

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

Received: 1 October 2024

Accepted: 21 February 2025

Published online: 08 March 2025


 Check for updates

Mia Riddley^{1,7}, Shannon Hepp^{1,7}, FNU Hardeep², Aruj Nayak², Meimei Liu³, Xin Xing³, Hailong Zhang⁴ & Jingqiu Liao^{1,5,6} 

Soil bacteria are vital to ecosystem resilience and resistance, yet ecological attributes and the drivers governing their composition and distribution, especially for taxa varying in ecological traits and inhabiting different ecosystems, are not fully understood. Here, we analyzed a large-scale bacterial community and environmental dataset of 622 soil samples systematically collected by us from six major terrestrial ecosystems across the United States. We show that soil bacterial diversity and composition significantly differ among ecotypes and ecosystems, partially determined by a few universal abiotic factors (e.g., soil pH, calcium, and aluminum) and several ecotype- or ecosystem-specific ecological drivers. Co-occurrence network analysis suggests that rare taxa have stronger ecological relevance to the community than abundant taxa. Ecological models revealed that deterministic processes shape assembly of abundant taxa and generalists, while stochastic processes played a greater role in rare taxa and specialists. Also, bacterial communities in the shrubland ecosystem appear to be more sensitive to environmental changes than other ecosystems, evidenced by the lowest diversity, least connected community network, and strongest local environmental selection driven by surrounding land use. Overall, this study reveals ecological mechanisms underlying the bacterial biogeography in terrestrial ecosystems nationwide and highlights the need to preserve rare biosphere and shrubland ecosystems amid environmental disturbance.

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. Soil bacteria are diverse and participate in a multitude of these ecological processes, contributing to

the resilience and resistance of terrestrial ecosystems¹. Growing evidence has shown that like macroorganisms, bacteria in a wide range of environments, including soils, display biogeographic patterns, such as the distance-decay relationship^{2,3}. Studying the biogeographic patterns of the collective output of all the bacteria

¹Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA 24061, USA. ²Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA. ³Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA. ⁴Department of Business Information Technology, Virginia Tech, Blacksburg, VA 24061, USA. ⁵Global Change Center, Virginia Tech, Blacksburg, VA 24061, USA. ⁶Center for Emerging, Zoonotic, and Arthropod-Borne Pathogens, Virginia Tech, Blacksburg, VA 24061, USA. ⁷These authors contributed equally: Mia Riddley, Shannon Hepp.  e-mail: liaoqj@vt.edu

present in soils (i.e., bacterial communities) offers a unique perspective to understanding the ecological consequences of environmental changes mediated by bacteria.

Bacterial taxa vary by ecological traits, representing different ecotypes. Two ecological traits that have attracted great attention recently are abundance and habitat range^{4–7}. The former trait defines abundant taxa (i.e., taxa that have a larger local abundance) and rare taxa (i.e., taxa that have a smaller local abundance), while the latter defines habitat generalists (i.e., taxa that have a wider habitat preference) and habitat specialists (i.e., taxa that have a more limited habitat range). These ecotypes can vary in metabolic capabilities and environmental adaptability, performing specialized functions that are integral to the health and sustainability of ecosystems⁸. In addition to ecotypes, soil bacterial communities can vary in diversity and composition across the terrestrial ecosystems that they inhabit, where unique environmental characteristics may lead to distinct positive and negative feedback occurring between bacteria and their environment⁹. Thus, a comprehensive understanding of soil bacterial biogeography from a mechanistic standpoint for different ecotypes and terrestrial ecosystems is essential for maintaining soil ecosystem balance and functionality and can enhance the predictability of ecological consequences of bacterial communities in response to environmental disturbance.

However, deciphering the community assembly mechanisms that structure bacterial diversity and biogeography remains a major challenge^{10–13}. Fundamental concepts such as niche-based and neutral theories¹⁴ have provided an ecological framework for understanding community assembly mechanisms. According to the niche theory, deterministic processes that impose environmental selection triggered by abiotic factors (e.g., nutrients) and biotic interactions (e.g., mutualism and competition) play a primary role in community assembly¹⁵. Environmental selection can lead to more phylogenetically or taxonomically dissimilar and similar community structures, termed heterogeneous and homogeneous selection, respectively¹⁶. Indeed, many studies have identified a variety of environmental variables important to bacterial community structure^{17–19}. For example, pH has been reported to be a universal indicator of bacterial diversity²⁰. In contrast to niche-based theory, neutral theory emphasizes the importance of stochastic processes, including homogenizing dispersal, dispersal limitation, and ecological drift, pointing towards a cause and effect on community structure²¹. Many studies have shown that stochastic processes, jointly with deterministic processes, can influence bacterial community assembly^{22,23}. Subtle environmental changes can lead to substantial shifts from stochastic to deterministic dominance in a bacterial community and vice versa²⁴. However, how microbial communities are governed by these ecological processes at a large spatial scale and how their importance differs by ecotypes and ecosystems remain largely unknown.

In this work, we characterized the nationwide biogeographic patterns of soil bacteria and investigated the underlying community assembly mechanisms for different ecotypes, including abundant and rare taxa and habitat generalists and specialists, as well as different terrestrial ecosystems. We analyzed bacterial community data of 622 soil samples representing six major terrestrial ecosystems (forest/woodland, shrubland, wetland, herbaceous, steppe/savanna, and barren) collected by us across the United States (US) within a consistent time frame, which was paired with 34 environmental variables capturing geolocation, soil properties, climate, and surrounding land use. With in-depth ecological analyses, we identified distinct patterns of biodiversity, spatial distribution, and abiotic and biotic drivers specific for ecotypes and across different ecosystems. Of note, rare taxa were found to have stronger ecological relevance to the community than abundant taxa based on co-occurrence network analysis. Using a phylogenetic-based approach²⁵, we further disentangled the co-working of deterministic and stochastic mechanisms underlying

bacterial community assembly. We found that deterministic processes were more important for abundant taxa and generalists, while stochastic processes were more vital for rare taxa and specialists. In addition, among all ecosystems, bacterial communities in shrubland exhibited the highest sensitivity to environmental disturbances, particularly those resulting from surrounding land use changes. Collectively, this study deepens a mechanistic and predictive understanding of the biogeography of soil bacteria at a nationwide scale and reveals the underlying ecological mechanisms at a resolution of ecotypes and terrestrial ecosystems.

Results

Diversity and composition of bacterial communities across the US

Among the 622 soil samples we collected, data for 618 samples passed the quality control and were included in the analysis (see Methods). A total of 3158 OTUs were identified, representing 31 bacterial phyla (Fig. 1a), with Actinobacteria and Proteobacteria being the most prevalent (Supplementary Fig. 1a). A count of 104 known classes, 205 orders, 333 families, 517 genera, and 559 species were detected. Four α -diversity metrics were calculated, including OTU richness, Shannon-Wiener diversity, Faith's phylogenetic diversity (PD), and Simpson's evenness. Bacterial richness and PD were approximately normally distributed with a range expanding from 60 – 531 and 8.9 – 42.8, respectively, and with an average of 288 and 25.5, respectively (Supplementary Fig. 1b). The distribution of Shannon-Wiener diversity and Simpson's evenness, however, was left-skewed (Supplementary Fig. 1b). Shannon-Wiener diversity ranged from 2.5 – 7.8 with an average of 6.89, and Simpson's evenness ranged from 0.411 – 0.931 with an average of 0.86. Since Shannon-Wiener diversity was highly significantly correlated with all other three α -diversity metrics (Spearman $\rho > 0.7$, $P < 0.05$ for all; Supplementary Fig. 1c), this metric was selected for downstream analysis to represent α -diversity. To determine the dissimilarity of bacterial community structure among samples, four β -diversity metrics, Bray-Curtis, Jaccard, unweighted UniFrac, and weighted UniFrac distances, were measured. All these metrics indicated large variation of β -diversity across samples (Supplementary Fig. 1d). For example, nearly 62% and 93% of Bray-Curtis and Jaccard measurements, respectively, were above 0.7 (Supplementary Fig. 1d). Since weighted UniFrac distance considers relative abundance and phylogenetic diversity and was highly significantly correlated with all other three β -diversity metrics (Mantel $\rho > 0.84$, $P < 0.05$ for all; Supplementary Fig. 1e), weighted UniFrac distance was used for downstream analysis to represent β -diversity. These results indicate that soil bacterial diversity is overall high, and bacteria community composition varies between locations across the US.

Among the 6 major terrestrial ecosystems that were sampled, the forest/woodland ecosystem is the most prevalent (298 samples), followed by herbaceous (134), wetland (54), shrubland (44), steppe/savanna (31), and barren (26) ecosystems (Supplementary Fig. 2a). Samples for these ecosystems varied by spatial scales and distributions (Supplementary Fig. 2b). Bacterial α -diversity, indicated by the Shannon-Wiener diversity index, was found to significantly differ among ecosystems, with forest/woodland and shrubland having the highest and lowest diversity, respectively (Kruskal-Wallis [KW] $P = 0.00866$; Fig. 1b). Also, bacterial α -diversity varied geographically across the US. The Midwest and Southeast, where many samples were collected from forest/woodland and herbaceous ecosystems, contained more diverse soil microbes compared to other regions (Fig. 1c). In addition, these ecosystems had significantly different bacterial community compositions between locations indicated by β -diversity (PERMANOVA $P = 0.001$), forming clusters of samples by ecosystems in the multidimensional scaling (MDS) analysis (Supplementary Fig. 3a). A total of 18, 53, 98, 73, 98, 76, 16, 447 bacterial phyla, classes, orders, families, genera, species, and OTUs were found to have

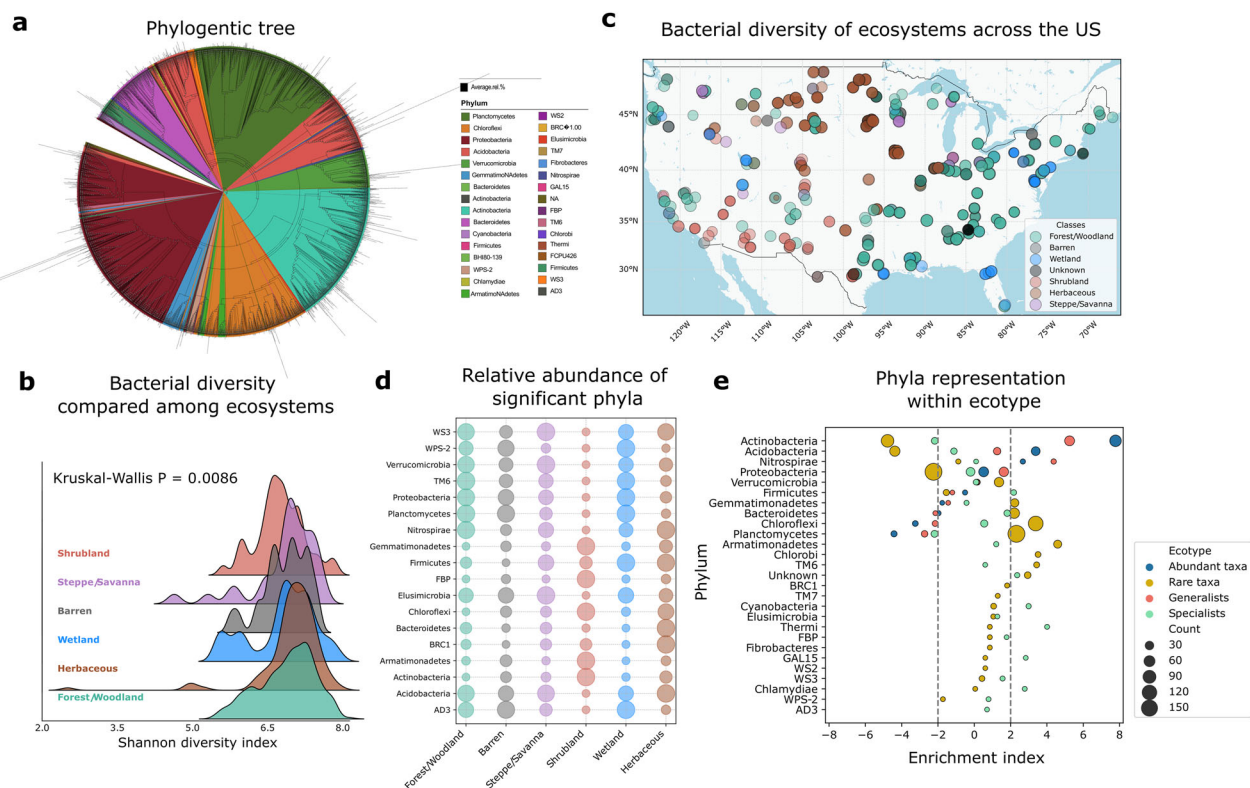


Fig. 1 | Diversity and composition of bacterial communities vary by terrestrial ecosystems and ecotypes across the US. **a** The maximum likelihood phylogenetic tree of 3158 OTUs constructed using 16S rRNA gene sequences with 1000 bootstraps. The tree is rooted by midpoint and the branches are color-coded by the phylum that each OTU represents. The outer annotation indicates the relative abundance of OTUs. **b** Ridge plot showing the bacterial diversity indicated by Shannon-Wiener diversity index compared among ecosystems, sorted by median. Kruskal-Wallis (KW) $P < 0.05$ indicates a significant difference. **c** Distribution of soil bacterial diversity across the US. Circles are color-coded by ecosystems and

circle size is proportional to Shannon-Wiener diversity index. **d** Bubble plot showing the mean relative abundance of phyla significantly different among ecosystems (adjusted KW $P < 0.05$). Mean relative abundance of each phylum was standardized to enhance ecosystem comparison, so circle sizes do not reflect true values. **e** Enrichment analysis for the phyla composition for each ecotype. The circle size denotes the number of OTUs representing a phylum for a given ecotype. Phyla with an enrichment index > 2 and < -2 indicates significant overrepresentation and underrepresentation within each ecotype, respectively.

significantly different relative abundance compared among ecosystems, respectively (adjusted KW $P < 0.05$ for all; Fig. 1d; Supplementary Data 1-5, Supplementary Fig. 3b). Of note, a number of taxa were found to be uniquely abundant in the shrubland ecosystem. For example, at the phylum level, Gemmatimonadetes, Chloroflexi, Armatimonadetes, and Actinobacteria had a significantly higher relative abundance in shrubland than other ecosystems (Fig. 1d). At the species level, *Nocardioides dilutus*, *Pseudonocardia halophobica*, *Virgisporangium ochraceum*, and *Geodermatophilus obscurus* had a significantly higher relative abundance in shrubland (Supplementary Fig. 3b). Overall, bacterial communities within different terrestrial ecosystems were significantly different in diversity and composition, which may be attributed to the unique environmental conditions that we observed among these ecosystems based on the MDS analysis and KW tests (Supplementary Fig. 3c and 2d, respectively).

OTUs detected in this study were classified into four ecotypes, abundant taxa, rare taxa, generalists, and specialists (see Methods). Abundant and rare taxa had high and low mean relative abundance across sites, respectively, while generalists and specialists had high and low site prevalence, respectively (Supplementary Fig. 4a). A total of 201 OTUs representing 10 phyla were classified as abundant taxa (Supplementary Fig. 4b). Based on the enrichment analysis, Actinobacteria, Acidobacteria, and Nitrospirae were significantly enriched among abundant taxa ($P < 0.05$ for all; Fig. 1e). A total of 870 OTUs representing 26 phyla were classified as rare taxa (Supplementary Fig. 4c), and among them, 7 phyla were significantly enriched, including

Gemmatimonadetes, Bacteroidetes, Chloroflexi, Planctomycetes, Armatimonadetes, Chlorobi, and TM6 ($P < 0.05$ for all; Fig. 1e). A total of 156 OTUs representing 10 phyla were classified as generalists (Supplementary Fig. 4d), and among them, Actinobacteria and Nitrospirae were significantly enriched ($P < 0.05$ for both; Fig. 1e). A similar number of OTUs (159) were classified as specialists, but they represented a set of more diverse phyla (22) compared to generalists (Supplementary Fig. 4e). There were 4 phyla significantly enriched among specialists, including Cyanobacteria, Thermi, GAL15, and Chlamydiae ($P < 0.05$ for all; Fig. 1e). Except for rare taxa, the number of OTUs representing each ecotype was significantly different among ecosystems (adjusted KW $P < 0.05$ for all). For example, herbaceous, steppe/savanna, and forest/woodland ecosystems harbored significantly more abundant taxa and generalists, while shrubland had the highest number of specialists but the lowest number of abundant taxa and generalists (Supplementary Fig. 5). These results highlighted distinct ecological roles of bacterial taxa across terrestrial ecosystems.

Influence of environmental factors on soil bacterial diversity

To understand the potential influence of environmental factors on soil bacterial α -diversity, we performed Spearman's rank correlation analysis between Shannon-Wiener diversity index and the 34 environmental variables representing geolocation, soil properties, climate, and surrounding land use (Fig. 2a). For all taxa, 11 out of 34 variables (32.4%) were significantly positively correlated with α -diversity, including soil pH, calcium, potassium, magnesium, molybdenum, wind

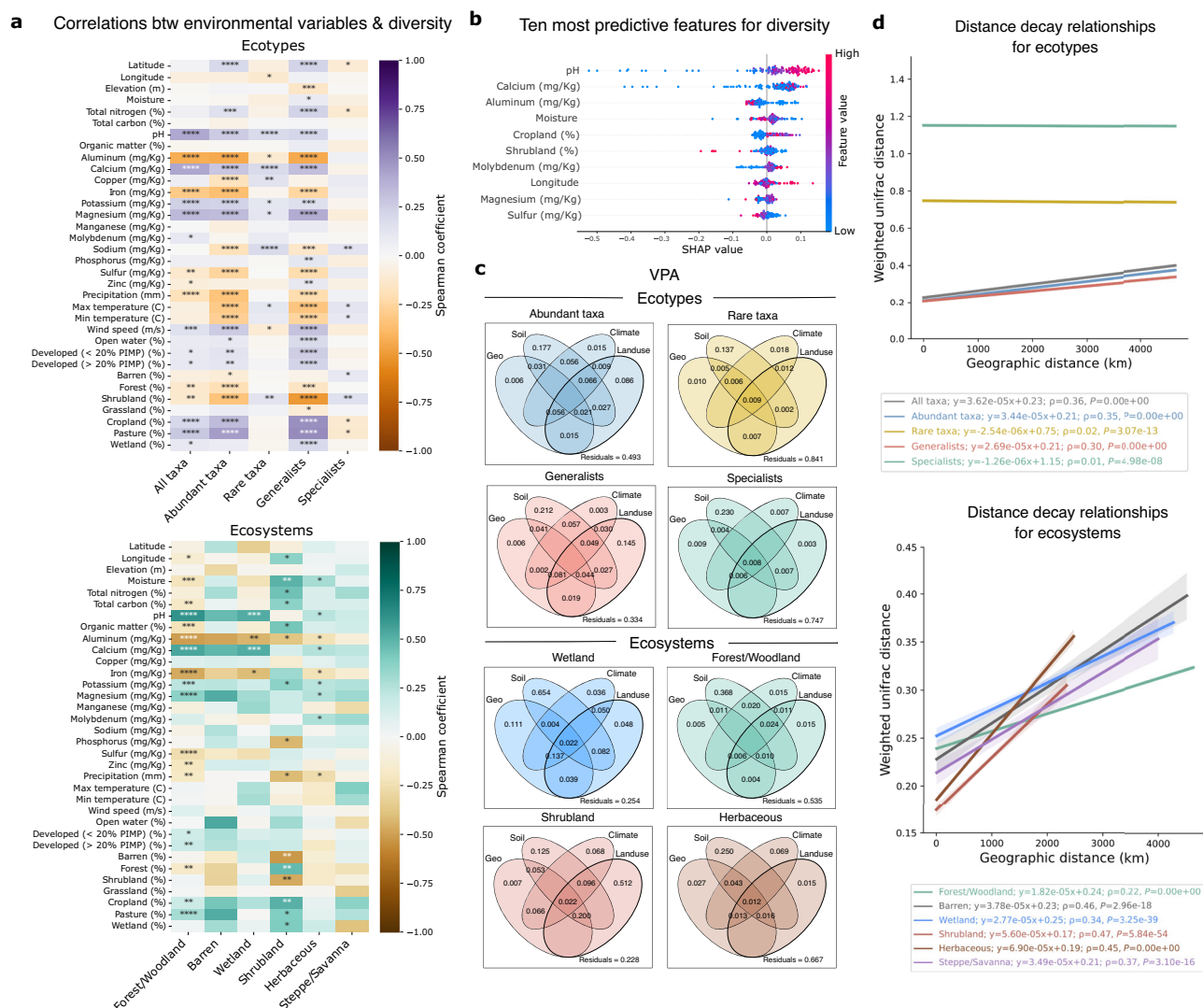


Fig. 2 | Soil bacterial α -diversity is shaped by abiotic environmental factors.

a Spearman's rank correlation between environmental variables and the α -diversity of all taxa, each ecotype, and within each ecosystem. PIMP: percentage imperviousness. Positive and negative correlation is indicated by purple and orange, in the upper heatmap, and green and yellow, respectively, in the bottom heatmap. ****, ***, **, *, and ns denote adjusted two-sided $P < 0.0001$, 0.001, 0.01, 0.05, and not significant, respectively. **b** The top ten most predictive environmental variables for bacterial α -diversity (SHAP-based; X axis), sorted by descending importance.

c Venn diagram of the variation partitioning analysis (VPA) showing the variation of the α -diversity of each ecotype and within each ecosystem explained by environmental factors. Residuals indicate unexplained variation. **d** Distance decay relationships for all taxa, ecotypes, and ecosystems. Bacterial composition dissimilarity is indicated by weighted UniFrac distance. ρ and P are Spearman's rank correlation coefficient and P -value, respectively, and the colored lines and shaded areas depict the best-fit trend lines and the 95% confidence interval (mean ± 1.96 s.e.m.) of the linear regression, respectively.

speed, and coverage of developed land (both > and <20% impervious cover), cropland, pasture, and wetland in surrounding areas, with pH, magnesium, and calcium showing the strongest correlation (adjusted Spearman $P < 0.05$ for all). In contrast, 7 variables (20.6%), including soil aluminum, iron, sulfur, zinc, precipitation, and proportions of forest and shrubland, were significantly negatively correlated, with aluminum and iron showing the strongest correlation (adjusted $P < 0.05$ for all). Stratifying by ecotypes, a consistent correlation pattern was observed in abundant taxa and generalists, in which most of the variables (> 65%) were significantly correlated with α -diversity and 21 of them overlapped between these two ecotypes. Compared to abundant taxa and generalists, much fewer environmental variables were significantly correlated with α -diversity of rare taxa and specialists (11 and 9, respectively) and their correlation coefficients were relatively small. Of note, soil sodium, maximum annual temperature, and the coverage of shrubland in surrounding areas were found to be significantly correlated with the α -diversity for all four ecotypes, but

their effect appeared to be opposite for abundant taxa and generalists compared to rare taxa and specialists. In addition, longitude was found to be uniquely correlated with the α -diversity of rare taxa, while moisture, phosphorus, zinc, grassland, and wetland were uniquely correlated with the α -diversity of generalists. These results suggest that environmental conditions play a more important role in shaping the α -diversity of abundant taxa and generalists than rare taxa and specialists.

Stratifying by ecosystems (Fig. 2a), the strongest correlation between environmental factors and bacterial α -diversity was observed in the forest/woodland ecosystem (18 significant variables) followed by shrubland (15 significant variables) (adjusted Spearman $P < 0.05$ for all). Only 9 and 4 variables were significantly correlated with the diversity in the herbaceous and wetland ecosystems, respectively. The barren and steppe/savanna ecosystems had no significant environmental variables identified, likely due to the weak statistical power caused by their relatively small sample size (Supplementary Fig. 2).

There were seven environmental variables significantly correlated with α -diversity in at least three ecosystems (adjusted $P < 0.05$ for all). Among them, soil pH, calcium, and potassium had a positive correlation consistently across ecosystems, and iron, aluminum, and precipitation had a consistent negative correlation, while moisture showed a mixed effect. Several environmental variables were uniquely significantly correlated with the α -diversity of a particular ecosystem, such as total nitrogen, phosphorus, and the coverage of barren, shrubland, and wetland in surrounding areas for the shrubland ecosystem; sulfur, zinc, and developed area coverage for the forest/woodland ecosystem; and molybdenum for the herbaceous ecosystem. Overall, these results suggest that soil bacterial α -diversity is influenced by environmental factors with combined effects from several universal drivers (e.g., pH, aluminum, calcium, and shrubland coverage in surrounding areas) and a number of drivers specifically acting on a particular ecotype or ecosystem. These findings, however, necessitate further validation using samples with even sample sizes, spatial distributions, and scales for each ecosystem type.

Given the strong relationships between environmental variables and soil bacterial α -diversity, we hypothesize that α -diversity is predictable using environmental variables at a nationwide scale. To test this hypothesis, we developed a machine learning (ML) model to predict α -diversity with the 34 environmental variables. We compared different ML algorithms with random sampling of various hyperparameters and identified gradient boosting regressor as the best algorithm (see Methods). We then conducted an exhaustive search on the regressor over all parameter values to fine-tune the hyperparameters. The most performant model (Supplementary Data 6) achieved an R^2 of 0.46 and a mean squared error (MSE) of 0.15 (Supplementary Fig. 6a). To interpret the best model prediction, we utilized SHAP²⁶ to assess the importance of each feature to the prediction. The top three most influential environmental variables were soil pH, calcium, and aluminum (Fig. 2b), which is consistent with their strong correlations with diversity observed in most ecotypes and ecosystems (Fig. 2a), suggesting universal role in predicting bacterial diversity. These findings underscore the predictive power of abiotic environmental factors for predicting soil bacterial diversity at a large spatial scale.

Since many variables significantly correlated with α -diversity were soil property variables, we hypothesized that soil properties were the most influential factor for soil bacterial diversity. To test this hypothesis, we conducted variation partitioning analysis (VPA) to quantify the contributions of geolocation, soil, climate, and land use variables to the α -diversity of all taxa as well as of each ecotype and within each ecosystem. For all taxa, soil properties exhibited the largest contribution (individually explaining 22.0% of the variation of the α -diversity), followed by surrounding land use (2.4%; Supplementary Fig. 6b). The individual contribution of climate and geolocation was minimal ($< 1\%$). A similar dominant contribution of soil properties was seen in abundant taxa, rare taxa, generalists, and specialists, with 17.7%, 13.7%, 21.2%, and 23.0% of the variation explained, respectively (Fig. 2c). The only other environmental factor individually explaining $> 10\%$ of the variation of the α -diversity was land use for generalists (14.5%; Fig. 2c). Much larger variation of the α -diversity was explained by environmental variables in abundant taxa (50.7%) and generalists (66.6%) than rare taxa (15.9%) and specialists (25.3%), consistent with the Spearman's rank correlation results. For ecosystems (Fig. 2c), VPA also identified soil properties as the most important factor for the α -diversity in wetland (65.4%), forest/woodland (36.9%), and herbaceous (25.0%) ecosystems. For the wetland ecosystem, in addition to soil properties, geolocation also exhibited relatively high importance (11.1%). For the shrubland ecosystem, the importance of land use was most notable (51.2%) followed by soil properties (12.5%). Consistent with the Spearman's rank correlation results (Fig. 2a), the barren and steppe/savanna ecosystems did not have any variation of the α -diversity explained by environmental variables. The unexplained

variation may be a consequence of unmeasured factors (e.g., organic matter quality) that are vital to shaping the structure of bacterial communities. Overall, soil properties appear to be the most influential environmental factor contributing to soil bacterial α -diversity across the US regardless of ecotypes and ecosystems, and land use patterns in surrounding areas tend to play a vital role to certain ecotypes (i.e., generalists) and ecosystems (i.e., shrubland).

While the influence of geolocation on soil bacterial α -diversity was not strong, it may be important to β -diversity. Indeed, using weighted UniFrac distance as the representative β -diversity index, we observed a distance-decay relationship in all taxa, abundant taxa, and generalists, in which β -diversity was significantly positively correlated with geographic distance (Spearman $\rho > 0.3$, $P < 0.05$, and regression slope > 0 for all; Fig. 2d). In comparison, a distance-decay relationship was not observed in rare taxa and specialists ($\rho \leq 0.02$ and regression slope < 0 for both), likely due to their high β -diversity at local scales. Notably, all ecosystems displayed a distance-decay relationship (Spearman $\rho > 0.2$, $P < 0.05$, regression slope > 0 for all; Fig. 2d). The strongest relationship was observed in the shrubland ecosystems ($\rho = 0.47$) followed by barren ($\rho = 0.46$), herbaceous ($\rho = 0.45$) steppe/savanna ($\rho = 0.37$), and wetland ($\rho = 0.34$) ecosystems. The distance-decay relationship was least strong in the forest/woodland ecosystem ($\rho = 0.22$), indicating frequent dispersal of bacteria. This may be attributed to the high presence of wildlife in the forest acting as bacterial dispersal vehicles. These results show that the dissimilarity of bacterial community composition in soils tends to increase with geographic distance for all ecosystems and ecotypes except for rare taxa and specialists at a nationwide scale, highlighting a vital role of local environmental selection and/or dispersal limitation in shaping soil bacterial β -diversity.

Bacterial co-occurrence network

To infer potential soil bacterial interactions and to understand the importance of taxa to the connectivity and structure of bacterial communities, we constructed a co-occurrence network for all OTUs (Fig. 3a) and measured the node degree, betweenness, and closeness centrality for each OTU. Degree centrality is defined as the number of edges each node (i.e., OTU in this case) has in a network, and can measure the level at which an OTU co-occurs with other OTUs²⁷. Betweenness centrality measures how often a node lies on the paths between other nodes and can be used to identify OTUs that interact most frequently with other members of the community network^{27,28}. 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^{28,29}. The density of the network constructed using all OTUs was 0.017. The majority of the edge weights fell between -0.1 and 0.1 (Supplementary Fig. 4), with a substantially larger proportion of positive edge weights (87.6%) observed than the negative ones (12.4%) (Fig. 3a). Chlorobi, Fibrobacteres, WS3, BRC1, and Thermi were the top five phyla with the highest mean node degree centrality (Fig. 3b). BHI80-139, GAL15, Thermi, Gemmatimonadetes, and Bacteroidetes were the top five phyla with the highest mean node betweenness centrality (Fig. 3b). BHI80-139, WS3, Fibrobacteres, Chlorobi, and Nitrospira were the top five phyla with the highest mean node closeness centrality (Fig. 3b). Members of a network with both high degree and betweenness centrality are typically the most connected taxa within the community and are considered “hubs”, which may have strong ecological relevance to the community²⁹. Thus, given the high degree and betweenness centrality observed in Thermi (a.k.a., *Deinococcus-Thermus*), this phylum was identified as a hub in soil bacterial communities across the US.

To understand the importance of different ecotypes to the community network, we analyzed the network stratified by ecotypes (Supplementary Fig. 8a-b). Results showed that rare taxa, generalists, and specialists had a significantly higher node degree centrality than

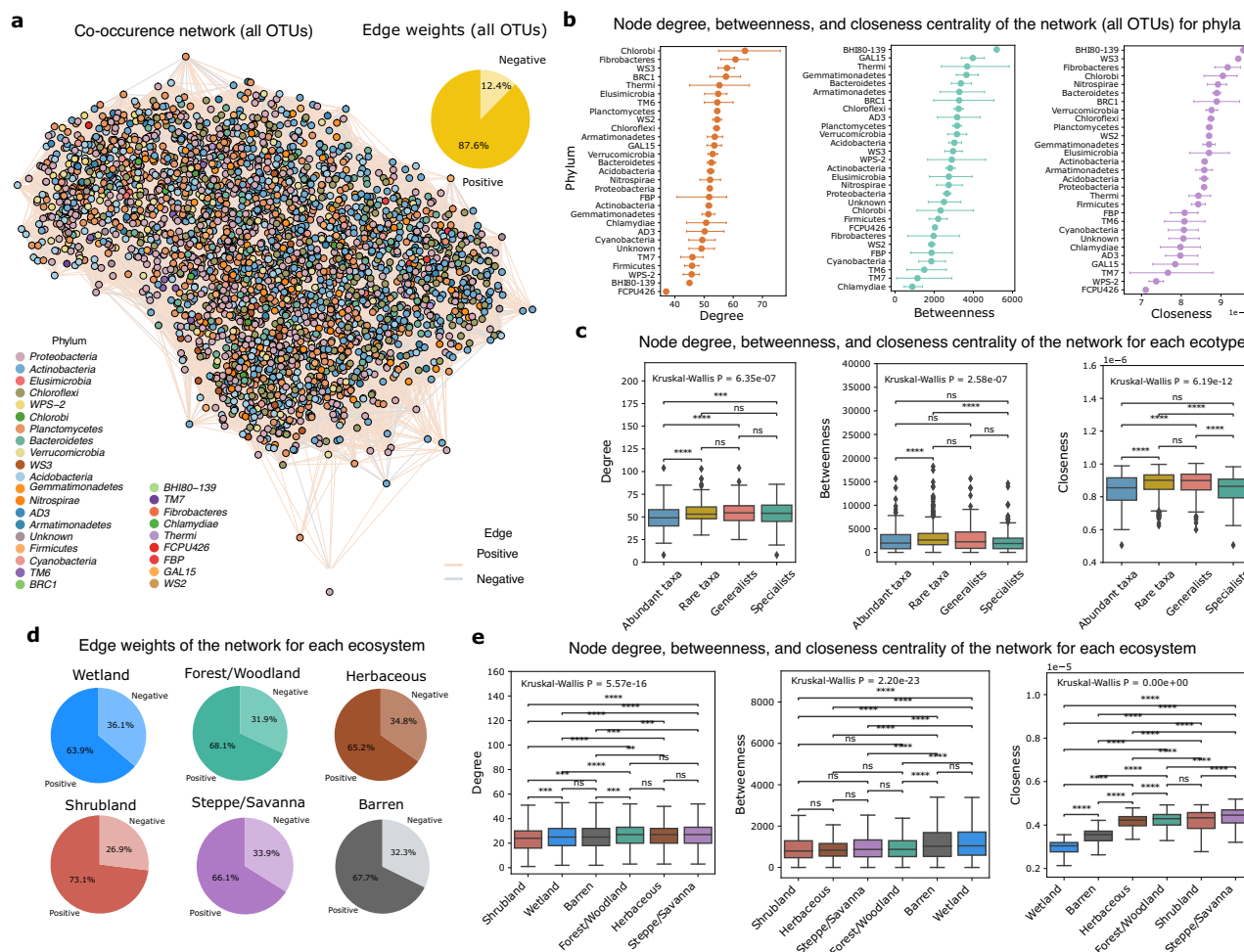


Fig. 3 | Soil bacterial co-occurrence network properties vary by ecotypes and terrestrial ecosystems. **a** Co-occurrence network constructed using all OTUs and proportion of positive and negative edge weights. Nodes are color-coded by phyla. Positive and negative edges are shown in orange and gray, respectively. **b** Top ten phyla with the highest node degree, closeness, and betweenness centrality of the network shown in (a), sorted in descending order. Error bar indicates 95% confidence interval (mean \pm 1.96 s.e.m.). **c** Node degree, betweenness, and closeness centrality of the network compared among ecotypes. The networks color-coded by ecotypes can be found in Supplementary Fig. 8a–b. $N = 201, 870, 156$ and 159 for abundant taxa, rare taxa, generalists, and specialists, respectively. **d** Proportion of

positive and negative edge weights of the network for each ecosystem. The network for each ecosystem is shown in Supplementary Fig. 9a–f. $N = 1189, 1161, 1067, 1058, 1043$, and 914 for wetland, barren, forest/woodland, steppe/savanna, herbaceous, and shrubland ecosystems, respectively. **e** Node degree, betweenness, and closeness centrality of the network compared among ecosystems, sorted in ascending order. Box plots show the interquartile range (IQR), with the line representing the median and whiskers extending to 1.5 times the IQR, Kruskal–Wallis $P < 0.05$ indicates a significant difference among groups, and ****, ***, **, *, and ns denote adjusted two-sided Mann–Whitney $U P < 0.0001, 0.001, 0.01, 0.05$, and not significant, respectively, in (c) and (e).

abundant taxa (median = 53.0, 54.5, 54.0, and 49.0, respectively; KW $P = 6.25e-07$; adjusted two-sided Mann–Whitney [MW] $U P < 0.05$ for all pairwise comparisons; Fig. 3c). Rare taxa also had a significantly higher node betweenness centrality than abundant taxa and specialists (median = 2626, 1970, and 1886, respectively; KW $P = 2.58e-07$; adjusted MW $U P < 0.05$ for all pairwise comparisons; Fig. 3c). In addition, rare taxa and generalists had a significantly higher node closeness centrality than abundant taxa and specialists (median = 9.01e-07, 8.99e-07, 8.55e-07, and 8.64e-07, respectively; KW $P = 6.19e-07$; adjusted two-sided MW $U P < 0.05$ for all pairwise comparisons; Fig. 3c). These results suggest that rare taxa play a more essential ecological role within the communities compared to abundant taxa, while the ecological relevance of generalists and specialists is not substantially different.

To understand how bacterial interactions may differ among ecosystems, we constructed a co-occurrence network individually for each ecosystem (Supplementary Fig. 9a–f). The networks in the wetland and barren ecosystems had a much lower density than other ecosystems (Supplementary Fig. 10). Different ecosystems showed a significant

difference in the proportions of positive and negative edge weights (Fisher's exact $P = 0.0005$). Among them, shrubland and wetland had the highest and lowest proportion of positive edge weights, respectively (63.9% and 73.1%, respectively; Fig. 3d). The node degree, betweenness, and closeness centrality of the networks also significantly differed among ecosystems (KW $P = 5.57e-16, 2.20e-23$, and $<1.00e-30$, respectively). Of note, the network in the shrubland ecosystem had the lowest node degree centrality (median = 24.0), significantly lower than all other ecosystems (median = 25.0 for barren and wetland and 27.0 for forest/woodland, herbaceous, and steppe/savanna; KW $P = 5.57e-16$; adjusted two-sided MW $U P < 0.05$ for all pairwise comparisons; Fig. 3e). Shrubland also had the lowest node betweenness centrality (median = 794.3), significantly lower than barren (1027.0) and wetland (1044.0) ecosystems (KW $P = 2.20e-23$; adjusted two-sided MW $U P = 1.79e-09$ and $2.37e-13$, respectively; Fig. 3e). Network node closeness centrality significantly varied among ecosystems (KW $P < 1.00e-30$; adjusted two-sided MW $U P < 0.05$ for all pairwise comparisons except between forest/woodland and shrubland), with wetland and steppe/savanna having the highest ($4.46e-06$)

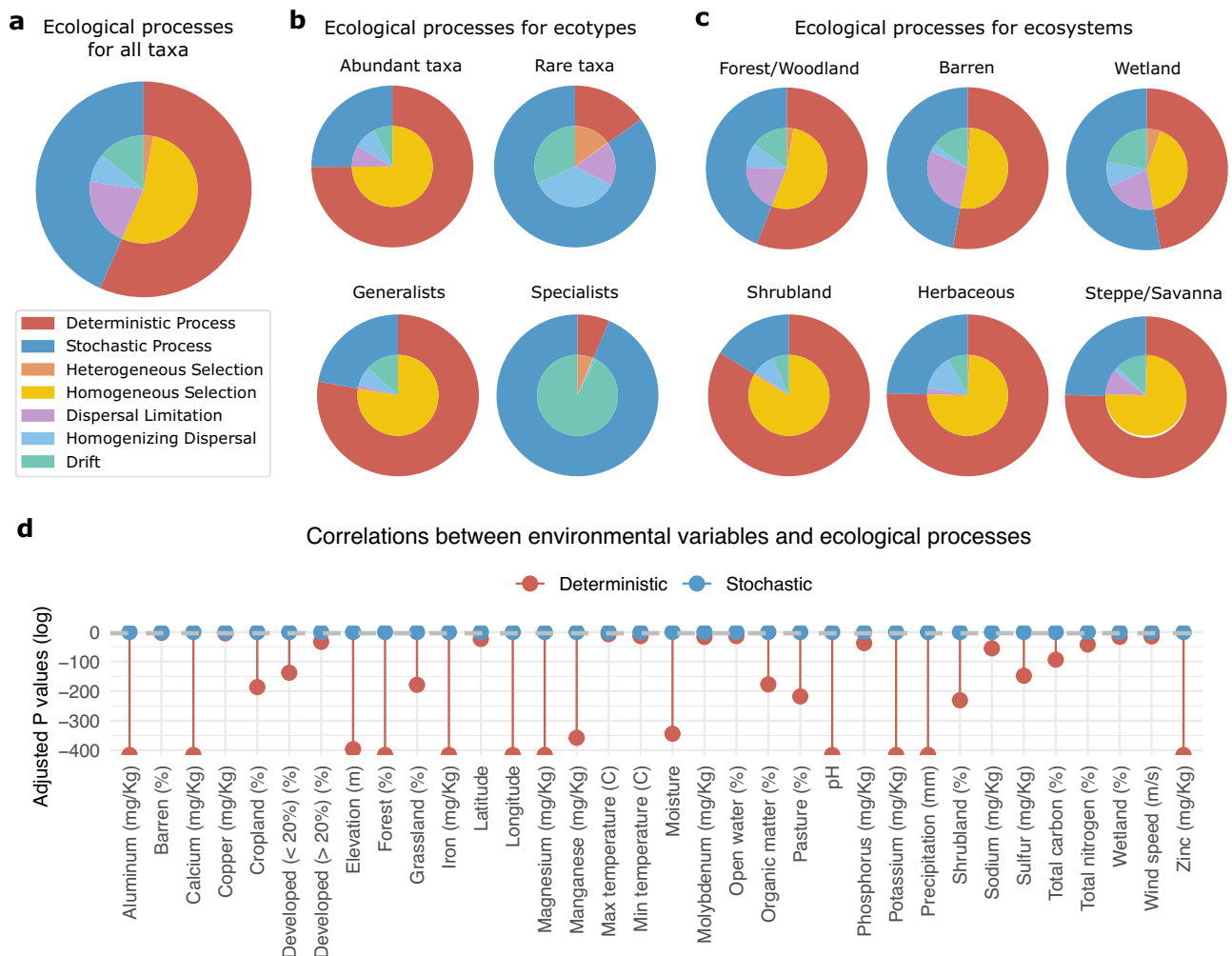


Fig. 4 | Deterministic and stochastic processes play different roles in the bacterial community assembly across ecotypes and terrestrial ecosystems.

a–c Quantified importance of ecological processes to the community assembly for (a) all taxa, (b) ecotypes, and (c) ecosystems. Deterministic processes include

heterogeneous selection and homogeneous selection. Stochastic processes include dispersal limitation, homogenizing dispersal, and drift. **d** Modified Mantel correlation between environmental variables and ecological processes. The grey dash line indicates adjusted two-sided $P < 0.05$ (log).

and lowest (3.04×10^{-6}) median, respectively (Fig. 3e). Overall, these results suggest that soil bacteria in different ecosystems undergo different levels of interactions, and bacteria in the shrubland ecosystem tend to be less connected compared to other ecosystems.

Ecological processes governing bacterial community assembly

As the role of the abiotic and biotic factors on soil bacterial diversity can be manifested with the deterministic processes governing community assembly, we further employed a two-step framework²⁵ based on β -nearest taxon index (β NTI) and a modified Raup–Crick (RC) metric (see Methods) to quantify the importance of deterministic processes (including homogeneous selection and heterogeneous selection) and stochastic processes (including dispersal limitation, homogenizing dispersal, and drift). Results showed that the community assembly for all taxa was jointly contributed by deterministic and stochastic processes, with deterministic processes being slightly more important (~60%; Fig. 4a). Homogeneous selection was identified as the primary deterministic process, while dispersal limitation, homogenizing dispersal, and drift contributed approximately equally to stochastic processes (Fig. 4a). Consistent with the results that environmental variables showed a substantially stronger association with the diversity of abundant taxa and generalists compared to rare taxa and specialists (Fig. 2a and c), more than 75% of the assembly for

abundant taxa and generalists was found to be influenced by deterministic processes, while more than 85% of that for rare taxa and specialists were contributed by stochastic processes (Fig. 4b). Specifically, homogeneous selection was identified as the main deterministic process for abundant taxa, while heterogeneous selection was the main deterministic process for rare taxa. Also, like for all taxa (Fig. 4a), the importance of stochastic processes for abundant taxa had a nearly equal contribution from dispersal limitation, homogenizing dispersal, and drift. For rare taxa, however, homogenizing dispersal and drift showed slightly higher importance than dispersal limitation, consistent with the distance decay relationship observed in this ecotype (Fig. 2d). Both environmental heterogeneity and ecological drift can produce high β -diversity among communities^{15,30,31}. Since the effects of homogenizing dispersal and dispersal limitation on β -diversity may cancel out given a similar contribution, the high local β -diversity observed for rare taxa (Fig. 2d) may be mainly caused by the accumulative effects from heterogeneous selection and drift. Generalists had a similar pattern of the importance of deterministic processes as abundant taxa, in which homogeneous selection was the main driver (Fig. 4b). In contrast, heterogeneous selection was identified as the dominant deterministic process for specialists (Fig. 4b). In addition, both homogenizing dispersal and drift were identified as the dominant stochastic processes for generalists, while

for specialists, nearly the whole importance of stochastic processes arised from drift. Overall, these results suggest that there is an evident joint influence of both deterministic and stochastic processes governing the bacterial community assembly across the US, with deterministic processes, especially homogenous selection, mainly acting on abundant taxa and generalists, and with stochastic processes mainly acting on rare taxa and specialists.

Like ecotypes, the ecological mechanisms underlying the bacterial community assembly across ecosystems were also different (Fig. 4c). For the forest/woodland, barren, and wetland ecosystems, the importance of deterministic and stochastic processes was about equal, with the former and the latter being slightly more important for the forest/woodland and barren ecosystems and the wetland ecosystem, respectively. In comparison, in shrubland, herbaceous, and steppe/savanna ecosystems, more than 75% of the bacterial community assembly was found to be influenced by deterministic processes. Specifically, homogenous selection was identified as the dominant deterministic process across all ecosystems. For forest/woodland and wetland ecosystems, dispersal limitation, homogenizing dispersal, and drift showed a nearly equal contribution. For barren and steppe/savanna ecosystems, dispersal limitation and drift were found to be more important than homogenizing dispersal. For the shrubland and herbaceous ecosystems, homogenizing dispersal and drift exhibited much higher importance than dispersal limitation. These results highlight the different roles that deterministic and stochastic processes play in governing bacterial community assembly in different ecosystems.

To determine which environmental factors may trigger the deterministic processes driving community assembly, we performed a modified Mantel test that handles asymmetric matrices (see Methods) for β NTI and environmental variables. Results show that most of the environmental variables were significantly correlated with deterministic processes (adjusted $P < 0.05$ for all; Fig. 4d). The most significant ones included seven soil property variables (aluminum, calcium, iron, magnesium, pH, potassium, and zinc), precipitation, proportion of forest in surrounding area, and longitude, 90% of which were found to be significantly correlated with bacterial α -diversity (Fig. 2a for all taxa). As expected, none of the environmental variables were significantly correlated with stochastic processes (adjusted $P > 0.05$ for all; Fig. 4d). These results suggest that environmental factors, chiefly soil properties, trigger deterministic processes acting on bacterial community assembly, which ultimately impact bacterial diversity that we observe in the soil environment. Of note, the importance of deterministic processes was found to be highest for bacterial communities in the shrubland ecosystem compared to other ecosystems (Fig. 4c), indicating strong local environmental selection, which could partially explain the strongest distance-decay relation observed in this ecosystem (Fig. 2d). The selection likely mainly comes from surrounding land use, as evidenced by $> 50\%$ variation of the α -diversity explained by land use variables in this ecosystem (Fig. 2c).

Discussion

Overall, soil bacteria in terrestrial ecosystems across the US were highly diverse within and between locations. Consistent with previous findings^{32–34}, Actinobacteria and Proteobacteria were predominant in the soil environment. Both phyla play a key role in global carbon cycling by decomposing soil organic matter, enhancing plant productivity, and are recognized for producing bioactive compounds vital for human and animal health^{34,35}. Notably, *Deinococcus-Thermus*, which is known as a phylum of extremophiles that exhibit strong resistance to environmental extremes³⁶, was identified as a hub in the network, indicating a critical ecological role to the bacterial communities in the soil environment. In addition, 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 (Fig. 3c). This finding is supported by the documented ecological roles that rare biosphere plays in ecosystems, including serving as a persistent microbial seed bank that contrasts the impact of local microbial extinction and immigration and providing a broad reservoir of ecological function and resiliency (redundancy and flexibility)³⁷.

In this study, deterministic processes were found to contribute to ~60% of community assembly of soil bacteria across the US. The impact of deterministic processes primarily arises from homogeneous selection, which drives more phylogenetically similar community structures between locations¹⁶. The remaining ~40% of assembly processes were contributed by stochastic processes, with dispersal limitation playing a more important role than homogenizing dispersal and drift. Consistent with this result, we observed evidence of a distance-decay relationship in soil bacterial communities at a nationwide scale, which is a classical biogeographic pattern commonly observed in macroorganisms³⁸. The joint effect of deterministic and stochastic processes acting on microbial community assembly has been widely recognized in the field of microbial ecology^{22,23}, but their relative importance varies by environment. For example, in freshwater lakes, deterministic processes appear to be much more important than stochastic processes⁷, while platisphere bacterial communities were reported to be dominantly driven by stochastic processes³⁹. Among the different terrestrial ecosystems examined in this study, deterministic processes were found to play a more important role in bacterial community assembly in all ecosystems, especially in shrubland, herbaceous, and steppe/savanna ecosystems, except for wetland. The more important role of stochastic processes in wetlands may be attributed to their distinct semi-aquatic features that increase the chance for dispersal. Indeed, we found a relatively weaker distance-decay relationship for bacteria communities in this ecosystem (Fig. 2d).

Striking differences in ecological mechanisms of community assembly were observed among different ecotypes. For the assembly of abundant taxa and generalists, the importance of deterministic processes was more than 75%, while for specialists and rare taxa, the importance of stochastic processes was more than 80% and few environmental variables were found to be significantly correlated with their diversity. Our results are consistent with the findings by Székely and Langenheder⁴⁰ but contradict those by Pandit et al.⁴¹ and our previous study regarding the assembly mechanisms for generalists and specialists in freshwater lake⁷. This controversy may be related to the level of correlation between habitat specialization and the abundance of taxa. In this study and Székely and Langenheder's (2014) study⁴⁰, generalists were common and abundant, whereas specialists tended to be in low abundance (Supplementary Fig. 4a). However, in Pandit et al.⁴¹ and our previous study⁷, generalists and specialists were distributed along the entire range of abundances. The ecological mechanism for abundant taxa identified in this study was consistent with our previous study on the bacterial communities in freshwater lake⁵, but for rare taxa, the results were inconsistent, in which stochastic processes appear to play a limited role in freshwater lake⁵. This suggests that community assembly mechanisms are more generalizable for abundant taxa, while for rare taxa, the mechanisms may be more dependent on the environmental settings. Overall, our results suggest that the biogeographical patterns of specialists and rare taxa in the soil environment are largely unpredictable. In contrast, abundant taxa and generalists appear to be less resistant to environmental disturbances and could be prone to biodiversity loss.

It is known that soil bacterial diversity is strongly influenced by environmental factors^{17–19}. We found that across the US, soil properties are the most influential factor followed by surrounding land use patterns, which jointly trigger the deterministic processes in bacterial community assembly. While precipitation and windspeed were significantly correlated with bacterial diversity, the importance of

climatic factors was overall small, contrasting with findings from other studies^{42–44}. This may be because the influence of climatic factors is dependent on the spatial and/or temporal scales. Of note, a few global drivers of soil bacteria diversity regardless of ecosystems and ecotypes were identified, including soil pH, calcium, and aluminum, detected by multiple statistical analyses, including machine learning models. The overriding importance of soil pH controlling soil bacterial diversity and community composition has been reported across a variety of spatial scales, including continental scales^{20,45,46}, land-use types^{47,48}, small, local scales^{17,37}, and across an elevational gradient⁴⁹. All these findings, including ours, suggest that pH is a universal predictor of soil bacterial diversity. The global effect of calcium and aluminum on soil bacterial diversity, however, is much less documented than pH. Calcium has been found to be responsible for forming micro-aggregates which aids in bacterial activity leading to increased diversity in soils based on availability^{50,51}. Aluminum, an important element aiding in plant growth, when present in excess, could limit nutrient uptake for soil bacteria and thus limits biodiversity due to acidification⁵². Indeed, the mean aluminum concentration of soil samples included in this study is high (40.64 mg/kg), seven times higher than the aluminum level (5 mg/kg) generally considered safe for plants in acidic soils⁵³. These interpretation of the effects of calcium and aluminum on soil bacterial diversity is consistent with the strong positive and negative correlations with α -diversity observed in this study, respectively (Fig. 2a). Our study suggests that soil calcium and aluminum concentrations, like pH, may also be used as universal predictors for soil bacterial diversity at a large spatial scale.

Notably, results in this study imply a high vulnerability of bacterial communities to environmental disturbance in the shrubland ecosystem. Specifically, bacterial communities in this ecosystem had the lowest diversity and least connected community network, and were undergoing strong environmental selection, which appear to be mainly triggered by surrounding land use. As a result, a number of species, including *Nocardioideus dilutus*, *Pseudonocardia halophobica*, *Virgisporangium ochraceum*, and *Geodermatophilus obscurus*, were found to be uniquely abundant in this ecosystem. These species may have adapted to the dry and nutrient-poor conditions because soils in scrublands are often nutrient-poor, sandy, or rocky, with low fertility and poor water retention⁵⁴. Shrubland provides important ecological, environmental, and socio-economic services, including supporting a diverse range of plant and animal species, serving as a significant carbon sink, providing ecological balance and fire adaptation, regulating water cycles, and providing resources for local communities^{55,56}. Our results suggest that environmental changes (e.g., intensive anthropogenic land use in surrounding areas) may lead to substantial diversity loss and ecological degradation in the shrubland ecosystem. Thus, it is critical to prioritize the protection of shrublands in ecosystem management amid environmental disturbance.

This study characterized the biogeographic patterns of bacterial communities in six major terrestrial ecosystems, including forest/woodland, shrubland, wetland, herbaceous, steppe/savanna, and barren, across the US. It also revealed key environmental factors and the importance of ecological processes governing community assembly, which consequently shape the heterogeneity of bacterial diversity and composition across different ecosystems and ecotypes. Future studies using amplicon sequence variants (ASVs) or metagenomic sequencing are needed to enhance the mechanistic understanding of microbial biogeography at a higher resolution. Of note, we propose that in addition to soil pH, calcium and aluminum may also be used as universal predictors of soil bacterial diversity. In addition, given the strong ecological relevance of rare taxa to the community and the vulnerability of shrubland ecosystems to environmental changes, we emphasize the importance of conserving the rare biosphere and shrubland ecosystems in the face of environmental disturbances.

These implications provide valuable insights that can improve the management and sustainability of ecosystem services by informing strategies for conservation and resource allocation.

Methods

Soil samples, 16S rRNA gene amplicon sequencing, and environmental data

A total of 622 soil samples previously collected from natural environments with minimum human disturbance across the contiguous US in 2018 were used in this study. The methods for sample collection were detailed in Liao et al.⁵⁷. In brief, samples were collected from topsoil (0–20 cm) following a standard protocol. To ensure an even distribution of sampling locations, the contiguous US was divided into 40 equal-sized sampling grids. Within each sampling grid, five sampling areas were identified, and within each sample area, five sampling sites were identified. At each site, three subsamples were collected and pooled. Based on the classification of the standardized terrestrial ecosystems established by the US Geological Survey (see the map in Fig. 8 in Sayre et al.⁵⁸), these samples cover six major terrestrial ecosystems, including forest/woodland (298 samples), herbaceous (134), wetland (54), shrubland (44), steppe/savanna (31), barren (26), and unknown ecosystems (29) (Supplementary Fig. 2). Forest/woodland ecosystems are dominated by trees forming a continuous canopy, typically covering more than 60% of the area. Herbaceous ecosystems are areas where non-woody plants, such as grasses, sedges, and forbs, are the primary vegetation, with minimal to no tree or shrub presence. Wetland ecosystems are characterized by saturated soils or standing water for significant periods, supporting hydrophytic vegetation. Shrubland ecosystems are areas where shrubs (i.e., woody plants shorter than trees with multiple stems) are the dominant vegetation, typically covering 25–60% of the area. Steppe/savanna ecosystems are open landscapes characterized by a mix of grasses and scattered trees or shrubs, with tree canopy cover ranging from 10–30%. Barren ecosystems are areas with minimal vegetation cover, often due to harsh environmental conditions like poor soils, extreme temperatures, or recent disturbances. Sites classified as unknown ecosystems are areas with a small pixel count ($\leq 20,000$) in the USGS classification system. Soil samples were extracted for total DNA using QIAGEN DNeasy PowerSoil Pro Kits and sequenced for the V4 region of the 16S rRNA gene using a MiSeq 2 × 250 bp paired-end read run. The methods for DNA extraction and sequencing were detailed in Liao et al.⁵⁹. The number of raw sequencing reads for all 622 samples ranged from 8599 to 59,425.

Environmental data used in this study were previously reported in Liao et al.⁵⁷. This dataset includes 3 geolocation (latitude, longitude, and elevation), 17 soil properties (moisture, total nitrogen, total carbon, pH, organic matter, aluminum, calcium, copper, iron, potassium, magnesium, manganese, molybdenum, sodium, phosphorus, sulfur, and zinc), 4 climatic (precipitation, wind speed, maximum and minimum temperatures), and 10 surrounding land use (open water, barren, forest, shrubland, grassland, cropland, pasture, wetland, and developed open space categorized as $> 20\%$ and $< 20\%$ impervious cover) variables.

Bacterial composition, diversity, and phylogenetic tree

Raw reads were processed using QIIME2 following the procedures described in Liao et al.⁵⁹ with minor modifications. In brief, reads were denoised using DADA2 and standardized by rarefaction to 5000 reads based on the α -rarefaction curve followed by proportioning. Four samples that had < 5000 reads were excluded from downstream analyses. Sequences were clustered de novo into operational taxonomic units (OTUs) using q2-vsearch at a similarity of 0.97 and taxonomic classification of OTUs was determined using classify-sklearn. While OTUs are often less precise than ASVs, OTUs were chosen over ASVs because of legacy data comparison, broader ecological groupings

(e.g., ecotypes), tolerance to sequencing errors, less sensitivity to hypervariable regions, and computational efficiency⁶⁰. To enhance the quality and interpretability of bacterial community analyses⁶¹, we took a conservative approach to remove OTUs with low frequency (present in <1% of samples) and low abundance (having total reads <10) that often result from sequencing errors, contamination, or random sampling artifacts, by referring to the thresholds used in Wilhelm et al.⁶². OTUs that are non-bacterial were also excluded from downstream analyses. After data preprocessing and filtering, a total of 3158 OTUs were produced (Supplementary Data 7). Four α -diversity metrics, including richness, Shannon-Wiener diversity index, Simpson's evenness, and Faith's phylogenetic diversity (PD), and four β -diversity metrics, including Bray-Curtis, Jaccard, unweighted UniFrac, and weighted UniFrac distances, were calculated using QIIME2 q2-diversity plugin. Relationships among α -diversity metrics and β -diversity metrics were assessed using Spearman's rank correlation tests and Mantel tests, respectively.

Kruskal-Wallis (KW) tests were employed to identify significant differences in the Shannon-Wiener diversity index, relative abundance of bacterial phyla, class, order, family, genus, species, and OTUs as well as environmental variables among ecosystems followed by a Benjamini-Hochberg (BH) false discovery rate (FDR) adjustment to account for multiple testing. Variables with an FDR-adjusted P value < 0.05 are considered significant. Multidimensional scaling (MDS) along with a permutational multivariate analysis of variance (PERMANOVA) test was used to compare the differences in overall environmental conditions and bacterial composition based on OTUs among ecosystems. A PERMANOVA P < 0.05 indicates a significant difference among groups. The distribution of the Shannon-Wiener diversity index for ecosystems was visualized using the Basemap Matplotlib Toolkit v.1.2.1 in Python v.3.6.8.

The phylogenetic tree of OTUs was constructed based on the full sequence alignment of the 16S rRNA gene using IQ-TREE with 1000 bootstraps⁶³. The best evolutionary model was determined based on the Bayesian information criterion (BIC) by the ModelFinder implemented in IQ-TREE. The tree, rooted by mid-point and annotated by the relative abundance of each OTU, was visualized using iTOL⁶⁴.

Ecotypes and associations with phyla and ecosystems

Four ecotypes, including abundant taxa, rare taxa, generalists, and specialists, were characterized. Abundant and rare taxa were defined based on the mean relative abundance across all samples. OTUs with a mean relative abundance of > 0.1% were defined as abundant taxa. This cutoff lies within the outlier area of the mean relative abundance distribution (Supplementary Fig. 11) and has been commonly used to define abundant taxa in other studies^{5,65}. With this cutoff, 201 out of 3158 OTUs (6.4%) were classified as abundant taxa. OTUs with a mean relative abundance of < 0.002% were defined as rare taxa. This cutoff was selected as it is approximately the median of the mean relative abundance for all OTUs (Supplementary Fig. 11), consistent with the approach to select cutoff for rare taxa in other studies^{5,6}. With this cutoff, 870 out of 3158 OTUs (27.5%) were classified as rare taxa. To confirm the rationality of the chosen cutoffs for defining abundant and rare taxa, multivariate cutoff level analysis (MultiCoLA), a strategy to systematically assess the influence of rarity definition on large community datasets, was conducted using MultiCoLA.1.4⁶⁶. Results showed that little variation in the data structure was observed for OTUs and taxonomic ranks up to a removal of 10% of the abundant taxa (Supplementary Fig. 12a) and 30% of the rare taxa (Supplementary Fig. 12b), respectively. These results suggest that the cutoffs chosen for defining abundant taxa (6.4% of total OTUs) and rare taxa (27.5% of total OTUs) were appropriate and not affected by arbitrariness.

Since our conservative approach to removing OTUs with low frequency and abundance may discard many rare species, this could bias the results for rare taxa in the study. To identify this potential bias, we re-analyzed the data by only removing singletons (i.e., OTUs

present in one sample with one read sequenced), which yielded 7869 OTUs. Using the same approach, we chose 0.0007% as the cutoff to classify rare taxa as it is approximately the median of the mean relative abundance for all these OTUs (Supplementary Fig. 13a). A total of 3709 OTUs were classified as rare taxa. We further calculated their Shannon-Wiener diversity 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; Supplementary Fig. 13b). Thus, we concluded that the classification of rare taxa was not biased by the method that we used to remove OTUs with low frequency and abundance.

Habitat generalists and specialists were defined based on niche breadth, which identifies different levels of specialization of species^{6,67}. Niche breadth was calculated using the Eq. (1) below for each OTU:

$$B_j = \frac{1}{\sum_{i=1}^N P_{ij}^2} \quad (1)$$

where B_j indicates niche breadth and P_{ij} is the relative abundance of species j present in a given habitat i . B_j can be greater than or less than the expected degree of niche breadth, indicating a broader or narrower range of habitats than expected^{6,41}. To quantify the degree to which B_j deviates from expectation, a permutation test was performed by shuffling the OTU relative abundance matrix by 1000 times to generate a null distribution of B_j for each OTU. OTUs with a 95% chance of having an observed B -value > expected and < expected were classified as generalists and specialists, respectively (P < 0.05). A total of 156 and 159 generalists and specialists were identified, respectively.

Shannon-Wiener diversity index and weighted UniFrac distances were computed for each ecotype for each sample using the scikit-bio library in Python v.3.6.8. The counts of each ecotype were compared among ecosystems using KW tests. To identify bacterial phyla that were significantly overrepresented and underrepresented within each ecotype, a binomial distribution model was used to compare the frequency of each phylum within each ecotype with the frequency among all OTUs using the Eq. (2) below^{59,68}.

$$i = \frac{n - pN}{\sqrt{p(1-p)N}} \quad (2)$$

where n is the observed count of OTUs belonging to a phylum for a given ecotype, N is the total count of OTUs for a given ecotype, p is the frequency of OTUs belonging to a phylum among all OTUs, and i is the enrichment index, which represents the multiplier of standard deviation in the binomial distribution. Phyla with $i > 2$ and $i < -2$ indicates a significant overrepresentation and underrepresentation, respectively (P < 0.05), for a given ecotype.

The influence of environmental factors on bacterial diversity

An FDR-adjusted Spearman's rank correlation analysis was conducted to assess the associations between bacterial α -diversity represented by the Shannon-Wiener diversity index and each environmental variable for all OTUs as well as for each ecosystem and ecotype. Environmental variables with an FDR-adjusted P < 0.05 were considered significant. Following this, variation partitioning analysis (VPA) was performed to quantify the relative contribution of each environmental variable group (i.e., geolocation, soil property, climate, or land use) to the variation of bacterial α -diversity. VPA was executed using the vegan package v.2.6-4 in R 4.2.2 and the adjusted R^2 , which represents the proportion of the variance for a dependent variable explained by an independent variable group, was visualized as a Venn diagram.

Distance decay relationships were assessed using Spearman's rank correlation analysis and linear regression on geographic distance against bacterial community dissimilarity indicated by weighted UniFrac distance for all OTUs as well as for each ecosystem and ecotype. A

larger positive Spearman correlation coefficient and a larger positive linear regression slope indicate a stronger distance decay relationship. Due to the mass of data points, only the best-fit line of the linear regression model is shown.

Machine learning models for bacterial diversity prediction

We developed an end-to-end model training, validation, and testing framework based on robust machine learning software, including scikit-learn, Light Gradient-Boosting Machine (LightGBM), and XGBoost⁶⁹. To find the best model to predict bacterial diversity, we first pre-selected a set of algorithms and hyperparameters, including random forest regressor, gradient boosting-based regressor, support vector regressor, and k-nearest neighbors-based regressors. For each algorithm, we uniformly sampled 100 parameter settings and ran a stratified 5-fold cross-validated search on each parameter setting. We identified gradient boosting-based regressor as the best algorithm as it had the highest average R^2 score. To fine-tune the hyperparameters, we ran an exhaustive search over all parameter values without sampling. To account for stochasticity introduced by the random splitting of samples into training and testing sets, we repeated this step 10 times. We selected the most performant gradient-boosting regressor with its hyperparameter set that had the highest interquartile mean of the R^2 scores out of the 10 repetitions. It utilized absolute error as the loss function to be optimized, mean squared error as the function to measure the quality of a split, 400 as the number of boosting stages, 150 as the maximum depth of the individual regression estimators, and 0.05 as the learning rate. To validate our model selection, we kept 20% of all samples as a holdout testing set and retrained the model exclusively with the remaining 80% of samples. R^2 and mean squared error (MSE) were reported based on a single evaluation of the holdout data. SHapley Additive exPlanations (SHAP)²⁶ was used to quantify the importance of features.

Co-occurrence network

The co-occurrence network for all OTUs as well as for each ecosystem was constructed using SpiecEasi 1.0.0 based on the relative abundance of OTUs²⁸. The neighborhood selection (MB) method and default settings of other parameters were used. Network density, edge weights, and node degree, betweenness, and closeness centrality were computed. For the co-occurrence network for each ecosystem, to avoid bias potentially caused by different sample sizes among ecosystems, an identical number of samples (i.e. the minimum sample size) was randomly selected from each ecosystem for the network construction. Fisher's exact tests were performed to assess if the frequency of positive and negative edge weights was associated with ecosystems. A Fisher's exact $P < 0.05$ indicates a significant association. KW tests were conducted to assess the difference in node degree, betweenness, and closeness centrality among ecosystems as well as among ecotypes. A KW $P < 0.05$ indicates a significant difference.

Ecological processes governing community assembly

We employed a two-step framework detailed in Stegen et al.²⁵ to infer ecological processes, i) deterministic processes, including homogeneous selection and heterogeneous selection, and ii) stochastic processes, including drift acting alone, dispersal limitation acting in concert with drift, and homogenizing dispersal, acting on phylogenetic composition turnover. In the first step, the beta mean nearest taxon distance (β MNTD) was first computed to quantify turnover in phylogenetic community composition across samples using the R comdistnt package based on the Eq. (3) below:

$$\beta\text{MNTD} = 0.5 \left[\sum_{i_k=1}^{n_k} f_{i_k} \min(\Delta i_k j_m) + \sum_{i_m=1}^{n_m} f_{i_m} \min(\Delta i_m j_k) \right] \quad (3)$$

Where f_{i_k} is the relative abundance of OTU i in community k , n_k is the number of OTUs in community k , and $\min(\Delta i_k j_m)$ is the minimum phylogenetic distance between OTU i in community k and OTU j in community m . A permutation test was further performed to quantify the degree to which β MNTD deviates from expectation, termed the β -nearest taxon index (β NTI). β NTI values > 2 or < -2 indicate a significant phylogenetic turnover greater than or less than expectation, implying heterogeneous and homogeneous selection, respectively, while $-2 \leq \beta\text{NTI} \leq 2$ indicates community assembly is governed by stochastic processes. In the second step, to quantify the importance of each stochastic process, the Bray–Curtis (BC) distance between samples was first calculated. To quantify the degree to which BC deviates from expectation, a permutation test was further performed and the deviation compared between observed BC and expectation was then standardized to vary between -1 and 1 , generating a modified Raup–Crick (RC) metric referred to as RC_{bray} . $\text{RC}_{\text{bray}} > 0.95$, $\text{RC}_{\text{bray}} < -0.95$, and $-0.95 \leq \text{RC}_{\text{bray}} \leq 0.95$ indicate dispersal limitation acting alongside drift, homogenizing dispersal, and drift acting alone, respectively. This framework was applied to quantify the importance of ecological processes for all OTUs as well as for each ecosystem and ecotype.

Correlations between environmental variables and ecological processes were assessed based on a modified Mantel test. In this test, $E = \{E_1, \dots, E_k\}$ represents environmental variables. β NTI between the i -th and j -th sample is denoted as βNTI_{ij} . D is defined as the set of pairs associated with deterministic processes: $D = \{(i, j) \mid |\beta\text{NTI}_{ij}| > 2\}$. S is defined as the set of pairs associated with stochastic processes: $S = \{(i, j) \mid -2 \leq \beta\text{NTI}_{ij} \leq 2\}$. For the k -th environmental variable, let (E_{ki}, E_{kj}) be the observed value pair for $(i, j) \in D$ or S . A Spearman correlation test was performed between βNTI_{ij} and (E_{ki}, E_{kj}) for $(i, j) \in D$ or S , followed by FDR correction.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 16S rRNA sequencing reads have been deposited at NCBI Sequence Read Archive (SRA) under accession number PRJNA749132. Source data for all graphs is available at https://github.com/leaph-lab/USsoil16S_MS70.

Code availability

Code to replicate all analyses is available at https://github.com/leaph-lab/USsoil16S_MS70.

References

- Gupta, A., Gupta, R. & Singh, R. L. Microbes and environment. In *Principles and Applications of Environmental Biotechnology for a Sustainable Future* (ed. Singh, R. L.) 43–84 (Springer Singapore, Singapore, 2017).
- Meyer, K. M. et al. Why do microbes exhibit weak biogeographic patterns? *ISME J.* **12**, 1404–1413 (2018).
- Xu, Z. et al. Geographical and environmental distance differ in shaping biogeographic patterns of microbe diversity and network stability in lakeshore wetlands. *Ecol. Indic.* **158**, 111575 (2024).
- Mo, Y. et al. Biogeographic patterns of abundant and rare bacterioplankton in three subtropical bays resulting from selective and neutral processes. *ISME J.* **12**, 2198–2210 (2018).
- Liao, J. et al. Similar community assembly mechanisms underlie similar biogeography of rare and abundant bacteria in lakes on Yungui Plateau, China. *Limnol. Oceanogr.* **62**, 723–735 (2017).
- Logares, R. et al. Biogeography of bacterial communities exposed to progressive long-term environmental change. *ISME J.* **7**, 937–948 (2013).

7. Liao, J. et al. The importance of neutral and niche processes for bacterial community assembly differs between habitat generalists and specialists. *FEMS Microbiol. Ecol.* **92**, fiw174 (2016).
8. Cohan, F. M. Bacterial species and speciation. *Syst. Biol.* **50**, 513–524 (2001).
9. Philippot, L., Chenu, C., Kappler, A., Rillig, M. C. & Fierer, N. The interplay between microbial communities and soil properties. *Nat. Rev. Microbiol.* **22**, 226–239 (2024).
10. Nemergut, D. R. et al. Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356 (2013).
11. Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
12. Choudoir, M. J., Doroghazi, J. R. & Buckley, D. H. Latitude delineates patterns of biogeography in terrestrial *Streptomyces*. *Environ. Microbiol.* **18**, 4931–4945 (2016).
13. Wu, L. et al. Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol.* **4**, 1183–1195 (2019).
14. Vellend, M. Conceptual synthesis in community ecology. *Q. Rev. Biol.* **85**, 183–206 (2010).
15. Chase, J. M. & Myers, J. A. Disentangling the importance of ecological niches from stochastic processes across scales. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 2351–2363 (2011).
16. Stegen, J. C., Lin, X., Fredrickson, J. K. & Konopka, A. E. Estimating and mapping ecological processes influencing microbial community assembly. *Front Microbiol.* **6**, 370 (2015).
17. Baker, K. L. et al. Environmental and spatial characterisation of bacterial community composition in soil to inform sampling strategies. *Soil Biol. Biochem.* **41**, 2292–2298 (2009).
18. Dequiedt, S. et al. Biogeographical patterns of soil bacterial communities. *Environ. Microbiol. Rep.* **1**, 251–255 (2009).
19. Lindström, E. S. & Langenheder, S. Local and regional factors influencing bacterial community assembly. *Environ. Microbiol. Rep.* **4**, 1–9 (2012).
20. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl Acad. Sci. USA* **103**, 626–631 (2006).
21. Zhou, J. & Ning, D. Stochastic community assembly: does it matter in microbial ecology? *Microbiol. Mol. Biol. Rev.* **81**, e00002–17 (2017).
22. Dong, M. et al. Microbial community assembly in soil aggregates: a dynamic interplay of stochastic and deterministic processes. *Appl. Soil Ecol.* **163**, 103911 (2021).
23. Caruso, T. et al. Stochastic and deterministic processes interact in the assembly of desert microbial communities on a global scale. *ISME J.* **5**, 1406–1413 (2011).
24. Feng, Y. et al. Two key features influencing community assembly processes at regional scale: Initial state and degree of change in environmental conditions. *Mol. Ecol.* **27**, 5238–5251 (2018).
25. Stegen, J. C. et al. Quantifying community assembly processes and identifying features that impose them. *ISME J.* **7**, 2069–2079 (2013).
26. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4766–4775 (2017).
27. Proulx, S. R., Promislow, D. E. L. & Phillips, P. C. Network thinking in ecology and evolution. *Trends Ecol. Evol.* **20**, 345–353 (2005).
28. Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
29. 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 (2021).
30. Segre, H. et al. Competitive exclusion, beta diversity, and deterministic vs. stochastic drivers of community assembly. *Ecol. Lett.* **17**, 1400–1408 (2014).
31. Chase, J. M. Stochastic community assembly causes higher biodiversity in more productive environments. *Science* **328**, 1388–1391 (2010).
32. Mendes, L. W. et al. Soil-borne microbiome: linking diversity to function. *Microb. Ecol.* **70**, 255–265 (2015).
33. Delgado-Baquerizo, M. et al. Soil microbial communities drive the resistance of ecosystem multifunctionality to global change in drylands across the globe. *Ecol. Lett.* **20**, 1295–1305 (2017).
34. Spain, A. M., Krumholz, L. R. & Elshahed, M. S. Abundance, composition, diversity and novelty of soil proteobacteria. *ISME J.* **3**, 992–1000 (2009).
35. Araujo, R. et al. Biogeography and emerging significance of Actinobacteria in Australia and Northern Antarctica soils. *Soil Biol. Biochem.* **146**, 107805 (2020).
36. Yuan, M. et al. Genome sequence and transcriptome analysis of the radioresistant bacterium *Deinococcus gobiensis*: insights into the extreme environmental adaptations. *PLoS ONE* **7**, e34458 (2012).
37. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**, 217–229 (2015).
38. Nekola, J. C. & White, P. S. The distance decay of similarity in biogeography and ecology. *J. Biogeogr.* **26**, 867–878 (1999).
39. Sun, Y. et al. Contribution of stochastic processes to the microbial community assembly on field-collected microplastics. *Environ. Microbiol.* **23**, 6707–6720 (2021).
40. Székely, A. J. & Langenheder, S. The importance of species sorting differs between habitat generalists and specialists in bacterial communities. *FEMS Microbiol. Ecol.* **87**, 102–112 (2014).
41. Pandit, S. N., Kolasa, J. & Cottenie, K. Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology* **90**, 2253–2262 (2009).
42. Jansson, J. K. & Hofmockel, K. S. Soil microbiomes and climate change. *Nat. Rev. Microbiol.* **18**, 35–46 (2020).
43. Bérard, A., Ben Sassi, M., Kaisermann, A. & Renault, P. Soil microbial community responses to heat wave components: drought and high temperature. *Clim. Res.* **66**, 243–264 (2015).
44. Sheik, C. S. et al. Effect of warming and drought on grassland microbial communities. *ISME J.* **5**, 1692–1700 (2011).
45. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-based assessment of Soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).
46. Griffiths, R. I. et al. The bacterial biogeography of British soils. *Environ. Microbiol.* **13**, 1642–1654 (2011).
47. Lauber, C. L., Strickland, M. S., Bradford, M. A. & Fierer, N. The influence of soil properties on the structure of bacterial and fungal communities across land-use types. *Soil Biol. Biochem.* **40**, 2407–2415 (2008).
48. Jenkins, S. N. et al. Actinobacterial community dynamics in long term managed grasslands. *Antonie Van Leeuwenhoek* **95**, 319–334 (2009).
49. Shen, C. et al. Soil pH drives the spatial distribution of bacterial communities along elevation on Changbai Mountain. *Soil Biol. Biochem.* **57**, 204–211 (2013).
50. Holland, J. E. et al. Liming impacts on soils, crops and biodiversity in the UK: a review. *Sci. Total Environ.* **610**, 316–332 (2018).
51. Grant, C. Residual effects of additions of calcium compounds on soil structure and strength. *Soil Tillage Res.* **22**, 283–297 (1992).
52. Chen, H., Ren, H., Liu, J., Tian, Y. & Lu, S. Soil acidification induced decline disease of *Myrica rubra*: aluminum toxicity and bacterial community response analyses. *Environ. Sci. Pollut. Res.* **29**, 45435–45448 (2022).
53. Groundcover. *Test Methods a Concern With Soil Aluminium*. <https://groundcover.grdc.com.au/agronomy/soil-and-nutrition/test-methods-a-concern-with-soil-aluminium> (2025).

54. Di Castri, F., Goodall, D. W. & Specht, R. L. *Ecosystems of the World. Ecosystems of the world CN - 577.38* (Elsevier scientific publ, Amsterdam Oxford New York, 1981).
55. Blondel, J. & Aronson, J. *Biology and Wildlife of the Mediterranean Region* (Oxford Univ. Press, Oxford, 2004).
56. Sala, O. E. et al. Global biodiversity scenarios for the year 2100. *Science* **287**, 1770–1774 (2000).
57. Liao, J. 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).
58. Sayre, R., Comer, P., Warner, H. & Cress, J. *A New Map of Standardized Terrestrial Ecosystems of the Conterminous United States*. <https://pubs.usgs.gov/pp/1768/> (2009).
59. Liao, J. et al. Comparative genomics unveils extensive genomic variation between populations of *Listeria* species in natural and food-associated environments. *ISME Commun.* **3**, 1–12 (2023).
60. Fasolo, A., Deb, S., Stevanato, P., Concheri, G. & Squartini, A. ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies. *PLoS ONE* **19**, e0309065 (2024).
61. Bokulich, N. A. et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59 (2013).
62. 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).
63. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
64. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
65. Liu, L., Yang, J., Yu, Z. & Wilkinson, D. M. The biogeography of abundant and rare bacterioplankton in the lakes and reservoirs of China. *ISME J.* **9**, 2068–2077 (2015).
66. Gobet, A., Quince, C. & Ramette, A. Multivariate cutoff level analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res.* **38**, e155–e155 (2010).
67. Levins, R. A. *Evolution in Changing Environments: Some Theoretical Explorations. Monographs in Population Biology* (Princeton Univ. Press, Princeton, NJ, 1974).
68. Cassie, R. M. Frequency distribution models in the ecology of plankton and other organisms. *J. Anim. Ecol.* **31**, 65 (1962).
69. Goh, Y.-X. et al. Evidence of horizontal gene transfer and environmental selection impacting antibiotic resistance evolution in soil-dwelling *Listeria*. *Nat. Commun.* **15**, 10034 (2024).
70. Liao, J. Data and code for ‘Differential roles of deterministic and stochastic processes in structuring soil bacterial ecotypes across terrestrial ecosystems in the United States. Zenodo <https://doi.org/10.5281/zenodo.14879726> (2025).

Acknowledgements

We thank LEAPH members for their valuable discussions. This work is funded by the Virginia Tech Global Change Center Seed Grant Program (JL). We also acknowledge the support from the Virginia Tech Via Ph.D. Fellowship (MR), NSF ATD 2124535 (XX), NSF DMS 241370 (ML), and Commonwealth Cyber Initiative (HZ).

Author contributions

J.L. designed the study. M.R., S.H., F.H., A.N., M.L., X.X., H.Z., and J.L. analyzed the data. J.L. and M.R. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-57526-x>.

Correspondence and requests for materials should be addressed to Jingqiu Liao.

Peer review information *Nature Communications* thanks Aline Frossard, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025