the authors note barren and steppe/savanna ecosystems had relatively small sample sizes, which potentially influenced conclusions based on these ecosystems (lines 285-287).

A limitation of the study is the large disparity in sample numbers between ecotypes with almost 300 for forest and woodland, over 100 for herbaceous, ~40-50 for wetland and shrubland, and around half of that for steppe/savanna and barren. At the very least a caveat addressing this limitation should be provided as part of the conclusions. Moreover what is the reason for the disparity?

I particularly liked the distance decay analyses. However, the lack of increase with geographic distance observed for rare taxa and specialists is clearly due to very high beta-diversity at local scales, and it is worth elaborating on this (lines 336-339). In addition, considering the high beta-diversity observed for rare taxa across different spatial scales, why would "homogenizing dispersal and drift showed higher importance than dispersal limitation"?

Co-occurrence network interpretations: (1) definitions for node degree, closeness and betweenness would make this work more accessible. (2) Closeness, or the proximity of OTUs to one another, has been interpreted to mean bacteria have "rapid interactions". However, this dataset does not contain a temporal component to justify such a claim. (3) Bacterial phyla with high between are interpreted to be keystone species, which is logical. However, the conclusions refer to keystone taxa as rare, whereas these phyla appear to be relatively abundant (e.g., figure 1g). (4) Regarding ecosystems co-occurrences

shown in in figure 3g, what are the actual medians/means? Differences, such as relatively high closeness and low node degree and betweenness in shrublands, are not apparent from the figures as presented. Also, which, if any, ecosystems are they significantly higher/lower than?

It is unclear how the results of this study are consistent with parasitism (line 496). This is statement is poorly worded, and more care should be taken to distinguish between results here, and potential links to other studies.

I am not convinced the results unequivocally show "a more important role of rare taxa in maintaining bacterial interactions" (lines 499-500).

It is somewhat odd that windspeed is mentioned for the first time in the discussion (line 509), and not at all in the results (aside from being buried in a figure). As climatic factors can be important, it would be helpful to highlight those in the results.

Are the aluminum concentrations in this study consistent with nutrient uptake inhibition? (line 521)

Conclusions: These are both vague and overreaching. It is unclear how the results have identified "several critical control points for ecosystem resilience and resistance", nor what the "vital implications" are "for the environmental conservation and management of ecosystem services". Conclusions need to be well justified by the results and interpretation and discussion of those results.

Minor points:

The green and blue colours in figure 1h are hard to distinguish.

The colours in data points and the legend for figure 1e are different.

The data underlying figure 1i are unclear. Are the data points OTU counts per sample? If so, why are there so few rare OTUs, for example, when overall more rare than abundant OTUs were determined (870 versus 201)? Is there are large difference in samples for each of the ecotype? (lines 224 on)

(Remarks on code availability)
I have not checked all the code or run any code, but there is a README file, and code is present and well organised (as R files and ipython notebooks).

Reviewer #2

(Remarks to the Author)
This paper reports on a biogeography study, focussing on diversity patterns of soil bacterial communities across different biomes of the United States (US). The authors collected a large number of samples (622) and analysed the bacterial alpha and beta-diversity using eDNA metabarcoding and a series of statistical tools. Bacterial taxa were differentiated by their relative abundance between abundant or rare taxa, and by their occurrence between generalists or specialists. Several soil abiotic properties were also analysed and were compared to the bacterial diversity patterns. The authors found a high diversity of bacterial taxa across the ecotype of the US, and differing patterns of rare vs abundant taxa and generalist vs specialist taxa. They identified shrubland to be potentially the most sensitive to environmental changes, as it was the ecotype where the lowest diversity was found, and where the bacterial diversity was the most limited in its dispersal. They also identified calcium and aluminium, along pH, to influence the soil bacterial community patterns at large scale.

The most noteworthy results, in my opinion, are the differing dispersals mechanisms (stochastic or deterministic) of the abundant vs rare taxa and of the generalists vs specialists (shown in figure 4b). These very interesting patterns, to my knowledge (without having done a recent systematic search on it), have not been shown at such a large scale.

The authors' comprehensive statistical analysis of the extensive dataset makes this study highly valuable and significantly important within the field. This type of study, focussing on the biogeography of soil microbial communities and including a large number of samples from a large geographic scale, is not rare and has increased in the recent years (for example: Fierer and Jackson, PNAS 2006; Steven etal., FEMS Microbiology 2013; Nelson et al., PNAS 2016 to cite a few well-cited papers). Therefore, while the study's sampling design is not novel, its originality lies in the depth of its statistical analyses, which focus on various taxa categorized by their ecological niches or relative abundances.

The statistical analysis of the dataset is sound, support the conclusion of the manuscript and I did not spot any major flow in the data analysis, interpretation and conclusions. However, while the data analysis (main part of the work done in that study) is well described, the sampling design (already published) would be described in greater details to ease the comprehension of the reader. For example, until which soil depth were the samples collected? How many replicates per sites?

Another main concern is that the authors stated to use DADA2 to perform denoising and clustering of the sequenced reads (lines 628-631). Why were OTUs produced and not ASVs (amplicon sequences variants)? Please correct the method description of the pipeline. An explanation is needed why the authors choose to work with OTUs and not ASVs which are nowadays widely recognized among the microbiologist community to more accurately mirrored the composition and structure of natural microbial communities.

(Remarks on code availability)
Specific comments on the MS:
Introduction: the introduction is clear, concise and well written. However, the knowledge gaps could be specified more, explaining more precisely which mechanisms are not fully understood nowadays.
Methods:
Line 611: What does unknown ecosystem stand for? The type of ecosystem should be known (at least from referring to a map).
Line 634: Why were really rare OTUs (present <1% of samples) discarded from the dataset? These could actually be worth to include in the analyses as rare taxa. How the threshold of 1% was chosen as the limit to include of not OTUs in the dataset? 3158 OTUs for a dataset including more than 600 samples is actually a relatively small number of OTUs. I am afraid that too many rare species were discarded, which could have altered and biased the results of the study. Were tests made about that?
Lines 672: check typos.
Lines 692-698: Please add a few references on the chosen and used analytical methods.
The results are generally clearly described. However, the figures included in the main texts include too many panels (up to 9 panels). Due to the high numbers of these panels, their size is very limited, making the figure hard to see. I would suggest limiting the number of panels to maximum 2 or 3 panels, so that figures can be presented in bigger sizes, and that the reader is not overwhelmed by too much information. Only the most relevant results should be shown in the main text, the rest can be shown in the supplementary material.
Discussion:
Lines 491-496. “Relations” found in the co-occurrence network can only be described as co-occurring taxa. We can not infer that these co-occurences are actually correlations or real interaction of the taxa.
Lines 517-518: This finding is not novel. It is microbial ecology common knowledge that pH is usually the main factor influencing the microbial community structure among sites.
Lines 525-526: Potentially other non-tested factors (i.e. not investigated in this study) could be important to shape the bacterial community structure, such as the organic matter quality. A statement should be added on that.
Lines 549-553: This is to me the most important outcome of the study. I would suggest making it more prominent in the discussion, maybe starting the discussion with this.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have substantially revised the manuscript and addressed my original concerns. I have listed some additional comments below primarily for the purposes of clarifying some additional points.

Introduction line 93: "soil ... offers a unique perspective". This claim requires some justification. Why would soils offer a unique perspective? (The processes and ecosystem services noted earlier are not unique to soils).

Line 217 "As expected, abundant and rare taxa had high and low relative abundance across sites". This reads as a circular argument, as if ecotypes were tested for the attributes they were defined based on. This could be clarified/rectified by removing "As expected", otherwise clarify differences in methods.

Lines 296-297 Even spatial distributions and scales are also important given the fact diversity varies with distance.

It is of course unavoidable that the distributions of terrestrial ecosystems will differ across large spatial scales due to factors such as climate, elevation, and geology. The difference in sampled ecosystems distributions is shown in figure 1c (and in part in figure 2d), but could perhaps be shown in another way to enable easy and effective comparison across ecosystems, e.g. by plotting the pairwise distances between samples as a histogram per ecosystem. This is just a suggestion, but would be quite quick to do.

One last thought regarding ecosystem distributions, is it possible to generate a map like in figure 1c but with all USGS classified ecosystems to illustrate actual ecosystem distributions and hence justifying differences in this study? (I note that there is a map in https://doi.org/10.1016/j.gecco.2019.e00860 but the ecosystem categories used do not quite line up with those in this study).

Line 510 Is the diversity referred to here beta or alpha diversity? Could the authors add a brief explanation as to why soil communities in the US are highly diverse compared with countries such as Brazil?

Line 516. "Hazards" does not have the right connotation for extremophiles. Environmental hazards could be replaced with "environmental extremes" or "conditions such as ...".

Line 564 Is "extinction" the right term here? It seems a large leap to suggest extinction of abundant soil bacteria, rather than loss from a given area, without further analysis and/or discussion focused on permanent loss of diversity within an ecosystem.

Line 604 The conclusion here that intensive anthropogenic land use surrounding shrublands will lead to diversity loss in shrubland should be reconciled alongside the result shown in figure 2a, which shows little correlation between diversity and land development near shrubland.

Typos/minor grammatical errors:

Line 57 "based on the co-occurrence network" - the network has not been mentioned before in the abstract, so replace "the" with "a" or modify accordingly.

Line 208 Missing "in", as in "relative abundance in shrubland".

Line 344 "observed" rather than "overserved"

Lines 515 This should read something like "which is known as a phylum of extremophiles"

Line 597 "Geodermatophilushave obscurus"

(Remarks on code availability)
Extensive code with graphical outputs are provided. I have viewed this but not rerun the code.

Reviewer #2

(Remarks to the Author)
I thank the authors for their thorough revision of the manuscript. I only have a couple of minor comments that the author should still address:
1. Concerning my previous comment on the use of OTUs vs ASVs.
Line 672-673: please cite a reference for the argued reasons.

2. Concerning my previous comment on the threshold of the rare OTUs.
Thank you for demonstrating the similarity of the Shannon Index between rare taxa using a “narrow” or a “large”(test) threshold. I think it would be meaningful to mention this test and show figure B in the supplementary.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all my concerns in their revised manuscript.

(Remarks on code availability)
As noted previously, the authors have provided detailed code.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Reviewer #1 (Remarks to the Author):

This is a large-scale study on the relationships between bacterial diversity and soil attributes, including soil chemistry and ecosystem type. Further to this the study provides evidence for how community assembly processes differ in these soils collectively, and among soil ecosystems, and bacteria defined based on ecotype (generalists, specialist, rare or abundant). The study will make a valuable contribution in terms of our understanding of bacterial diversity in soil. The paper is for the most part clearly written, and was enjoyable to read. Multiple analytical approaches were used to draw consistent and robust conclusions with respect to environmental drivers.

Limitations include an uneven distribution of samples per ecosystem type, and use of a molecular data type that provides limited taxonomic resolution (16S rRNA gene amplicons) and limits the study to comparison based on phylogeny. Is also not novel as a stand-alone fact that soil diversity is influenced by environmental factors such as pH. However, these limitations do not majorly undermine the value of the studies findings, which provide a detailed analysis of bacterial diversity, and insights into community assembly processes, in a wide range of US soils.

We appreciate the reviewer's assessment and their recognition of the value, significance, and robustness of our study. We agree with the limitations raised by the reviewer. Our response to them is summarized below:

1. To ensure a spatially uniform sampling design, samples were not evenly collected within each ecosystem. Nevertheless, because of the methods that we employed in the data analysis, we believe our conclusions regarding the differences in microbial diversity and composition, distance-decay relationships, co-occurrence networks, and community assembly mechanisms compared among ecosystems are not biased by the uneven distribution of samples per ecosystem. The only analysis that may be biased is the correlation analysis between microbial diversity and environmental factors since sample size influences statistical power. We acknowledged this limitation and the need for further validation of these correlations using an even sample size per ecosystem (see Lines 293-284 and 296-297). A detailed explanation is provided below in response to the comment specific to this limitation.
2. We recognize the limitation of using 16S rRNA gene amplicon sequencing and have acknowledged the need to further understand the ecological mechanisms underlying soil bacterial diversity and community assembly at a higher resolution using metagenomic sequencing in the conclusions (see Lines 612-614).
3. While it is not novel to identify pH as a key environmental factor, we believe this is an important finding since we add one more piece of evidence (at a nationwide scale) that pH is a universal driver of soil microbial diversity. It also demonstrates the consistency of our study with findings from other studies. We have referred to these studies to support our findings (Lines 574-578). Importantly, besides pH, we also identified other two potential universal drivers, calcium, and aluminum (**Fig. 2a-b**) using multiple approaches, including machine learning models, which are novel in terms of both methodologies and findings. In addition, we quantified the importance of these environmental factors and associated them with bacterial community assembly processes (**Fig. 4d**), which we deemed uncommon in studies on soil diversity.

Other comments:

Sampling across ecosystems is a key part of this study. As such, the ecosystems would benefit from a description. How were ecosystems defined in terms of plant species presence?

We have added a detailed description of ecosystems, including information about plant species presence, in the revised manuscript (see Lines 637-648).

It should be stated upfront, how many sites there were per ecosystem type (it's not entirely clear from figure 1e), and how many samples there were per ecosystem type and site. This is particularly important to note upfront given the authors note barren and steppe/savanna ecosystems had relatively small sample sizes, which potentially influenced conclusions based on these ecosystems (lines 285-287).

We have clarified the sample size for each ecosystem in the second paragraph of the result section (Lines 191-193), in addition to mentioning it in the method section (Lines 634-636).

A limitation of the study is the large disparity in sample numbers between ecotypes with almost 300 for forest and woodland, over 100 for herbaceous, ~40-50 for wetland and shrubland, and around half of that for steppe/savanna and barren. At the very least a caveat addressing this limitation should be provided as part of the conclusions. Moreover what is the reason for the disparity?

We thank the reviewer for noting this limitation. While it is ideal to have an even distribution of samples per ecosystem, we believe the majority of our conclusions regarding the comparisons among ecosystems are robust due to the methods that we employed in the data analysis. For the differences in alpha diversity and relative abundance of bacterial taxa compared among ecosystems, we used Kruskal-Wallis's test, which is a nonparametric statistical test that allows for uneven sample sizes among groups. To compare beta diversity among ecosystems, we used PERMANOVA, which is also not restricted to equal sample sizes. The distance-decay relationships are built upon linear regression models. As long as the models are robust, the slope of the regression line, which indicates the distance-decay relationships, should not be influenced by the sample size. The importance of ecological processes to bacterial community assembly, is a relative metric calculated based on proportions. It is thus acceptable to compare their importance across ecosystems, even with uneven sample sizes.

The two analyses that may be influenced by uneven sample sizes across ecosystems are the correlations between bacterial diversity and environmental factors, and the co-occurrence networks, in which the underlying algorithm is also correlation. To avoid bias potentially caused by different sample sizes among ecosystems in the co-occurrence network analysis, we randomly selected an identical number of samples from each ecosystem to construct the network (see Lines 795-798). This effort ensures a fair comparison of co-occurrence networks among ecosystems. For the correlation analysis on bacterial diversity and environmental factors, we decided to use the real sample size rather than randomly selecting an identical number of samples for each ecosystem because the latter would lead to a significant drop in the statistical power, which is a missed opportunity for identifying key environmental factors for ecosystems with relatively large sample size. To address this limitation, we provided a caveat for the result that no environmental variables were identified for steppe/savanna and barren (see Lines 293-284). We also have acknowledged the need for further validation of these correlations using an even sample size per ecosystem (see Lines 296-297).

Regarding the disparity in sample sizes among ecosystems, it is due to the nature of the sampling design, in which we had a spatially even distribution of samples collected across the US. In brief, we divided the contiguous US into 40 equal-sized sampling grids, and an identical number of samples were collected within each sampling grid (see Lines 629-633). Since ecosystems are not evenly distributed across space, this sampling plan yielded different sample sizes among ecosystems. Detailed sampling design can be found in Liao et al. (2021).

I particularly liked the distance decay analyses. However, the lack of increase with geographic distance

observed for rare taxa and specialists is clearly due to very high beta-diversity at local scales, and it is worth elaborating on this (lines 336-339). In addition, considering the high beta-diversity observed for rare taxa across different spatial scales, why would "homogenizing dispersal and drift showed higher importance than dispersal limitation"?

We thank the reviewer for appreciating our distance-decay analyses and noticing the high beta-diversity at local scales for rare taxa and specialists. We have added this notion in the revised manuscript (Line 345).

Even though homogenizing dispersal was found to be more important than dispersal limitation for rare taxa, their difference was not large (**Fig. 4b**). Thus, their effects on the beta diversity of rare taxa probably cancel out. Both environmental heterogeneity and ecological drift can produce high beta diversity among communities (Chase, 2010; Chase & Myers, 2011; Segre et al., 2014). Given the importance of heterogeneous selection and drift to the assembly of rare taxa (**Fig. 4b**), we believe the high beta-diversity observed for rare taxa is mainly caused by the accumulative effects from these two ecological processes. We have added this interpretation in the revised manuscript (Lines 453-457).

Co-occurrence network interpretations: (1) definitions for node degree, closeness and betweenness would make this work more accessible. (2) Closeness, or the proximity of OTUs to one another, has been interpreted to mean bacteria have "rapid interactions". However, this dataset does not contain a temporal component to justify such a claim. (3) Bacterial phyla with high between are interpreted to be keystone species, which is logical. However, the conclusions refer to keystone taxa as rare, whereas these phyla appear to be relatively abundant (e.g., figure 1g). (4) Regarding ecosystems co-occurrences shown in in figure 3g, what are the actual medians/means? Differences, such as relatively high closeness and low node degree and betweenness in shrublands, are not apparent from the figures as presented. Also, which, if any, ecosystems are they significantly higher/lower than?

(1) We added the definitions for node degree, closeness, and betweenness centrality and provided the corresponding references in the revised manuscript (Lines 362-376)
(2) We changed the definition of closeness centrality to “Closeness centrality measures how distant a node is to other nodes and can be used to identify the most central OTUs within a community network” according to Kurtz et al (2015) and Zamkovaya et al (2020) (Lines 366-367).
(3) We modified the interpretation to “Members of a network with both high degree and betweenness centrality are typically the most connected taxa within the community and are considered “hubs” according to Zamkovaya et al (2020). Based on this, Thermi (a.k.a., Deinococcus-Thermus) phylum was identified as a hub in soil bacterial communities, which may play an essential ecological role in the community networks (see Lines 375-379 in the Results and Lines 514-517 in the Discussion). Of note, Thermi was not abundant in soils (**Supplementary Fig. 1a**). Also, mead relative abundance was standardized to enhance the comparison of each phylum among ecosystems in **Fig. 1g** (**Fig. 1d** in the revised manuscript). Thus, the circle sizes do not represent true mean relative abundance. We have clarified this in the figure caption.
(4) Since the Kruskal-Wallis test compares the medians of groups, we added the medians for both **Fig. 3d** and **g** (**Fig. 3c** and **e** in the revised manuscript) in the main text (Lines 397-405 and 417-427). We also removed outliers in **Fig. 3g** (now **Fig. 3e**) to improve clarity. In addition, we added the *post hoc* two-sided Mann-Whitney *U* tests for pairwise comparisons between ecosystems for **Fig. 3g** (now **Fig. 3e**) and described the results in detail in the main text (Lines 417-427).

It is unclear how the results of this study are consistent with parasitism (line 496). This is statement is poorly worded, and more care should be taken to distinguish between results here, and potential links to other studies.

Due to the concern raised by both reviewers, we decided to remove the discussion on positive interactions since co-occurrence may not be a good proxy for ecological interactions and our interpretation may be farfetched.

I am not convinced the results unequivocally show "a more important role of rare taxa in maintaining bacterial interactions" (lines 499-500).

We have changed to interpretation to “rare taxa were found to display stronger ecological relevance to the community than abundant taxa, evidenced by significantly higher node degree, betweenness, and closeness centrality” (Lines 517-519).

It is somewhat odd that windspeed is mentioned for the first time in the discussion (line 509), and not at all in the results (aside from being buried in a figure). As climatic factors can be important, it would be helpful to highlight those in the results.

We have highlighted significant climatic factors, including windspeed, precipitation, and temperature, in the result section (Lines 244-246 and 248-249).

Are the aluminum concentrations in this study consistent with nutrient uptake inhibition? (line 521)

Yes, the mean aluminum concentration of soil samples included in this study is 40.64 (mg/kg), seven times higher than the aluminum level (5 mg/kg) generally considered safe for plants in acidic soils (Groundcover, 2021). We have added this discussion in the revised manuscript (Lines 582-584).

Conclusions: These are both vague and overreaching. It is unclear how the results have identified "several critical control points for ecosystem resilience and resistance", nor what the "vital implications" are "for the environmental conservation and management of ecosystem services". Conclusions need to be well justified by the results and interpretation and discussion of those results.

We thank the reviewer for pointing this issue out. We have revised the conclusions by removing the statement about the critical control points and focusing on the key results on environmental factors, rare taxa, and shrubland ecosystems. We believe the revised conclusions are now well justified and clearer (see Lines 614-621).

Minor points:

The green and blue colours in figure 1h are hard to distinguish.

We have changed the green and blue colors in **Fig. 1h** (**Fig. 1e** in the revised manuscript) to make them more distinguishable. We have also reduced the number of panels in **Fig. 1**, so now this panel is larger and easier to see.

The colours in data points and the legend for figure 1e are different.

This is caused by the transparency that we added to the colors to distinguish overlapping samples (**Fig. 1c** in the revised manuscript). We confirm that the colors of individual points match the colors in the legend.

The data underlying figure 1i are unclear. Are the data points OTU counts per sample? If so, why are there so few rare OTUs, for example, when overall more rare than abundant OTUs were determined (870 versus 201)? Is there are large difference in samples for each of the ecotype? (lines 224 on)

Yes, the data points show OTU counts per sample in **Fig. 1i** (**Supplementary Fig. 5** in the revised manuscript). This is because while abundant taxa are common within each sample, many abundant taxa are shared among many samples. In contrast, each sample harbors fewer rare taxa than abundant taxa, but rare taxa within a sample tend to be unique (not found in many other samples). As a result, the total number of OTUs classified as rare taxa is much larger than abundant taxa,

Reviewer #1 (Remarks on code availability):

I have not checked all the code or run any code, but there is a README file, and code is present and well organised (as R files and ipython notebooks).

We thank the reviewer for this comment. We have updated the code accordingly for the updated analyses in the revised manuscript.

REFERENCE

Liao, J., Guo, X., Weller, D.L. *et al.* Nationwide genomic atlas of soil-dwelling *Listeria*reveals effects of selection and population ecology on pangenome evolution. *Nat Microbiol* **6**, 1021–1030 (2021).

Chase, J. M. Stochastic community assembly causes higher biodiversity in more productive environments. *Science*, **328**, 1388–1391 (2010).

Chase, J. M., & Myers, J. A. Disentangling the importance of ecological niches from stochastic processes across scales. *Philos Trans R Soc Lond B Biol Sci*, **366**, 2351–2363 (2011).

Segre, H., Ron, R., De Malach, N., Henkin, Z., Mandel, M., & Kadmon, R. Competitive exclusion, beta diversity, and deterministic vs. stochastic drivers of community assembly. Ecol Lett, **17**, 1400–1408 (2014).

Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11**, e1004226 (2015).

Zamkovaya, T., Foster, J. S., de Crécy-Lagard, V. & Conesa, A. A network approach to elucidate and prioritize microbial dark matter in microbial communities. *ISME J* **15**, 228–244 (2020).

Test methods a concern with soil aluminum | Groundcover. https://groundcover.grdc.com.au/agronomy/soil-and-nutrition/test-methods-a-concern-with-soil-aluminium (2021).

Reviewer #2 (Remarks to the Author):

This paper reports on a biogeography study, focussing on diversity patterns of soil bacterial communities across different biomes of the United States (US). The authors collected a large number of samples (622) and analysed the bacterial alpha and beta-diversity using eDNA metabarcoding and a series of statistical tools. Bacterial taxa were differentiated by their relative abundance between abundant or rare taxa, and by their occurrence between generalists or specialists. Several soil abiotic properties were also analysed and were compared to the bacterial diversity patterns. The authors found a high diversity of bacterial taxa across the ecotype of the US, and differing patterns of rare vs abundant taxa and generalist vs specialist taxa. They identified shrubland to be potentially the most sensitive to environmental changes, as it was the ecotype where the lowest diversity was found, and where the bacterial diversity was the most limited in its dispersal. They also identified calcium and aluminium, along pH, to influence the soil bacterial community patterns at large scale.

The most noteworthy results, in my opinion, are the differing dispersals mechanisms (stochastic or deterministic) of the abundant vs rare taxa and of the generalists vs specialists (shown in figure 4b). These very interesting patterns, to my knowledge (without having done a recent systematic search on it), have not been shown at such a large scale.

The authors' comprehensive statistical analysis of the extensive dataset makes this study highly valuable and significantly important within the field. This type of study, focussing on the biogeography of soil microbial communities and including a large number of samples from a large geographic scale, is not rare and has increased in the recent years (for example: Fierer and Jackson, PNAS 2006; Steven etal., FEMS Microbiology 2013; Nelson et al., PNAS 2016 to cite a few well-cited papers). Therefore, while the study's sampling design is not novel, its originality lies in the depth of its statistical analyses, which focus on various taxa categorized by their ecological niches or relative abundances.

We thank the reviewer's recognition of the value of our study and the novelty of our data analysis.

The statistical analysis of the dataset is sound, support the conclusion of the manuscript and I did not spot any major flow in the data analysis, interpretation and conclusions. However, while the data analysis (main part of the work done in that study) is well described, the sampling design (already published) would be described in greater details to ease the comprehension of the reader. For example, until which soil depth were the samples collected? How many replicates per sites?

We have added more details regarding sampling design and the definition of ecosystems in the method section (see Lines 629-633 and 637-648).

Another main concern is that the authors stated to use DADA2 to perform denoising and clustering of the sequenced reads (lines 628-631). Why were OTUs produced and not ASVs (amplicon sequences variants)? Please correct the method description of the pipeline. An explanation is needed why the authors choose to work with OTUs and not ASVs which are nowadays widely recognized among the microbiologist community to more accurately mirrored the composition and structure of natural microbial communities.

We agree with the reviewer’s opinion on the advantages of ASVs. While OTUs are often less precise than ASVs, we chose OTUs over ASVs because of legacy data comparison, broader ecological groupings (e.g., ecotypes used in this study), tolerance to sequencing errors, less sensitivity to hypervariable regions, and computational efficiency. We have added an explanation in the revised manuscript (Lines 670-673) and pointed out the need to further study microbial biogeography at a higher resolution using ASV or metagenomic sequencing in the conclusions (Lines 612-614).

Specific comments on the MS:

Introduction: the introduction is clear, concise and well written. However, the knowledge gaps could be specified more, explaining more precisely which mechanisms are not fully understood nowadays.

We thank the reviewer's assessment of the introduction. We have rewritten the knowledge gaps to make it more specific: *yet ecological attributes and the drivers governing their composition and distribution, especially for taxa differing in ecological traits and inhabiting different terrestrial ecosystems, are not fully understood* (Lines 47-48).

Methods:
Line 611: What does unknown ecosystem stand for? The type of ecosystem should be known (at least from referring to a map).

Sites classified as unknown ecosystems are areas with a small pixel count (<=20,000) in the USGS classification system. We added an explanation to this in the revised manuscript (Lines 647-648).

Line 634: Why were really rare OTUs (present <1% of samples) discarded from the dataset? These could actually be worth to include in the analyses as rare taxa. How the threshold of 1% was chosen as the limit to include of not OTUs in the dataset? 3158 OTUs for a dataset including more than 600 samples is actually a relatively small number of OTUs. I am afraid that too many rare species were discarded, which could have altered and biased the results of the study. Were tests made about that?

We fully understand the concern raised by the reviewer. To enhance the quality and interpretability of microbial community analyses (Bokulich, et al., 2013), we took a conservative approach to remove OTUs with low frequency and abundance that often result from sequencing errors, contamination, or random sampling artifacts, by referring to the thresholds used in Wilhelm et al. (2023) (note: they removed OTUs present in less than 10 samples out of 778, which is about 1.3%). We have added the reasoning and these references in the revised manuscript (Lines 673-677).

Since the cutoff chosen to define rare taxa is based on the mean relative abundance across samples for all OTUs, using a less conservative way to remove OTUs will change the distribution of the mean relative abundance for all OTUs, influencing the cutoff chosen to define rare taxa and their mean relative abundance. Thus, we hypothesize that the diversity of OTUs classified as rare taxa will not significantly change using a less conserved method to remove OTUs compared to that calculated in this study. To test this hypothesis, we re-analyzed the data by only removing singletons (i.e., OTUs present in one sample with one read sequenced), which yielded 7,869 OTUs. We chose 0.0007% as the cutoff for rare taxa since it is approximately the median of the mean relative abundance for all these OTUs (**Fig. a below**), consistent with the approach to classifying rare taxa used in this study. With this cutoff, a total of 3,709 OTUs were classified as rare taxa. We further calculated the Shannon-Wiener diversity of these rare taxa and compared it with that of rare taxa classified in this study. We observed no significant difference between these two groups (median = 0.072 and 0.068, respectively; Mann-Witney *U P* = 0.18; **Fig. b below**). In addition, we used multivariate cutoff level analysis (MultiCoLA), a strategy to systematically assess the influence of rarity definition on large community datasets, and confirmed the rationality of the chosen cutoffs for defining abundant and rare taxa used in this study (Lines 714-721). Overall, we concluded that our results were not biased by the methods that we used to remove OTUs with low frequency and abundance.

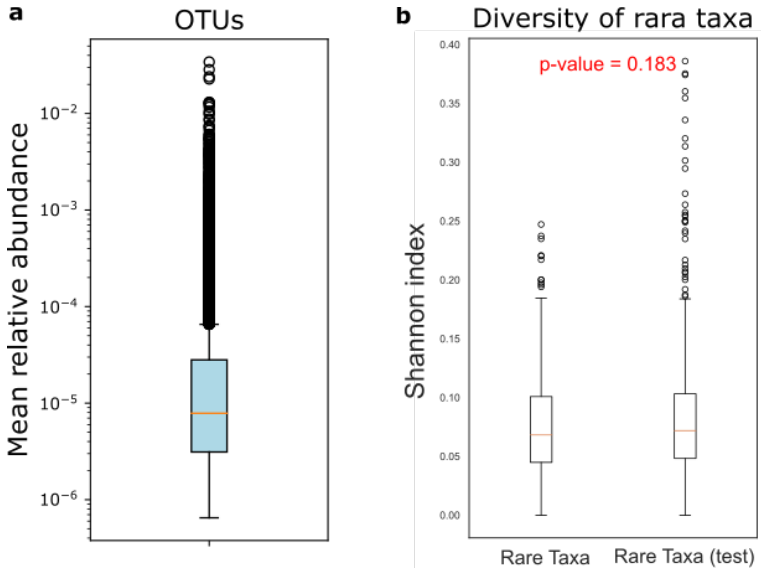

Fig. a, Mean relative abundance of OTUs across samples (test dataset). b, Diversity of rare taxa compared between the dataset used in this study and the test dataset. The dataset used in this study: OTUs present in < 1% of samples and with total reads < 10 were removed. Test dataset: singletons (i.e., OTUs present in one sample with one read sequenced) were removed. Box plots show the interquartile range (IQR), with the line representing the median and whiskers extending to 1.5 times the IQR. *P*, two-sided Mann-Witney *U* test.

Lines 672: check typos.

Corrected.

Lines 692-698: Please add a few references on the chosen and used analytical methods.

We have added a few references for the binomial distribution model (Line 741).

The results are generally clearly described. However, the figures included in the main texts include too many panels (up to 9 panels). Due to the high numbers of these panels, their size is very limited, making the figure hard to see. I would suggest limiting the number of panels to maximum 2 or 3 panels, so that figures can be presented in bigger sizes, and that the reader is not overwhelmed by too much information. Only the most relevant results should be shown in the main text, the rest can be shown in the supplementary material.

We agree with the reviewer's comment. Given the depth of our data analysis, we wanted to have key results about all OTUs, ecotypes, and ecosystems presented in the main figures, which led to many panels in each figure. After careful consideration, we have re-arranged the panels. Now, the number of panels for **Fig. 1**, **2**, **3**, and **4** has decreased to 5, 4, 5, and 4, and the number of supplementary figures has increased to 12 in the revised manuscript. We believe the panels that remained in the main figures represent the most relevant results in this study.

Discussion:
Lines 491-496. "Relations" found in the co-occurrence network can only be described as co-occurring taxa. We can not infer that these co-occurences are actually correlations or real interaction of the taxa.

Due to the concern raised by both reviewers, we decided to remove the discussion on positive interactions since co-occurrence may not be a good proxy for ecological interactions and our interpretation may be farfetched.

Lines 517-518: This finding is not novel. It is microbial ecology common knowledge that pH is usually the main factor influencing the microbial community structure among sites.

While it is not novel to identify pH as a key environmental factor, we believe this is an important finding since we add one more piece of evidence (at a nationwide scale) that pH is a universal driver of soil microbial diversity. It also demonstrates the consistency of our study with findings from other studies. We thus decided to keep this discussion and have referred to these other studies to support our findings (Lines 574-578).

Lines 525-526: Potentially other non-tested factors (i.e. not investigated in this study) could be important to shape the bacterial community structure, such as the organic matter quality. A statement should be added on that.

We have added this statement in the revised manuscript (Lines 331-332).

Lines 549-553: This is to me the most important outcome of the study. I would suggest making it more prominent in the discussion, maybe starting the discussion with this.

We have moved the discussion on the bacterial community assembly up (Lines 525-564).

REFERENCE

Bokulich, N. A., et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*, **10**, 57–59 (2013).

Wilhelm, R. C., Amsili, J. P., Kurtz, K. S. M., van Es, H. M. & Buckley, D. H. Ecological insights into soil health according to the genomic traits and environment-wide associations of bacteria in agricultural soils. *ISME Commun* **3**, 1–11 (2023).

Reviewer #1 (Remarks to the Author):

The authors have substantially revised the manuscript and addressed my original concerns. I have listed some additional comments below primarily for the purposes of clarifying some additional points.

We thank the reviewer's assessment and have addressed the additional comments as detailed below.

Introduction line 93: "soil ... offers a unique perspective". This claim requires some justification. Why would soils offer a unique perspective? (The processes and ecosystem services noted earlier are not unique to soils).

We added “Soil is a complex, dynamic ecosystem acting as a living interface between the atmosphere, lithosphere, and hydrosphere, which sustains a wide range of ecological processes (e.g., nitrogen and carbon cycling) critical for life on Earth.” to highlight the importance and uniqueness of soils (Lines 88-90).

Line 217 "As expected, abundant and rare taxa had high and low relative abundance across sites". This reads as a circular argument, as if ecotypes were tested for the attributes they were defined based on. This could be clarified/rectified by removing "As expected", otherwise clarify differences in methods.

We removed “As expected” (Line 219).

Lines 296-297 Even spatial distributions and scales are also important given the fact diversity varies with distance.

We revised this sentence to “These findings, however, necessitate further validation using samples with even sample sizes, spatial distributions, and scales for each ecosystem type.” (Lines 298-299).

It is of course unavoidable that the distributions of terrestrial ecosystems will differ across large spatial scales due to factors such as climate, elevation, and geology. The difference in sampled ecosystems distributions is shown in figure 1c (and in part in figure 2d), but could perhaps be shown in another way to enable easy and effective comparison across ecosystems, e.g. by plotting the pairwise distances between samples as a histogram per ecosystem. This is just a suggestion, but would be quite quick to do.

We thank the reviewer's suggestion. We created these histograms (see **Fig. b** below for convenience) and have included them in the revised manuscript (Lines 195-196; **Supplementary Fig. 2b**).

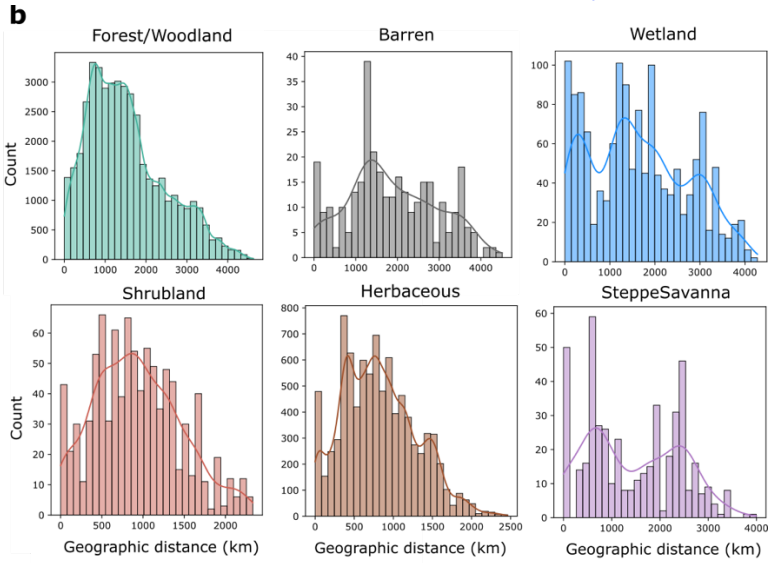

Fig. b, Histograms showing the distribution of pairwise geographic distance (km) between samples for each ecosystem.

One last thought regarding ecosystem distributions, is it possible to generate a map like in figure 1c but with all USGS classified ecosystems to illustrate actual ecosystem distributions and hence justifying differences in this study? (I note that there is a map in https://doi.org/10.1016/j.gecco.2019.e00860 but the ecosystem categories used do not quite line up with those in this study).

We thank the reviewer for this comment. Since USGS has already generated the map that the reviewer suggested (see Fig. 8 in https://pubs.usgs.gov/pp/1768/), we find it more appropriate to directly cite this map than to repeat the effort. We referred to this USGS map and provided appropriate reference (Ref. 60) in the revised manuscript (Lines 636-637).

Line 510 Is the diversity referred to here beta or alpha diversity? Could the authors add a brief explanation as to why soil communities in the US are highly diverse compared with countries such as Brazil?

We refer to both alpha and beta diversity. We have clarified this by saying “soil bacteria in the terrestrial ecosystems of the contiguous US were highly diverse within and across locations.” (Lines 512-513). We realize that the higher bacterial diversity in the US compared to other countries could be an artifact due to different sampling and data processing methods being used. It may not be scientifically sound to make this claim solely based on several publications without robust meta-data analysis. Thus, we have removed this interpretation in the revised manuscript.

Line 516. "Hazards" does not have the right connotation for extremophiles. Environmental hazards could be replaced with "environmental extremes" or "conditions such as ...".

We thank the reviewer for pointing this out. We replaced it with “environmental extremes” (Lines 517-518).

Line 564 Is "extinction" the right term here? It seems a large leap to suggest extinction of abundant soil bacteria, rather than loss from a given area, without further analysis and/or discussion focused on permanent loss of diversity within an ecosystem.

We agree with the reviewer. We replaced “extinction” with “biodiversity loss” (Line 566).

Line 604 The conclusion here that intensive anthropogenic land use surrounding shrublands will lead to diversity loss in shrubland should be reconciled alongside the result shown in figure 2a, which shows little correlation between diversity and land development near shrubland.

We apologize for the confusion. In **Fig. 2a** (bottom panel), each land-use variable on the y-axis was measured as the proportion of each land-use type in the surrounding areas for a given sampling site. Variables on the x-axis show different ecosystems. As shown in this heatmap, six out of ten land-use variables (i.e., barren, forest, shrubland, cropland, pasture, and wetland) exhibit significant correlations with bacterial alpha diversity within the shrubland ecosystem. In addition, in **Fig. 2c**, VPA results show that more than 50% of the variation of the bacterial alpha diversity within the shrubland ecosystem can be explained by land-use patterns in surrounding areas. Thus, we believe our conclusion regarding the potential effects of anthropogenic land use in surrounding areas on bacterial diversity in the shrubland ecosystem is well supported by these results.

Typos/minor grammatical errors:

Line 57 "based on the co-occurrence network" - the network has not been mentioned before in the abstract, so replace "the" with "a" or modify accordingly.

We replaced “the” with “a” (Line 57).

Line 208 Missing "in", as in "relative abundance in shrubland".

We added “in” (Line 211).

Line 344 "observed" rather than "overserved"

Revised (Line 347).

Lines 515 This should read something like "which is known as a phylum of extremophiles"

Revised accordingly (Line 517).

Line 597 "Geodermatophilushave obscurus"

Revised to “*Geodermatophilus obscurus*” (Lines 212 and 599).

Reviewer #1 (Remarks on code availability):

Extensive code with graphical outputs are provided. I have viewed this but not rerun the code.

We thank the reviewer for the assessment.

Reviewer #2 (Remarks to the Author):

I thank the authors for their thorough revision of the manuscript. I only have a couple of minor comments that the author should still address:

1. Concerning my previous comment on the use of OTUs vs ASVs.
Line 672-673: please cite a reference for the argued reasons.

We added a reference for this (Line 675).

2. Concerning my previous comment on the threshold of the rare OTUs.
Thank you for demonstrating the similarity of the Shannon Index between rare taxa using a “narrow” or a “large”(test) threshold. I think it would be meaningful to mention this test and show figure B in the supplementary.

We described this test (Lines 725-735) and included both figure A and B as supplementary figures (**Supplementary Fig. 13a, b**) in the revised manuscript.