nature microbiology

Article

https://doi.org/10.1038/s41564-023-01573-x

Environmental stress mediates groundwater microbial community assembly

Received: 22 March 2023

Accepted: 28 November 2023

Published online: 11 January 2024

Check for updates

Daliang Ning[1,2], Yajiao Wang[1,2], Yupeng Fan[1,2], Jianjun Wang[1,3], Joy D. Van Nostrand[1], Liyou Wu[1,2], Ping Zhang[1,4], Daniel J. Curtis[1], Renmao Tian[1,5], Lauren Lui[6], Terry C. Hazen[7,8,9], Eric J. Alm[10], Matthew W. Fields[11], Farris Poole[12], Michael W. W. Adams[12], Romy Chakraborty[13], David A. Stahl[14], Paul D. Adams[6,9], Adam P. Arkin[6,9], Zhili He[1,15] & Jizhong Zhou[1,2,13,16,17]

Community assembly describes how different ecological processes shape microbial community composition and structure. How environmental factors impact community assembly remains elusive. Here we sampled microbial communities and >200 biogeochemical variables in groundwater at the Oak Ridge Field Research Center, a former nuclear waste disposal site, and developed a theoretical framework to conceptualize the relationships between community assembly processes and environmental stresses. We found that stochastic assembly processes were critical (>60% on average) in shaping community structure, but their relative importance decreased as stress increased. Dispersal limitation and 'drift' related to random birth and death had negative correlations with stresses, whereas the selection processes leading to dissimilar communities increased with stresses, primarily related to pH, cobalt and molybdenum. Assembly mechanisms also varied greatly among different phylogenetic groups. Our findings highlight the importance of microbial dispersal limitation and environmental heterogeneity in ecosystem restoration and management.

Disentangling ecological drivers controlling community assembly is crucial but challenging, especially in microbial ecology[1–3]. Community assembly refers to the process(es) by which species from a regional pool colonize and interact to establish and maintain local communities[4]. By focusing on deterministic assembly, niche-based theories emphasize the differences among species in their interactions with one another and the environment[5]. In contrast, neutral theories address the importance of stochastic assembly, assuming functional equivalence of all

[1]Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA. [2]School of Biological Sciences, University of Oklahoma, Norman, OK, USA. [3]State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China. [4]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA. [5]Institute for Food Safety and Health, Illinois Institute of Technology, Bedford Park, IL, USA. [6]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]Department of Earth and Planetary Sciences, Bredesen Center, Department of Civil and Environmental Sciences, Center for Environmental Biotechnology, and Institute for a Secure and Sustainable Environment, University of Tennessee, Knoxville, TN, USA. [8]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [9]Department of Bioengineering, University of California, Berkeley, CA, USA. [10]Department of Biological Engineering, Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA. [11]Center for Biofilm Engineering and Department of Microbiology and Cell Biology, Montana State University, Bozeman, MT, USA. [12]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA. [13]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [14]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA. [15]Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China. [16]School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, OK, USA. [17]School of Computer Science, University of Oklahoma, Norman, OK, USA. e-mail: jzhou@ou.edu

individuals/species[6]. To unify niche and neutral perspectives, Vellend proposed a conceptual framework that defines selection, dispersal, diversification and drift as four fundamental processes for ecological community assembly[7]. Strictly speaking, 'neutral' and 'stochastic' are not interchangeable, so are 'niche' and 'deterministic'[8]. In this study, we adopt practical definitions linked to Vellend's framework: 'deterministic processes' include selection, as well as non-random dispersal and diversification that are related to species-specific traits; 'stochastic processes' are featured by random changes in birth, death, migration or speciation; thus, in microbial communities, drift, dispersal and diversification are largely stochastic[1].

Community assembly processes have been extensively studied in plant and animal ecology, but much less in microbial ecology until recently (Supplementary Fig. 1). With the rapid development and broad application of high-throughput sequencing and associated experimental and computational technologies[9], unravelling assembly processes controlling the structure of microbial communities has attracted great attention[1]. However, given the extraordinary diversity of microbial communities, quantifying different assembly processes is challenging. Towards this challenge, various statistical approaches have been developed on the basis of multivariate analyses[3], neutral theory models[10,11] and null models[12–14]. Their applications to microbial communities revealed important insights into microbial assembly mechanisms, for example, homogeneous selection of grassland soil bacteria enhanced by climate warming[14], the predominant role of dispersal limitation in groundwater microbial assembly[15–17] and dramatically increased stochasticity of microbial functional diversity during biostimulation in polluted groundwater[13,18]. However, how community assembly processes change along space, time and environmental stresses remain elusive.

Stress is typically referred to as abiotic or biotic constraints on the productivity of species and on the development of ecosystem(s)[19], for example, salinity[20–22], toxic contaminants[23], drought[24,25], fire[26] and so on. Environmental stresses are major drivers of community variation[19,27]. While various previous studies reported that stresses led to deterministic assembly of macroorganisms[24,27–29] and microorganisms[20,22,23,25,26,30–37], microbial community assembly was also found to be more stochastic in many stressful environments[23,38–41]. To reconcile such divergent findings, it is critical to develop general frameworks to encapsulate the relationships between microbial community assembly processes and stress gradients. Thus, we examined the planktonic fractions of bacterial communities in groundwater at the Oak Ridge Field Research Center site (Oak Ridge, Tennessee, USA). Because of past nuclear waste disposal, the groundwater has been contaminated with extremely high levels of uranium, nitrate, technetium and various heavy metals. The pH in the groundwater varies greatly, ranging from 3.1 to 10.5 (refs. 42,43). Such wide ranges of diverse stressors are rarely found elsewhere. With extensive team effort, the site biogeochemistry and microbial community diversity of ~100 representative wells were comprehensively characterized, with more than 200 biogeochemical variables measured[42,43], which are very rare in typical microbial ecology studies. Thus, this groundwater ecosystem provides an unprecedented opportunity to discern the relationships between community assembly processes and environmental stresses. Two specific questions were addressed in this study: (1) How do different assembly processes of the groundwater communities change with environmental stresses and across different microbial lineages? (2) Which environmental factor(s) are most important in mediating the changes in community assembly processes in response to stresses? In this Article, we propose a theoretical framework to conceptualize the relationships between community assembly processes and environmental stresses (Fig. 1), followed by evaluations of the groundwater microbial communities. Our results revealed that stochasticity, dispersal limitation and 'drift' decrease, but determinism and heterogeneous selection increase with environmental stress.

Results

Theoretical framework

Four major schemas are proposed to describe the general relationships between community assembly processes and environmental stresses.

Schema A: stochasticity and environmental stress. In general, it is expected that stochastic assembly decreases as stress increases, whereas deterministic assembly increases[22,26,33,35]. High stress usually imposes strong selective pressure[19,44,45]. As stress increases, many species will be more suppressed, but those with higher tolerance or adaptation to the stressor(s) will thrive, leading to more deterministic community assembly[30–36] (Fig. 1a). In contrast, under low/no stress environments, species generally can grow faster, resulting in much higher frequencies of birth, death and migration, hence stochasticity is expected to decrease as stress increases (Fig. 1a). Explicit or implicit evidence to support this schema can be found in some studies on microbial[22,26,33,35,46] and other organismal communities[24,28].

Schema B: selection and environmental stress. Deterministic assembly is mainly attributed to selection (abiotic filtering and biotic interactions)[1]. If stress benefits a certain species with some specific functional traits, higher stress should lead to trait convergence (that is, similar trait(s) of co-occurring species)[47] as well as phylogenetic clustering in communities, assuming phylotype has trait coherence. If the stress is homogeneous in the environment, the local communities will become more functionally and phylogenetically similar as stress increases, resulting in greater homogeneous selection[1,48] (Fig. 1b(i)). This situation often exists in environments experiencing a single major stressor, such as a heat wave[49], a certain toxicant[50] or abnormal pH[51]. In contrast, if the stress is heterogeneous in the environment, the local communities may become more functionally and phylogenetically dissimilar with higher stresses because heterogeneous stress may select functionally and phylogenetically different species, leading to greater heterogeneous selection[1,48] (Fig. 1b(ii)). This situation can be common in heterogeneous environments with complex pollutants and/or other stressors[46], although it is less documented.

Schema C: dispersal and environmental stress. Higher stress, in general, is predicted to retard microbial colonization and establishment (for example, refs. 52,53). As a result, the dispersal of microorganisms will be more limited as stress increases (Fig. 1c(i)). In contrast, homogenizing dispersal requires nearly unlimited dispersal, thus its changes with stress could be opposite to those of dispersal limitation or non-significant (Fig. 1c(ii)). Homogenizing dispersal may exist in manipulated systems (for example, a well-mixed bioreactor) but rarely in natural ecosystems. It should be noted that microbial dispersal could either be stochastic or deterministic (see Supplementary Discussion A for details).

Schema D: drift and environmental stress. Higher stress usually decreases total population size and diversity[19], thus reducing probabilistic events of birth/death and resulting in less drift (Fig. 1d). This may be the main reason for decreased stochasticity under higher stress, but there is no quantitative evidence to support this. It should be noted that when assessed by null model approaches[12,14], the fraction which is not governed by selection or dispersal could largely reflect the influence of drift, but might also include effects of diversification, weak selection and/or weak dispersal[1,14]; hereafter, the fraction is designated as 'drift' for convenience.

Geochemical characteristics and stresses

The site is located within the Bear Creek Valley watershed in Oak Ridge. In the contaminated areas, the groundwater is polluted by radionuclides (for example, U and Tc), nitrate, sulfide and various heavy metals (Cd, Ni, Cr and so on) with abnormal pH (down to 3.1), mainly from the former S-3 ponds. The ponds were the primary accumulation site for nuclear waste disposal. Although they were closed and capped in 1988,

https://doi.org/10.1038/s41564-023-01573-x

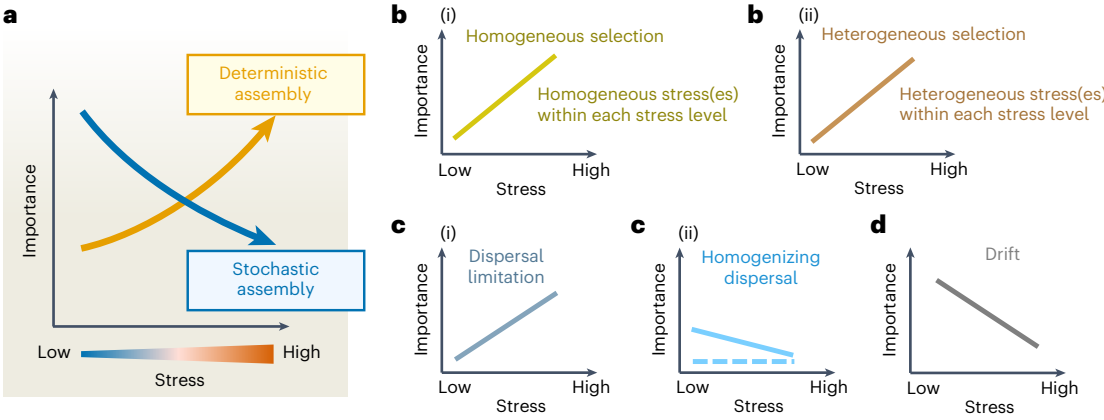


Fig. 1 | Schematic representation of the relationships between community assembly processes and stress. a, Deterministic assembly generally increases while stochastic assembly decreases as stress increases. **b**, Higher stress may result in greater importance of homogeneous selection (i) or heterogeneous selection (ii). **c**, In general, higher stress increases dispersal limitation (i) and decreases homogenizing dispersal (ii). Nevertheless, if homogenizing dispersal is negligible, its variation with stress may not be detectable (dashed line in (ii)). **d**, The influence of ecological drift generally decreases as stress increases. Definitions of terms[1]: 'selection' refers to major niche-based processes that shape community structure due to fitness differences among different organisms, including effects of abiotic conditions (environmental filtering) and biotic interactions (for example, competition, facilitation, mutualism, predation and host filtering and so on). 'Homogeneous selection' refers to selection under homogeneous abiotic and biotic environmental conditions leading to more-similar structures among communities. 'Heterogeneous selection' refers to selection under heterogeneous environmental conditions leading to more-dissimilar structures among communities. 'Dispersal' refers to movement and successful colonization (establishment) of an individual organism from one location to another via both active and passive mechanisms. 'Dispersal limitation' means that movement of individuals to and/or establishment of individuals (colonization) in a new location is restricted. 'Homogenizing dispersal' indicates a very high rate of dispersal among communities, which homogenizes communities such that their structures are very similar. 'Drift' refers to random changes in the relative abundances of different species within a community due to the inherent stochastic processes of birth, death and reproduction.

contaminants from these ponds leached out, creating a groundwater contaminant plume across the site[42,43]. We measured over 200 variables, including geographic factors, temperature and various geochemical characteristics in gaseous, liquid and solid phases, covering pH, conductivity, C, N, O, S, P, Cl, Br and 176 measurements of 56 metals in different phases (Supplementary Table 1). The results demonstrated that the groundwater microbial communities were under multiple stressors in the contaminated areas, including abnormal osmotic pressure from high salinity (conductivity up to 20,620 μS cm^{-1}), extreme acidity (pH down to 3.1) or alkalinity (pH up to 10.0), extremely high concentrations of nitrate (up to 14,446 mg l^{-1}), uranium (up to 16.6 mg l^{-1}) and other metals (Supplementary Table 1 and Extended Data Fig. 1). We formulated a stress index (SI) as the observed concentration of each stress indicator divided by its reference level from water regulations, guidelines or a toxicity database (Supplementary Tables 2 and 3). After log transformation, the maximum stress index (MSI) in each sample showed a significant negative correlation ($r = -0.603$, $P < 10^{-9}$) with the observed number of reference taxa that are common and abundant in uncontaminated areas (Extended Data Fig. 2), suggesting that MSI can be used as a rough classification of stress levels. On the basis of MSI, the 91 samples were divided into 7 stress levels ($n = 13$ per level; Supplementary Table 3 and Fig. 2a). The MSI values and stress levels were mapped to illustrate their spatial distribution (Fig. 2a,b). The wells around the contamination source ponds had the highest stress level, with maximum SIs for nitrate (9 samples), U (3 samples) or H^+ (1 sample; Extended Data Fig. 1a–c and Supplementary Table 3). The wells downstream of the source pond also had high stress (MSI > 10), but the MSI decreased with distance from the source pond. Besides MSI, we also explored two other stress metrics based on stressor number and reference taxa richness and three other options to define stress levels (Methods and Supplementary Fig. 2).

Deterministic versus stochastic assembly

The planktonic fractions of bacterial communities from groundwater were analysed by 16S ribosomal (r)RNA gene sequencing, with a total of 28,644 observed operational taxonomic units (OTUs) (97% similarity). The environmental stresses showed a negative correlation with bacterial richness based on iChao1 estimation ($r = -0.284$, $P = 0.006$; Extended Data Fig. 3), while there were no significant correlations based on Shannon index ($r = -0.165$, $P = 0.117$). The stresses also showed significant impacts on taxonomic (permutational multivariate analysis of variance (ANOVA) $F = 1.66$, $P = 0.004$) and phylogenetic ($F = 2.93$ $P = 0.056$; Supplementary Table 4) β-diversity, suggesting that deterministic processes such as environmental filtering might have impacted the diversity patterns.

To further test our hypothesis with respect to stochastic-vs-deterministic assembly, we employed several well-established complementary approaches. First, multivariate analysis was utilized to estimate community variation explained by environmental variables after controlling for spatial influence, the so-called environment effect[3]. Taking advantage of our comprehensive measurements, the environment effect was estimated at each stress level, dramatically changing from 16.5% to 66.9% as stress increased, leading to a significant positive correlation with MSI ($R^2 = 0.904$, $P = 0.001$; Fig. 3a). In addition to the multivariate analysis, various approaches were used to estimate stochasticity by comparing observed with random expectations simulated by neutral theory or other null models, including neutral taxa percentage (Extended Data Fig. 4a), normalized stochasticity ratio (NST, Fig. 3b), stochastic turnover ratio (Extended Data Fig. 4b) or stochastic process influence (Extended Data Fig. 4c). The results of all approaches demonstrated a significant decrease in stochastic assembly as stress increased ($R^2 = 0.740$–0.967, $P < 0.014$; Fig. 3a,b and Extended Data Fig. 4). At the four lower stress levels (MSI < 9.5), the estimated stochasticity values ranged from 55% to 85% on average, and the 25% bootstrapping quartiles of stochasticity were always higher than 50%. Correspondingly, the average environment effect was lower than 22% at the four low stress levels. These results suggested the predominant roles of stochastic processes in shaping the groundwater bacterial communities without severe contamination. The highest stress level (MSI > 145) showed the highest environmental effect (66.9%) and the

https://doi.org/10.1038/s41564-023-01573-x

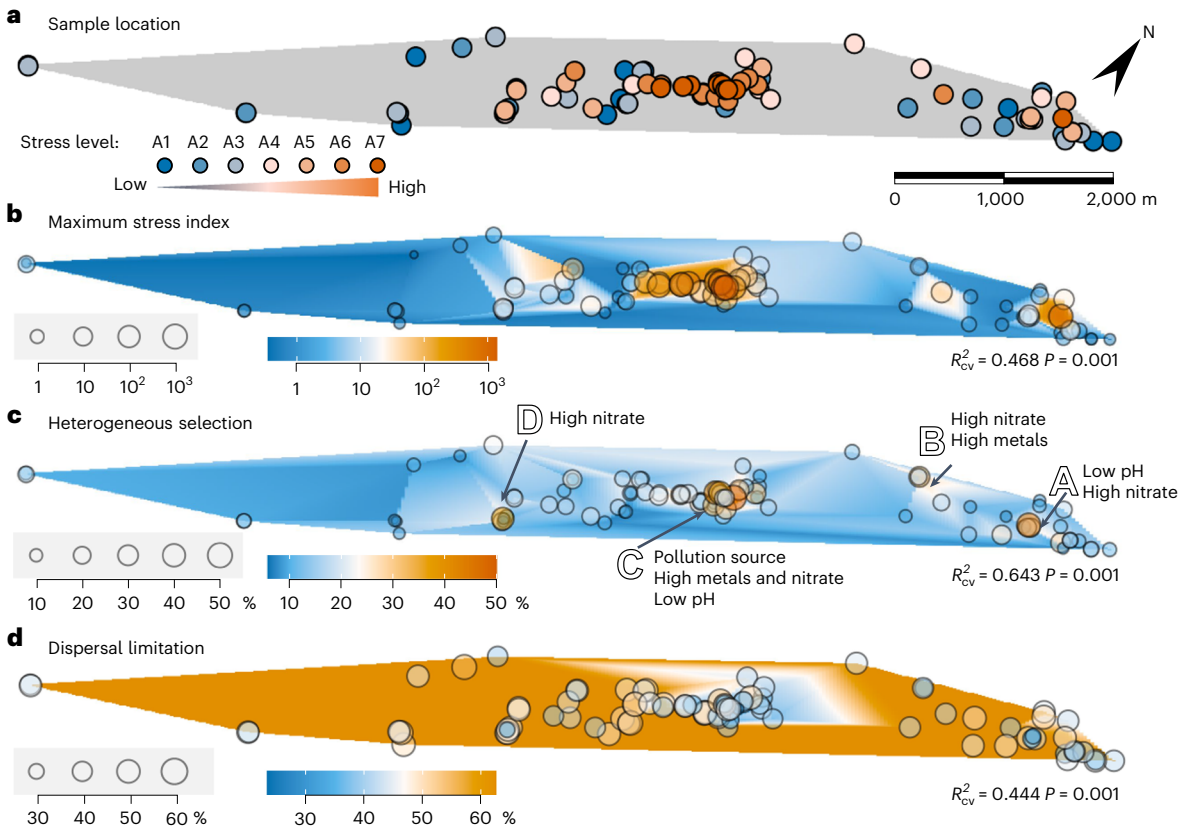


Fig. 2 | Maps of sampling positions, stress and relative importance of community assembly processes. a, Sample location. **b**, MSI, with higher values representing higher stress. **c**, Heterogeneous selection. **d**, Dispersal limitation. In **b**–**d**, the bubble size represents the value at each sampling position; the colour bars reflect predicted values; R^2_{cv} is cross-validated R^2 of the model used to generate each map; *P* values (one-sided) are based on permutational test; A–D indicate hotspots of heterogeneous selection.

lowest stochasticity (down to 41%). Nevertheless, the estimated stochasticity was still 41%–67% based on different methods, even still significantly ($P < 0.001$) higher than 50% based on NST (Fig. 3b) and the method of inferring community assembly mechanisms by the phylogenetic-bin-based null model (iCAMP; Extended Data Fig. 4c), indicating that stochasticity decreased but was still essential in shaping bacterial assembly in this highly contaminated environment. Besides MSI-based assessment, the other options showed similar patterns (Supplementary Fig. 3). Overall, the results clearly supported Schema A (Fig. 1a) of our theoretical framework.

Relative influence of different assembly processes

The relative importance of different assembly processes was estimated with iCAMP, which showed improved quantitative performance from previous approaches[14]. Across the whole site, heterogeneous selection played a major role (20.3 ± 15.0%; Fig. 3c), while homogeneous selection had a relatively small role in controlling spatial community turnovers (10.3 ± 8.0%, mean ± s.d.; Extended Data Fig. 5a). Heterogeneous selection obviously increased as the stress rose ($R^2 = 0.826$, $P = 0.004$) and became the second most influential (30.0 ± 4.3%; Fig. 3d) process at the highest stress level, whereas homogeneous selection did not show any significant change across different stress levels ($R^2 = 0.026$, $P = 0.727$; Extended Data Fig. 5a). Overall, dispersal limitation dominated the community turnovers (44.2 ± 13.4%) and showed the strongest relative influence across all stress levels (from 39% to 48% on average; Fig. 3c), corresponding to the strong physical filtering/trapping effect of subsurface environment on microorganisms in groundwater. Despite its dominance across all stress levels, dispersal limitation significantly decreased as stress increased ($R^2 = 0.697$, $P = 0.019$; Fig. 3e), in accordance with shorter distances and better dispersal conditions in highly contaminated areas[54]. In contrast, homogenizing dispersal showed the lowest influence (1.9 ± 2.1%; Fig. 3c) and no significant change with stress ($P = 0.250$; Extended Data Fig. 5b). 'Drift' also played a significant role (23 ± 4.6%; Fig. 3c) in shaping community diversity, and its importance significantly decreased ($R^2 = 0.738$, $P = 0.013$) as stress increased (Fig. 3f). Besides MSI-based assessment, other options showed similar trends of heterogeneous selection, dispersal limitation and 'drift' as stress increased (Supplementary Fig. 3). All the results are consistent with our predictions based on Schemas B and D but opposite to those based on Schema C (Fig. 1b–d).

Assembly processes in different phylogenetic groups

In the iCAMP analysis, the observed OTUs were divided into 262 phylogenetic groups (bins) on the basis of the phylogenetic signal threshold; each was assessed separately for the relative importance of different assembly processes (Fig. 4a). Overall, dispersal limitation, heterogeneous selection and 'drift' dominated 60%, 31% and 6.5% of the bins, corresponding to 68%, 13% and 17% relative abundance, respectively. In contrast, homogeneous selection dominated only 2.7% of the bins, and homogenizing dispersal did not overwhelm other processes in any bin. When counting in the abundance of each bin, the impact of heterogeneous selection and dispersal limitation were mainly (>60%) attributed to the responses of bins in γ- and α-Proteobacteria, Verrucomicrobiota, Bacteroidota and Nitrospirota (Fig. 4b). The three most abundant bins governed by heterogeneous selection were from yet uncultivated phyla, including a candidate phylum WPS-2, an unclassified phylum and a candidate phylum RCP2-54 (Supplementary Table 5a). For the dominant taxon in each of these three bins, the BLAST-based

https://doi.org/10.1038/s41564-023-01573-x

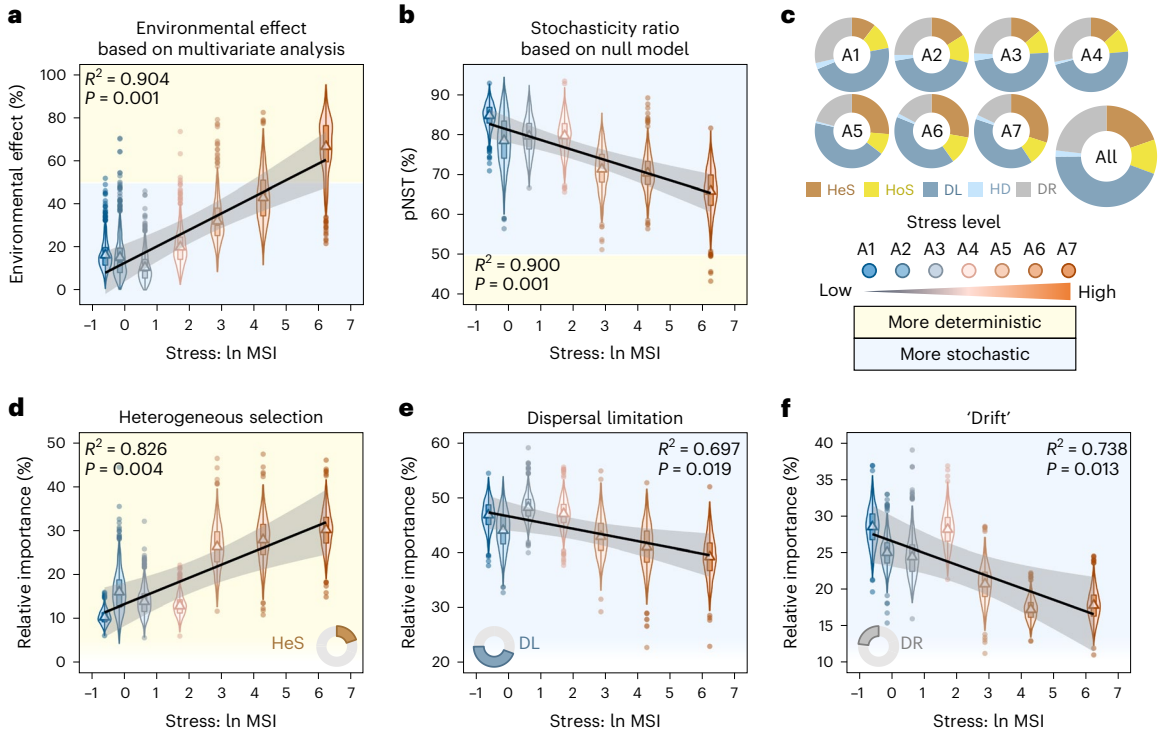


Fig. 3 | Variation in community assembly processes at different stress levels. a,**b**, Determinism and stochasticity of groundwater bacterial assembly, reflected by environmental effect as estimated via multivariate analysis (**a**) and phylogenetic normalized stochasticity ratio (pNST) based on null model analysis of phylogenetic beta diversity βMNTD (**b**) (see Extended Data Fig. 4 for the results of other methods). **c**–**f**, Relative importance of assembly processes at different stress levels, including their mean expectation (**c**) and the variations in heterogeneous selection (HeS; **d**), dispersal limitation (DL; **e**) and 'drift' (DR; **f**) (see Extended Data Fig. 5 for homogeneous selection (HoS) and homogenizing dispersal (HD)). The violin and boxplots in **a**,**b**,**d**–**f** are based on bootstrapping results at each stress level (n = 13 per level; bootstrapping 1,000 times). Colours of violin and boxplots indicate the stress levels. In boxplots: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; dots, outliers; triangles, mean value at each stress level. Black line, grey shadow, R^2 and P values are the trendline, 95% confidence interval, coefficient of determination and significance, respectively, based on linear regression of the mean values as a function of log-transformed MSI.

high-identity hits (>99%) in NCBI databases are all as yet uncultivated and from very similar habitats, suggesting high probabilities to discover unknown species with essential functions in this site[55,56]. In contrast, the top three bins governed by dispersal limitation and those by 'drift' are from the phyla of Proteobacteria and Bacteroidota. Their dominant taxa generally had high-identity BLAST hits from various distinct habitats, implicating larger niche width (Supplementary Table 5b).

Spatial variations in different assembly processes

The influence of different assembly processes on the groundwater microbial communities was mapped to visualize their spatial variations. The map of heterogeneous selection showed four hot areas (Fig. 2c), including the area around the contamination source ponds and several areas featured with low pH, high nitrate, high concentrations of heavy metals (for example, U) or their combinations. Dispersal limitation was generally influential across the site (Fig. 2d). The sampling positions with higher scores of dispersal limitation were found around the boundaries corresponding to the sharp change in dissolved oxygen (Extended Data Fig. 1d), indicating physical barriers for microbial dispersal in groundwater. In summary, the hotspots of heterogeneous selection and dispersal limitation, although solely estimated from microbiome data, showed reasonable consistency with the spatial distribution of environmental stressors and hydrogeological conditions.

Environmental variables affecting community assembly

Since community assembly was largely governed by heterogeneous selection and dispersal limitation with clear patterns along stress gradients, some environmental and spatial variables should play important roles in controlling groundwater microbial community assembly. A cross-validated Mantel test was developed to show associations between individual variables and each assembly process. The estimated relative importance of different assembly processes was first transformed to central log ratios for use in the Mantel test to ameliorate compositional data issues[57]. The significant ($R^2_{CV} > 0.01$, $P < 0.05$) pairwise correlations were visualized with a network (Fig. 5). Heterogeneous selection obviously showed much more and stronger associations with measured environmental variables (35%), including pH and various metals in the supernatant and suspended solid, with average $R^2_{CV} = 0.043$ and maximum $R^2_{CV} = 0.126$ (Supplementary Table 6). In contrast, 'drift', homogeneous selection and homogenizing dispersal were only significantly correlated with <8.5% of the variables, with all R^2_{CV} being <0.03; dispersal limitation did not show significant correlation with any individual variable (all $R^2_{CV} < 0$, although some $P < 0.05$).

Considering strong collinearity among measured variables (grey lines in Extended Data Fig. 6) and having fewer samples than variables, orthogonal partial least squares (OPLS) was applied to reveal major determinants underlying the variation in each assembly process (Supplementary Table 7). The OPLS models of different processes are significant ($P \leq 0.005$ for R^2_Y, the variance of the dependent variable Y explained by the model) without obvious overfitting ($P \leq 0.01$ for Q^2_Y, the predictive accuracy of the model), except for the model of homogenizing dispersal ($P = 0.15$ for R^2_Y, $P = 0.09$ for Q^2_Y). The OPLS model explained 48.3% of the variation in heterogeneous selection, with pH, Co, Mo and Eu as top important variables (variable importance

https://doi.org/10.1038/s41564-023-01573-x

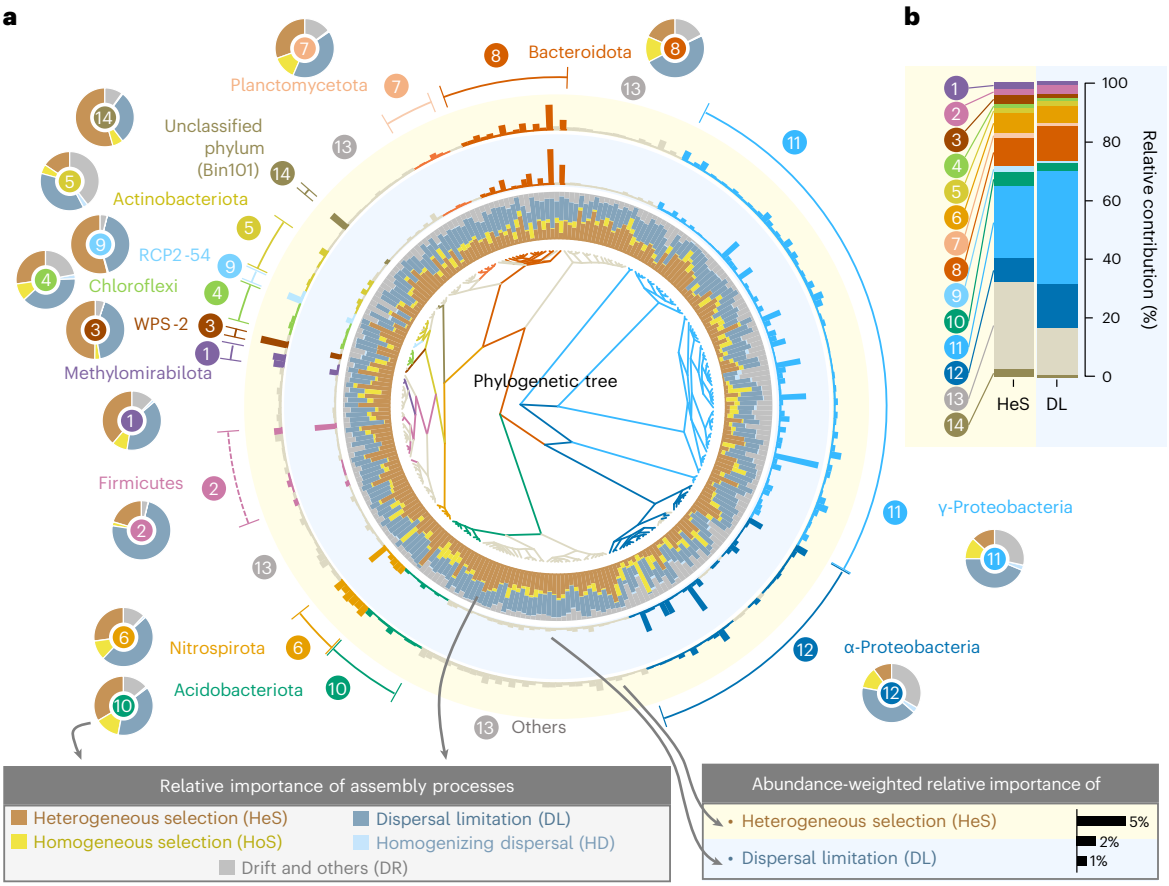


Fig. 4 | Variations in assembly mechanisms across different phylogenetic groups. a, Phylogenetic tree (centre), relative importance of different processes (inner annulus) and abundance-weighted importance of DL (middle annulus) and HeS (outer annulus) in each phylogenetic bin from iCAMP analysis, as well as relative importance of different processes in each phylum (or class of Proteobacteria, donut plots by phylum or class names). **b**, Relative abundance of different phyla or classes governed by HeS or DL. Colours indicate phyla or classes in the phylogenetic tree branches, in bars in the middle and outer annuli (**a**) and in stacked bars (**b**), as well as different processes in the inner annulus and donut plots (**a**).

in projection, VIP > 2), implicating abnormal pH and deficiency in Co and Mo as the major drivers of selection variation (detailed in Supplementary Discussion B). In contrast, the model only explained 11.3% of the variation in dispersal limitation, in which some chemical properties (for example, Mo and Zn, VIP > 2) were more important than geographic distance (VIP = 0.345). This might be because microbial dispersal in groundwater is primarily controlled by site geology and complicated groundwater flow pathways where some chemical substances and microbes could be constrained by similar site geological conditions.

Discussion

Understanding how ecological communities respond to environmental stresses is a critical topic in ecology[19]. Although the impacts of various stresses on microbial community diversity, structure and interactions have been examined in microbial ecology (for example, refs. 19,32,58), knowledge on how community assembly processes are affected by stress remains rudimentary. This study proposed a theoretical framework to conceptualize the general relationships between community assembly processes and environmental stresses and tested them with groundwater bacterial communities over a wide range of stress conditions. Consistent with our theoretical prediction, the importance of deterministic processes increased with stress (Fig. 1a). These results are also supported by several previous studies, demonstrating that microbial community assembly under stress is largely deterministic[20,22,23,25,26,30–37]. However, it should be noted that, theoretically, the opposite alternatives could also exist. For instance, if the stress is so harmful or even fatal that no species can persist or get any advantage, the observed species co-occurrence could be very random, hence higher stress could lead to more stochastic assembly[59].

While the importance of deterministic assembly was established by numerous studies in the early age of microbial ecology[1], the importance of stochastic processes in microbial assembly has been appreciated just in the recent decade[1,14,18,38,60]. This study showed that stochastic processes are still essential even under high stress, which is in agreement with other groundwater studies[13,17,18,61–63]. Considering the large randomness of microbial dispersal in groundwater due to filtering, adsorbing and trapping effects in subsurface porous structures, this is understandable. Besides, many studies on microbial assembly in other ecosystems also suggest that stochastic assembly could be very important even under extreme environments[38–41]. Thus, compared with the dominance of deterministic processes in stressful environments[20,22,26,37], the increase in determinism with stresses should be more general.

While other assembly processes (for example, selection) varied with stress gradients as we predicted (see Supplementary Discussion B for details)[19,21,24,27–29,42,51,64–78], dispersal limitation was negatively related to stress, opposite to Schema C. This is reasonable when microbial dispersal relies more on physical transport conditions (see Supplementary Discussion C for details)[15–17,54,79,80]. Also, knowledge on community assembly mechanisms under stress conditions is important for better subsurface modelling[81] (see Supplementary Discussion D for details). In addition, great caution is needed in data interpretation due to some

https://doi.org/10.1038/s41564-023-01573-x

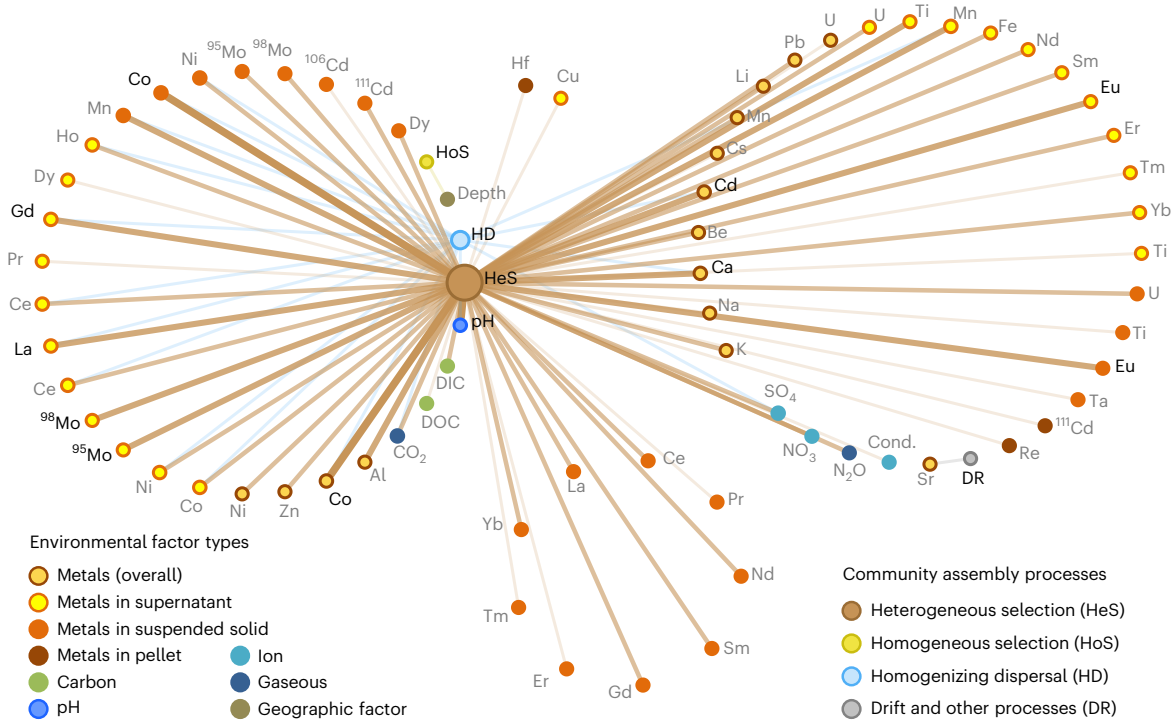


Fig. 5 | Environmental variables significantly correlated with each assembly process. Links show significant correlation ($R^2_{CV} > 0.01$ and $P < 0.05$); P values were calculated on the basis of the Mantel test (one-sided) and adjusted using the false discovery rate method; line width reflects the R^2_{CV}. Node area is related to its degree (associated node number). Nodes annotated with bold names represent assembly processes and top associated variables with $R^2_{CV} > 0.07$. Values of R^2_{CV} and significance are listed in Supplementary Table 6. Nodes are grouped by modules from the network counting in all significant correlations among environmental variables and assembly processes (Extended Data Fig. 6). DIC, dissolved inorganic carbon; DOC, dissolved organic carbon; Cond., conductivity.

caveats related to dispersal[2,3,82,83], methodology[1,13,14] and biofilms (see Supplementary Discussion A for details).

In conclusion, the results presented in this study are important for both basic and applied ecology research. Several theoretical schemas have been developed, which should be useful in formulating testable and compelling hypotheses for experimental studies. Since this framework was tested with groundwater microbiome data, further studies are needed to determine whether it is applicable to other ecosystems and/or under different stress conditions. Our findings also have important implications for ecosystem restoration and environmental management. As demonstrated in this study, stochastic processes, particularly dispersal limitation, play critical roles in shaping groundwater microbiomes, even under high stress conditions. Thus, to achieve a desired ecosystem state, any groundwater restoration programme must consider approaches to overcome dispersal limitation, for example, via inoculation. In addition, heterogeneous selection is important in mediating groundwater microbiome structure and is strongly associated with several key environmental factors such as pH, microelements or toxic metals. Thus, bioremediation strategies should also consider its importance to achieve restoration goals by altering the heterogeneity of environmental conditions and/or specific key environmental factors such as pH, nutrients and electron donors, as previously demonstrated[18,84].

Methods

Experimental site and sampling

The Department of Energy's Oak Ridge Field Research Center site consists of 243 acres of contaminated area and 402 acres of an uncontaminated background area in Oak Ridge, Tennessee. More detailed information can be found in previous publications[42,43] and the website http://www.esd.ornl.gov/orifrc. Groundwater samples were collected from 97 representative wells (91 wells with more complete data were used in this study; Fig. 2a) on the basis of a careful design to reflect the large geochemical gradient across the site[42].

Physical and chemical analyses

Temperature, pH, dissolved oxygen, conductivity and redox, were determined by an In-Situ Troll 9500 system (In-Situ Inc.). Sulfide and ferrous iron concentrations were measured using the US Environmental Protection Agency (EPA) methylene blue method (Hach; EPA Method 8131) and the 1,10-phenanthroline method (Hach; EPA Method 8146), respectively. Dissolved gases (He, H_2, N_2, O_2, CO, CO_2, CH_4 and N_2O) were measured on an SRI 8610C gas chromatograph (SRI Instruments) with argon carrier gas, using a method derived from EPA RSK-175 and the US Geological Survey Reston Chlorofluorocarbon Laboratory procedures. Dissolved organic carbon (DOC) and dissolved inorganic carbon (DIC) concentrations were determined with a Shimadzu TOC-V CSH analyser (EPA Method 415.1). The concentrations of anions (bromide, chloride, nitrate, phosphate and sulfate) were determined using a Dionex 2100 system (Thermo Fisher) with an AS9 column and a carbonate eluent (US EPA Methods 300.1 and 317.0). Metals and trace elements in the groundwater were determined on an inductively coupled plasma mass spectrometry (ICP-MS) instrument (Elan 6100; PerkinElmer) using a method similar to EPA Method 200.7. Metals were analysed in suspended solid, supernatant and pellet[85] samples. Each sample was vigorously shaken before taking 6 ml as 'suspended solid' sample. Another aliquot (6 ml) of each sample was centrifuged at 7,000 × g for 15 min to get the supernatant and pellet parts. The three parts were acidified to 2% using concentrated HNO_3 and analysed by ICP-MS (7500ce, Agilent). Physical and chemical data were stored using Excel in Microsoft 365.

Of the 202 measured environmental and spatial variables, 20 variables have no missing value, 175 variables have only 1 missing value and 7 variables have 4–12 missing values. For each variable not analysed in a sample, the missing value was estimated by multiple linear

regression using other variables measured in the sample. The R^2 values of the models were as high as 0.99 ± 0.04 (Supplementary Table 8). Thus, we accepted the estimates of missing values and used them in further analysis.

To map the measured variables, Akima's method was applied for bivariate interpolation and smooth surface fitting using the linear model function 'interp' in the R package 'akima' (v.0.5-1.2)[86]. To evaluate the interpolation, the Monte Carlo method was applied to calculate the cross-validated R^2 (R^2_{CV}) of the interpolation, with 90% of the samples randomly selected as the training set each time. The linear interpolation of 25 environmental variables and 5 ubiquitous taxa showed $R^2_{CV} > 0.25$ and $P < 0.05$ (permutational test), and was thus considered relatively reliable. Then, the geographic distribution of the variables was visualized with colour-filled contour maps using the function 'filled.contour' in the R package 'graphics' (v.4.2.2). For a variable without reliable linear interpolation, an optimum predictive model was built with the 30 reliably interpolated variables as predictors, by forward selection of the linear model or a random forest[87] model based on R^2_{CV}. The models fed with log-transformed and/or untransformed predictor values were compared for optimum prediction with maximum R^2_{CV}. Meanwhile, the dependent variable values were transformed by logit function to avoid the predicted values exceeding reasonable range. With the optimum model, the variable values at unsampled positions were predicted; only when the model of a variable showed $R^2_{CV} > 0.25$ and $P < 0.05$ were the predicted values used to draw a colour-filled contour map. We emphasize that the maps only help to visualize the general trends, and the interpolation or prediction at a specific position could have poor accuracy.

Stress levels

On the basis of various physical and chemical measurements, an integrated metric was developed to reflect different stress levels. The high concentrations of metals, H^+, OH^-, some anions and high conductivity all increase stress (so-called stress indicators; listed in Supplementary Table 2) and were used to evaluate stress levels. We calculated the SI of each stressor by dividing the observed value by its reference level. The reference levels were from the standards in drinking water regulations or guidelines; if not available, the concentrations without symptoms or toxic effects in a toxicity database were considered (Supplementary Table 2). Similar to Liebig's law, the stress level of a sample should be better determined by the MSI. To evaluate whether MSI can roughly represent stress to common microorganisms, Pearson correlation was performed between log-transformed MSI and reference taxa number. The 'reference taxa' were identified as taxa with mean relative abundance >0.1% and occurrence frequency >50% in 'uncontaminated' (all SI < 1) samples (Supplementary Table 9). Considering that similar sample size at each stress level is better for statistical analysis, we classified the 91 samples into 7 stress levels (Supplementary Table 3 and option A in Supplementary Fig. 2). To explore within-stress-level environmental dispersion at different stress levels, the permutational test of multivariate homogeneity of group dispersions (PERMDISP)[88] was performed using the functions 'betadisper' and 'permutest.betadisper' in the R package 'vegan' (v.2.5-7)[89]. PERMDISP is based on within-group Euclidean distances of all measured environmental variables standardized to zero mean and unit variance. To test the robustness of our conclusions, the samples were also grouped to different stress levels by three other options (Supplementary Notes and Fig. 2): Option B: grouping by the number of stressors that are defined as stress indicators with SI > 1 in a sample; Option C: grouping by the suppressed rate of reference taxa richness calculated as relative difference from the maximum reference taxa richness in uncontaminated samples; and Option D: grouping by hierarchical clustering of samples based on stressor concentrations and then selecting samples in each group to make the environmental dispersion the same in different groups.

Sequencing

The sequencing data are from our previous study[42]. An aliquot (4 l) of each sample was sequentially filtered through 10-μm- and 0.2-μm-pore-size filters. DNA was extracted from 0.2 μm filters using a modified Miller method[42], and the V4 region of 16S rRNA genes was amplified and sequenced on a MiSeq sequencer (Illumina). The primers were 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′). The sequencing data were processed using USEARCH (v.6.0.307) and custom scripts at https://github.com/spacocha/16S_pre-processing_scripts/ to generate OTUs at 97% identify[42]. The phylogenetic tree was built using FastTree (v.2.1.11)[90] after sequence alignment with PyNAST (v.1.2.2)[91]. Taxonomic classification was implemented in QIIME2 (v.2021.2) with the sklearn-based taxonomy classifier (v.0.23.1)[92], using Silva SSU 138 (ref. 93) as a reference database. Bacterial richness was estimated by the improved Chao 1 index (iChao1)[94]. Taxonomic and phylogenetic beta diversity were measured using Bray–Curtis dissimilarity and β-mean nearest taxon distance (βMNTD)[95]. We also generated amplicon sequence variant tables using UNOISE in USEARCH (v.11.0.667), which showed similar trends as OTUs, although less significant (Supplementary Fig. 4).

Phylogenetic signal

To infer niche selection with phylogenetic diversity metrics, species niche preference should have phylogenetic signal, for example, significant association between niche difference and phylogenetic distance. To test phylogenetic signal, niche value was calculated as the abundance-weighted mean of each environmental factor across all samples for each OTU[95] and its difference between two OTUs represents the between-taxa niche difference. In this way, the pairwise niche differences were calculated using the 'dniche' function in the R package iCAMP (v.1.6.1). To visualize the range of phylogenetic distances showing significant correlation with niche differences, the relationship curves were drawn and Mantel correlogram analysis[12,96] and stepwise Mantel test were performed using the function 'big.mantel.correlog' in the R package iCAMP. The 'big.mantel.correlog' is based on 'mantel' and 'mantel.correlog' in the R package 'vegan'[89] and improved to handle big datasets by using the R package 'bigmemory' (v.4.6.1) (see more details in Extended Data Fig. 7).

Multivariate analysis

Since phylogenetic signal was found significant within a short phylogenetic distance (Extended Data Fig. 7), βMNTD, which showed most values (>90%) to be within a distance of 0.2, was calculated with the function 'bmntd.big' in the R package 'NST' (v.3.1.10)[13] and used as the dependent variable in distance-based redundancy analysis[97]. All measured environmental variables were scaled to zero mean and unit variance. Geographic distance was transformed using principal coordinates of neighbourhood matrix[98] with default setting in the function 'pcnm' of the R package 'vegan'[89]. The distance-based redundancy analysis was performed for each stress level separately, with forward model selection based on adjusted R^2 and P values using the functions 'capscale' and 'ordiR2step' in 'vegan'.

Neutral theory model

A 'neutral taxon' was identified as an OTU whose occurrence frequency is within the 95% confidence interval of the frequency predicted by its relative abundance in the regional pool according to the neutral theory model[10,11,14]. Then, the abundance-weighted percentage of neutral taxa (NP) in a sample was calculated as the relative abundance sum of neutral OTUs in the sample[10]. The regional pool was not simply calculated as the arithmetic average of observed samples, considering the geographically uneven sampling and the large difference in biomass among wells. To estimate the regional pool, the density of each taxon in each well was assessed by multiplying observed relative abundance and acridine orange direct counts. Then, the log-transformed density

of each taxon was subjected to bivariate interpolation with Akima's method to assess its densities in evenly distributed grids without measured wells. Subsequently, the regional pool, counting all the grids, was input as 'meta.com' when NP was calculated using the function 'snm.comm' in the package iCAMP (v.1.6.1)[14].

Null-model-based analysis

A previous framework based on null models of the entire community (QPEN)[12,48] was applied to assess the relative importance of different community assembly processes. For each pair of communities (samples), if the phylogenetic dissimilarity measured by βMNTD was significantly higher (β nearest taxon index, βNTI > 2) or lower (βNTI < −2) than the null model expectation, the community turnover was considered to be governed by heterogeneous or homogeneous selection, respectively. If the phylogenetic dissimilarity was not differentiable from the null expectation (−2 ≤ βNTI ≤ 2) and taxonomic dissimilarity (Bray–Curtis) was significantly higher (the modified Raup–Crick metric based on Bray–Curtis dissimilarity, $RC_{Bray} > 0.95$) or lower ($RC_{Bray} < -0.95$) than the null model expectation, the community turnover was regarded as controlled by dispersal limitation or homogenizing dispersal, respectively. If neither phylogenetic nor taxonomic dissimilarity was differentiable from null expectations ($-2 \le \beta NTI \le 2$ and $-0.95 \le RC_{Bray} \le 0.95$), the turnover was regarded as governed by stochastic drift, diversification, weak selection and dispersal[1], named 'undominated'[12,48] or 'drift'[14] for short. In each group (for example, a stress level), the relative importance of a certain process was assessed by the percentage of pairwise turnovers governed by the process. For comparison with other stochasticity indexes, the turnovers governed by dispersal limitation, homogenizing dispersal or 'drift' processes were deemed as stochastic turnovers and their percentage in a group was utilized as a measure of stochasticity, named stochastic turnover ratio.

A phylogenetic-bin-based framework (iCAMP) was recently developed from QPEN and demonstrated improved quantitative performance[14]. Considering that different taxa in the same community can be dominated by different assembly processes, iCAMP divides the observed taxa into different phylogenetic groups (bins) and identifies the assembly processes governing each bin's turnovers in a similar way as QPEN. For each bin, the fraction of pairwise turnovers with significantly higher (β net relatedness index of a bin (βNRI_{bin}) > 1.96) or lower ($\beta NRI_{bin} < -1.96$) phylogenetic dissimilarity (β mean pairwise distance) than the null expectation reflects the influence of heterogeneous or homogeneous selection, respectively. The fraction of turnovers with non-significant phylogenetic dissimilarity ($-2 \le \beta NRI_{bin} \le 2$) but significantly higher ($RC_{bin} > 0.95$) or lower ($RC_{bin} < -0.95$) taxonomic dissimilarity (Bray–Curtis) than the null expectation reflects the influence of dispersal limitation or homogenizing dispersal, respectively. The remainder ($-2 \le \beta NRI_{bin} \le 2$ and $-0.95 \le RC_{bin} \le 0.95$) represents the influence of 'drift'. The relative influence of each assembly process in each bin was summarized and weighted with the bins' relative abundances to identify the major bins governed by each process, which was then used to quantify the process influence at the community level[14]. The QPEN, iCAMP, bin-level statistics and related significance test were performed using the functions 'qpen', 'icamp.big', 'icamp.bins' and 'icamp.boot' in the R package iCAMP (v.1.6.1), respectively.

Assembly process influence map

In a previous study[48], the influence of an assembly process at each location was estimated as its mean relative importance in all community turnovers between this location and all other locations. We made some modification considering the uneven geographic distribution of measured wells in this study. Bivariate interpolation was applied to estimate the relative importance of each process in community turnovers among evenly distributed grids from the values among measured wells. Then, the process influence at each well was calculated as its mean relative importance in turnovers between the well and all other grids. Subsequently, the process influence was visualized as a colour-filled contour map as described above for mapping other measured variables.

Associations among processes and environmental variables

The relative importance of each assembly process estimated by iCAMP was transformed to centred log ratio to avoid a compositional data issue[57]. Then, the transformed values of each process were analysed for correlation with each log-transformed environmental variable using Mantel test with cross validation. Pairwise correlations among environmental variables were assessed using a cross-validated linear model. The Monte Carlo method was applied for the cross validation, with 90% of the data used for training and 10% used as test data. To visualize the correlations, significant associations with $R^2_{CV} > 0.01$ and $P < 0.05$ were used to draw a network, and the modules were identified using greedy optimization[99]. Considering the collinearity, the contributions of different environmental variables to each process were assessed by OPLS[100] using the function 'opls' in the R package 'ropls' (v.1.30.0). The significance test was modified to be based on the permutation of samples rather than pairwise values (similar to permutation in Mantel test). In the OPLS output, R^2_Y and R^2_X represent the percentage of Y (that is, relative importance of each assembly process in this study) and X (that is, environmental variables in this study) dispersion (that is, the sum of squares) explained by the model, respectively. Q^2_Y reflects the overall predictive performance of the model, calculated by cross validation; if Q^2_Y is not significant ($P_{Q2Y} > 0.05$), the model may have been overfitted.

Statistics and reproducibility

Statistical analyses were completed using the R software (v.4.2.2) and specific methods detailed above. Since the assumptions of parametric tests (for example, normality, equal variances and so on) were not valid in many analyses, we used bootstrapping, permutational test and Mantel test for comparisons and association analyses, and we performed cross validation to improve the reliability of the results as described above. Our previous study showed that null model-based approaches such as NST require sufficient biological replicates (for example, >6) to ensure statistical power[13]. Thus, the sample size in each stress level was set to ≥8 in all the four different options (Supplementary Notes). As mentioned above, we excluded 6 wells with too many missing data (>6 environmental variables not available) and used the data from 91 wells in this study.

Reporting summary

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

Data availability

Data are accessible from a KBase narrative (https://narrative.kbase.us/narrative/145709), the current static version of which is https://kbase.us/n/145709/13/. The 16S rRNA gene sequencing data are from our prior study[42] and are available in MG-RAST with accession code mgp8190. The taxonomy classifier trained on Silva SSU 138 is from QIIME2 (v.2021.2), available at https://data.qiime2.org/2021.2/common/silva-138-99-515-806-nb-classifier.qza. Source data are provided with this paper.

Code availability

All custom scripts and the latest version of the R package iCAMP are available from GitHub (https://github.com/DaliangNing/iCAMP1).

References

1. Zhou, J. & Ning, D. Stochastic community assembly: does it matter in microbial ecology? *Microbiol. Mol. Biol. Rev.* https://doi.org/10.1128/mmbr.00002-17 (2017).

2. Nemergut, D. R. et al. Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356 (2013).
3. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.* **10**, 497–506 (2012).
4. HilleRisLambers, J., Adler, P. B., Harpole, W. S., Levine, J. M. & Mayfield, M. M. Rethinking community assembly through the lens of coexistence theory. *Annu. Rev. Ecol. Evol. Syst.* **43**, 227–248 (2012).
5. Fargione, J., Brown, C. S. & Tilman, D. Community assembly and invasion: an experimental test of neutral versus niche processes. *Proc. Natl Acad. Sci. USA* **100**, 8916–8920 (2003).
6. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* Vol. 32 (Princeton Univ. Press, 2001).
7. Vellend, M. Conceptual synthesis in community ecology. *Q. Rev. Biol.* **85**, 183–206 (2010).
8. Tucker, C. M., Shoemaker, L. G., Davies, K. F., Nemergut, D. R. & Melbourne, B. A. Differentiating between niche and neutral assembly in metacommunities using null models of β-diversity. *Oikos* **125**, 778–789 (2016).
9. Zhou, J. et al. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *mBio* **6**, e02288-14 (2015).
10. Burns, A. R. et al. Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* **10**, 655–664 (2016).
11. Sloan, W. T. et al. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8**, 732–740 (2006).
12. Stegen, J. C. et al. Quantifying community assembly processes and identifying features that impose them. *ISME J.* **7**, 2069–2079 (2013).
13. Ning, D., Deng, Y., Tiedje, J. M. & Zhou, J. A general framework for quantitatively assessing ecological stochasticity. *Proc. Natl Acad. Sci. USA* **116**, 16892–16898 (2019).
14. Ning, D. et al. A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat. Commun.* **11**, 4717 (2020).
15. Beaton, E. D. et al. Local and regional diversity reveals dispersal limitation and drift as drivers for groundwater bacterial communities from a fractured granite formation. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2016.01933 (2016).
16. Danczak, R. E., Johnston, M. D., Kenah, C., Slattery, M. & Wilkins, M. J. Microbial community cohesion mediates community turnover in unperturbed aquifers. *mSystems* https://doi.org/10.1128/mSystems.00066-18 (2018).
17. Sheng, Y. Z. et al. Distinct assembly processes shape bacterial communities along unsaturated, groundwater fluctuated, and saturated zones. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2020.143303 (2021).
18. Zhou, J. et al. Stochasticity, succession, and environmental perturbations in a fluidic ecosystem. *Proc. Natl Acad. Sci. USA* **111**, E836–E845 (2014).
19. Steinberg, C. E. W. *Stress Ecology: Environmental Stress as Ecological Driving Force and Key Player in Evolution* (Springer, 2012).
20. Zhang, K. et al. Salinity is a key determinant for soil microbial communities in a desert ecosystem. *mSystems* **4**, e00225-18 (2019).
21. Danczak, R. E. et al. Ecological assembly processes are coordinated between bacterial and viral communities in fractured-shale ecosystems. *mSystems* https://doi.org/10.1128/msystems.00098-20 (2020).
22. Zhang, G. et al. Salinity controls soil microbial community structure and function in coastal estuarine wetlands. *Environ. Microbiol.* **23**, 1020–1037 (2021).
23. Rocca, J. D. et al. The Microbiome Stress Project: toward a global meta-analysis of environmental stressors and their effects on microbial communities. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2018.03272 (2019).
24. Chase, J. M. Drought mediates the importance of stochastic community assembly. *Proc. Natl Acad. Sci. USA* **104**, 17430–17434 (2007).
25. Sengupta, A. et al. Disturbance triggers nonlinear microbe–environment feedbacks. *Biogeosciences* **18**, 4773–4789 (2021).
26. Lee, S.-H., Sorensen, J. W., Grady, K. L., Tobin, T. C. & Shade, A. Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. *ISME J.* **11**, 1447–1459 (2017).
27. Weiher, E. et al. Advances, challenges and a developing synthesis of ecological community assembly theory. *Phil. Trans. R. Soc. Lond. B* **366**, 2403–2413 (2011).
28. Chase, J. M. Stochastic community assembly causes higher biodiversity in more productive environments. *Science* **328**, 1388–1391 (2010).
29. Harpole, W. S. & Suding, K. N. A test of the niche dimension hypothesis in an arid annual grassland. *Oecologia* **166**, 197–205 (2011).
30. Goberna, M., Navarro-Cano, J. A., Valiente-Banuet, A., Garcia, C. & Verdu, M. Abiotic stress tolerance and competition-related traits underlie phylogenetic clustering in soil bacterial communities. *Ecol. Lett.* **17**, 1191–1201 (2014).
31. Patel, V., Sharma, A., Lal, R., Al-Dhabi, N. A. & Madamwar, D. Response and resilience of soil microbial communities inhabiting in edible oil stress/contamination from industrial estates. *BMC Microbiol.* https://doi.org/10.1186/s12866-016-0669-8 (2016).
32. Lima-Morales, D. et al. Linking microbial community and catabolic gene structures during the adaptation of three contaminated soils under continuous long-term pollutant stress. *Appl. Environ. Microbiol.* **82**, 2227–2237 (2016).
33. Tolkkinen, M. et al. Multi-stressor impacts on fungal diversity and ecosystem functions in streams: natural vs. anthropogenic stress. *Ecology* **96**, 672–683 (2015).
34. Lee, K. W. K. et al. Biofilm development and enhanced stress resistance of a model, mixed-species community biofilm. *ISME J.* **8**, 894–907 (2014).
35. He, Q. et al. Temperature and microbial interactions drive the deterministic assembly processes in sediments of hot springs. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2021.145465 (2021).
36. Gao, Q. et al. The microbial network property as a bio-indicator of antibiotic transmission in the environment. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2020.143712 (2021).
37. Liu, H. et al. Deterministic process dominated belowground community assembly when suffering tomato bacterial wilt disease. *Agronomy* **12**, 1024 (2022).
38. Gao, C. et al. Fungal community assembly in drought-stressed sorghum shows stochasticity, selection, and universal ecological dynamics. *Nat. Commun.* https://doi.org/10.1038/s41467-019-13913-9 (2020).
39. Verbruggen, E., van der Heijden, M. G. A., Weedon, J. T., Kowalchuk, G. A. & Roling, W. F. M. Community assembly, species richness and nestedness of arbuscular mycorrhizal fungi in agricultural soils. *Mol. Ecol.* **21**, 2341–2353 (2012).
40. 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).

41. Marasco, R. et al. Rhizosheath microbial community assembly of sympatric desert speargrasses is independent of the plant host. *Microbiome* https://doi.org/10.1186/s40168-018-0597-y (2018).
42. Smith, M. B. et al. Natural bacterial communities serve as quantitative geochemical biosensors. *mBio* https://doi.org/10.1128/mBio.00326-15 (2015).
43. He, Z. et al. Microbial functional gene diversity predicts groundwater contamination and ecosystem functioning. *mBio* **9**, e02435-17 (2018).
44. Kelly, L., Ding, H. M., Huang, K. H., Osburne, M. S. & Chisholm, S. W. Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *ISME J.* **7**, 1827–1841 (2013).
45. Kuang, J. et al. Predicting taxonomic and functional structure of microbial communities in acid mine drainage. *ISME J.* **10**, 1527–1539 (2016).
46. Jiang, Y. M., Huang, H. Y., Tian, Y. R., Yu, X. & Li, X. K. Stochasticity versus determinism: microbial community assembly patterns under specific conditions in petrochemical activated sludge. *J. Hazard. Mater.* https://doi.org/10.1016/j.jhazmat.2020.124372 (2021).
47. Ács, É. et al. Trait-based community assembly of epiphytic diatoms in saline astatic ponds: a test of the stress-dominance hypothesis. *Sci. Rep.* **9**, 15749 (2019).
48. Stegen, J. C., Lin, X., Fredrickson, J. K. & Konopka, A. E. Estimating and mapping ecological processes influencing microbial community assembly. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2015.00370 (2015).
49. Bei, Q. et al. Extreme summers impact cropland and grassland soil microbiomes. *ISME J.* https://doi.org/10.1038/s41396-023-01470-5 (2023).
50. Yuan, J. et al. Microbial and abiotic factors of flooded soil that affect redox biodegradation of lindane. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2021.146606 (2021).
51. Tripathi, B. M. et al. Soil pH mediates the balance between stochastic and deterministic assembly of bacteria. *ISME J.* **12**, 1072–1083 (2018).
52. Huot, O. B. & Tamborindeguy, C. Drought stress affects *Solanum lycopersicum* susceptibility to *Bactericera cockerelli* colonization. *Entomol. Exp. Appl.* **165**, 70–82 (2017).
53. Luo, C. H., Lu, F., Shao, L. M. & He, P. J. Application of eco-compatible biochar in anaerobic digestion to relieve acid stress and promote the selective colonization of functional microbes. *Water Res.* **68**, 710–718 (2015).
54. Watson, D. B., Kostka, J. E., Fields, M. W. & Jardine, P. M. *The Oak Ridge Field Research Center Conceptual Model* (Natural and Accelerated Bioremediation Research Field Research Center, 2004).
55. Tian, R. et al. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* **8**, 51 (2020).
56. Kothari, A. et al. Native plasmid-encoded mercury resistance genes are functional and demonstrate natural transformation in environmental bacterial isolates. *mSystems* **4**, e00588-19 (2019).
57. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2017.02224 (2017).
58. Hernandez, D. J., David, A. S., Menges, E. S., Searcy, C. A. & Afkhami, M. E. Environmental stress destabilizes microbial networks. *ISME J.* **15**, 1722–1734 (2021).
59. Jiao, S., Zhang, B., Zhang, G., Chen, W. & Wei, G. Stochastic community assembly decreases soil fungal richness in arid ecosystems. *Mol. Ecol.* **30**, 4338–4348 (2021).
60. Zhou, J. et al. Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community. *mBio* **4**, e00584-12 (2013).
61. Yan, L. J. et al. Environmental selection shapes the formation of near-surface groundwater microbiomes. *Water Res.* https://doi.org/10.1016/j.watres.2019.115341 (2020).
62. Fillinger, L., Hug, K. & Griebler, C. Selection imposed by local environmental conditions drives differences in microbial community composition across geographically distinct groundwater aquifers. *FEMS Microbiol. Ecol.* https://doi.org/10.1093/femsec/fiz160 (2019).
63. Stegen, J. C. et al. Groundwater–surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover. *Nat. Commun.* https://doi.org/10.1038/ncomms11237 (2016).
64. Ge, X. et al. Iron- and aluminium-induced depletion of molybdenum in acidic environments impedes the nitrogen cycle. *Environ. Microbiol.* **21**, 152–163 (2019).
65. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl Acad. Sci. USA* **103**, 626–631 (2006).
66. Sauve, S., Hendershot, W. & Allen, H. E. Solid-solution partitioning of metals in contaminated soils: dependence on pH, total metal burden, and organic matter. *Environ. Sci. Technol.* **34**, 1125–1131 (2000).
67. Chen, X. B., Wright, J. V., Conca, J. L. & Peurrung, L. M. Effects of pH on heavy metal sorption on mineral apatite. *Environ. Sci. Technol.* **31**, 624–631 (1997).
68. Demtröder, L., Narberhaus, F. & Masepohl, B. Coordinated regulation of nitrogen fixation and molybdate transport by molybdenum. *Mol. Microbiol.* **111**, 17–30 (2019).
69. Wen, X. et al. Characterization of vegetable nitrogen uptake and soil nitrogen transformation in response to continuous molybdenum application. *J. Plant. Nutr. Soil Sci.* **181**, 516–527 (2018).
70. Stanton, D. E., Batterman, S. A., Von Fischer, J. C. & Hedin, L. O. Rapid nitrogen fixation by canopy microbiome in tropical forest determined by both phosphorus and molybdenum. *Ecology* **100**, e02795 (2019).
71. Rousk, K., Degboe, J., Michelsen, A., Bradley, R. & Bellenger, J.-P. Molybdenum and phosphorus limitation of moss-associated nitrogen fixation in boreal ecosystems. *New Phytol.* **214**, 97–107 (2017).
72. Glass, J., Axler, R., Chandra, S. & Goldman, C. Molybdenum limitation of microbial nitrogen assimilation in aquatic ecosystems and pure cultures. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2012.00331 (2012).
73. Cvetkovic, A. et al. Microbial metalloproteomes are largely uncharacterized. *Nature* **466**, 779–782 (2010).
74. Barras, F. & Fontecave, M. Cobalt stress in *Escherichia coli* and *Salmonella enterica*: molecular bases for toxicity and resistance. *Metallomics* **3**, 1130–1134 (2011).
75. Paulo, L. M., Ramiro-Garcia, J., van Mourik, S., Stams, A. J. M. & Sousa, D. Z. Effect of nickel and cobalt on methanogenic enrichment cultures and role of biogenic sulfide in metal toxicity attenuation. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2017.01341 (2017).
76. Behera, M., Dandapat, J. & Rath, C. C. Effect of heavy metals on growth response and antioxidant defense protection in *Bacillus cereus*. *J. Basic Microbiol.* **54**, 1201–1209 (2014).
77. De, J., Ramaiah, N. & Vardanyan, L. Detoxification of toxic heavy metals by marine bacteria highly resistant to mercury. *Mar. Biotechnol.* **10**, 471–477 (2008).
78. Nies, D. H. Efflux-mediated heavy metal resistance in prokaryotes. *FEMS Microbiol. Rev.* **27**, 313–339 (2003).
79. Fodelianakis, S. et al. Dispersal homogenizes communities via immigration even at low rates in a simplified synthetic bacterial metacommunity. *Nat. Commun.* https://doi.org/10.1038/s41467-019-09306-7 (2019).

80. Muller, A. L. et al. Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. *ISME J.* **8**, 1153–1165 (2014).
81. Lui, L. M. et al. Mechanism across scales: a holistic modeling framework integrating laboratory and field studies for microbial ecology. *Front. Microbiol*. https://doi.org/10.3389/fmicb.2021.642422 (2021).
82. Vellend, M. et al. Assessing the relative importance of neutral stochasticity in ecological communities. *Oikos* **123**, 1420–1430 (2014).
83. Lowe, W. H. & McPeek, M. A. Is dispersal neutral? *Trends Ecol. Evol.* **29**, 444–450 (2014).
84. Wu, W.-M. et al. In situ bioreduction of uranium(VI) to submicromolar levels and reoxidation by dissolved oxygen. *Environ. Sci. Technol.* **41**, 5716–5723 (2007).
85. Thorgersen, M. P. et al. Molybdenum availability is key to nitrate removal in contaminated-groundwater environments. *Appl. Environ. Microbiol.* **81**, 4976–4983 (2015).
86. Akima, H. & Gebhardt, A. akima: Interpolation of Irregularly and Regularly Spaced Data v.0.5-12 (CRAN, 2015).
87. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
88. Anderson, M. J. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**, 245–253 (2006).
89. Oksanen, J. et al. vegan: Community Ecology Package v.2.5-7 (CRAN, 2020).
90. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
91. Caporaso, J. G. et al. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).
92. Bokulich, N. A. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).
93. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
94. Chiu, C.-H., Wang, Y.-T., Walther, B. A. & Chao, A. An improved non-parametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics* **70**, 671–682 (2014).
95. Stegen, J. C., Lin, X., Konopka, A. E. & Fredrickson, J. K. Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J.* **6**, 1653–1664 (2012).
96. Wang, J. et al. Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J.* **7**, 1310–1321 (2013).
97. Legendre, P. & Anderson, M. J. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* **69**, 1–24 (1999).
98. Borcard, D. & Legendre, P. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Modell.* **153**, 51–68 (2002).
99. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
100. Trygg, J. & Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **16**, 119–128 (2002).

Acknowledgements

This study by ENIGMA (Ecosystems and Networks Integrated with Genes and Molecular Assemblies; http://enigma.lbl.gov), a Science Focus Area Program at Lawrence Berkeley National Laboratory, is based on work supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research under contract number DE-AC02-05CH11231 (to P.D.A., A.P.A., J.Z., T.C.H., E.J.A., M.W.F., M.W.W.A., R.C. and D.A.S.). The development of the theoretical framework and statistical methods was partly supported by NSF Grants EF-2025558 and DEB-2129235 to J.Z. and D.N. The study was also supported by the Office of the Vice President for Research at the University of Oklahoma (to J.Z.).

Author contributions

All authors contributed intellectual input and assistance to this study and the paper preparation. J.Z. conceived the research questions. J.Z. and D.N. developed the theoretical framework. E.J.A., T.C.H., M.W.F., M.W.W.A., R.C., Z.H., J.Z., P.D.A. and A.P.A. designed and organized the experiment. Y.W., Y.F., J.W., J.D.V.N., L.W., P.Z., D.J.C., R.T., L.L. and F.P. generated or collected the data. D.N. integrated the data and performed statistical analyses with the assistance of Y.W.; J.Z. and D.N. wrote the paper with inputs from D.A.S., Z.H., J.W., L.L. and D.J.C.

Competing interests

The authors declare no competing interests.

Additional information

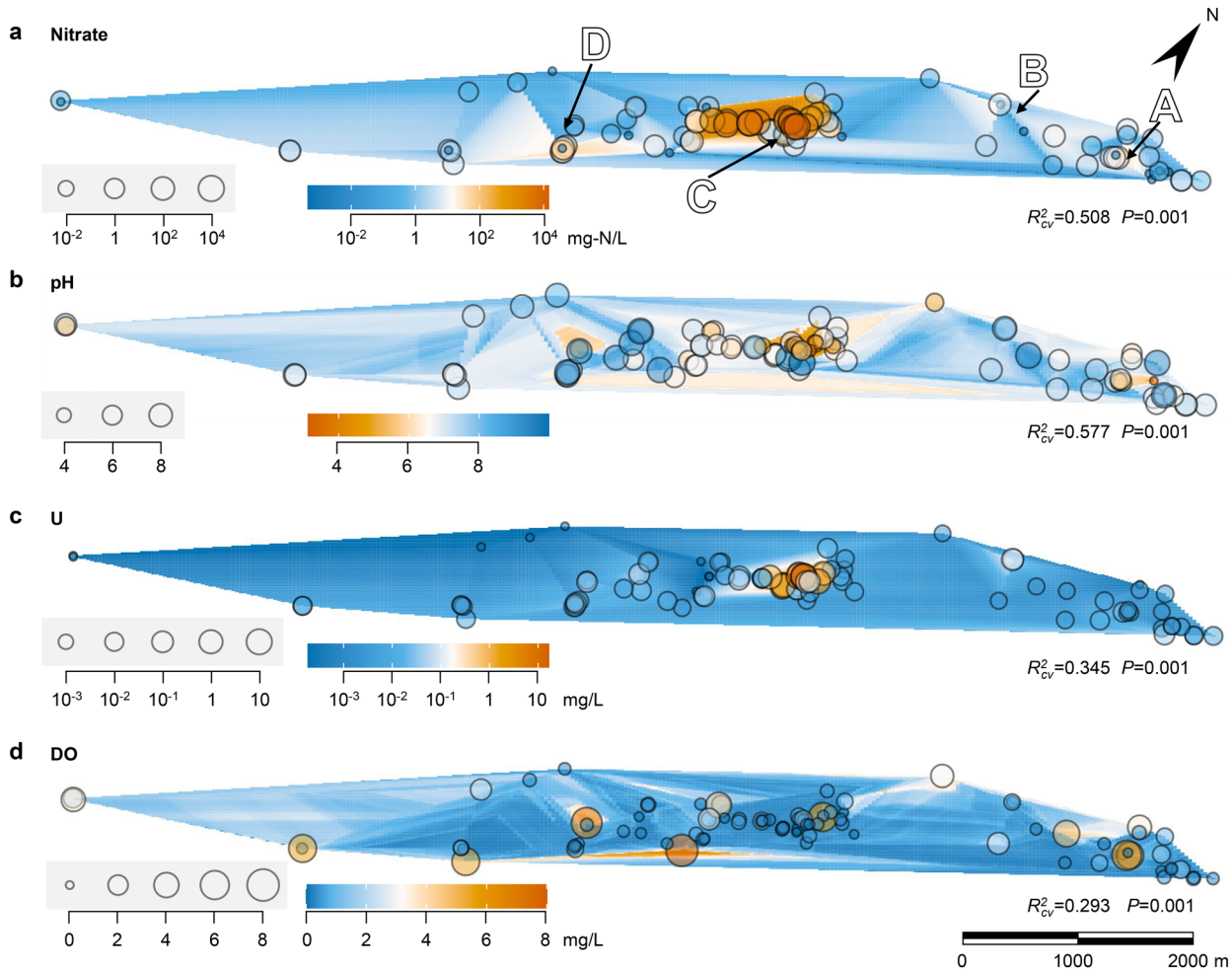
Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01573-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01573-x.

Correspondence and requests for materials should be addressed to Jizhong Zhou.

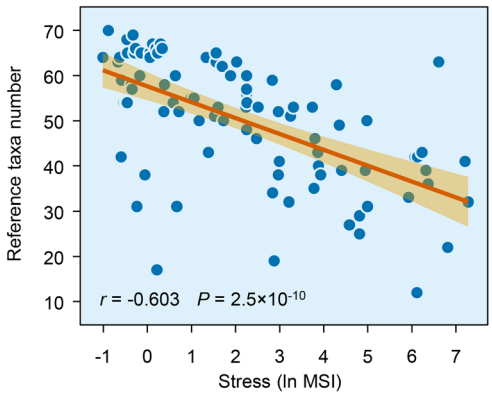
Peer review information *Nature Microbiology* thanks Malak Tfaily, Michael Wilkins and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-023-01573-x



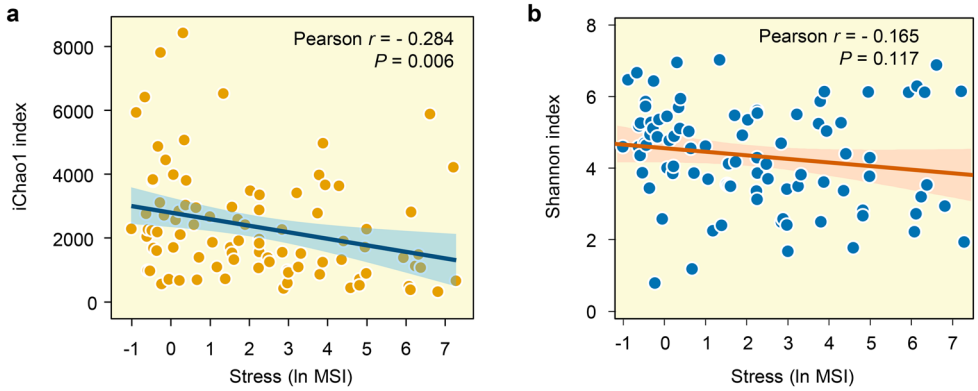
Extended Data Fig. 1 | Maps of the concentrations of representative environmental variables in groundwater. a, Nitrate-nitrogen. **b**, pH. **c**, Uranium (U). **d**, Dissolved oxygen (DO). The bubble size represents the value at each sampling position; R^2_{CV} is cross-validated R^2 of the model used to generate each map; *P* values (one-sided) are based on permutational test; the uppercase A - D indicate hotspots of heterogeneous selection, the same as in Fig. 2.

https://doi.org/10.1038/s41564-023-01573-x



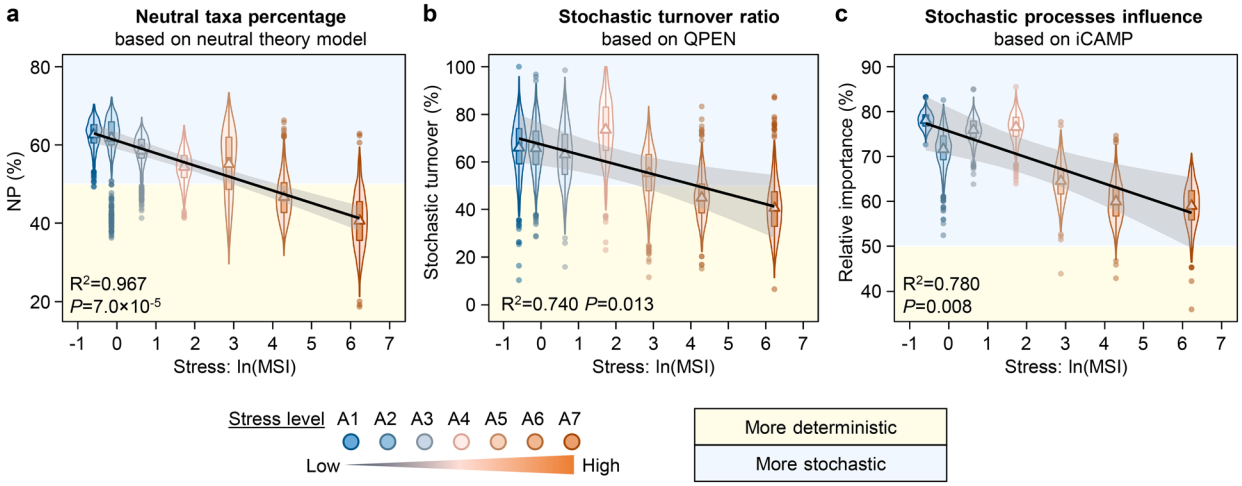
Extended Data Fig. 2 | Correlation between reference taxa number and log-transformed maximum stress index (ln MSI). Each data point represents a sample. The line and shadow are the trendline and 95% confidence interval based on linear regression. *r* and *P* show Pearson correlation coefficient and significance (two-sided). The 'reference taxa' are defined as relatively abundant and common taxa in uncontaminated samples.

https://doi.org/10.1038/s41564-023-01573-x



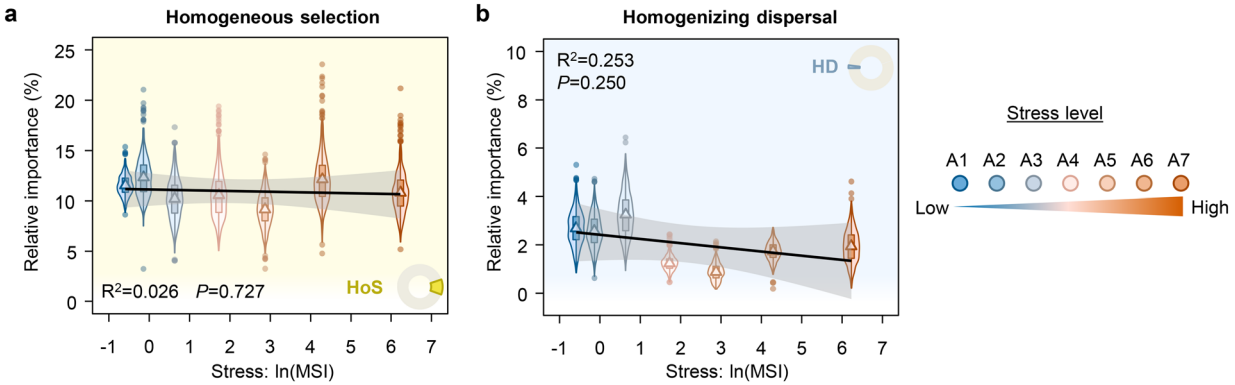
Extended Data Fig. 3 | Variation of bacterial alpha diversity along the stress gradient. a, Richness estimated by iChao1 index; **b**, Shannon index. The line and shadow show the trendline and 95% confidence interval based on linear regression. *r* and *P* are Pearson correlation coefficient and significance (two-sided). The significant Pearson correlation indicates a general decrease in bacterial richness as stress increased in the groundwater.

https://doi.org/10.1038/s41564-023-01573-x



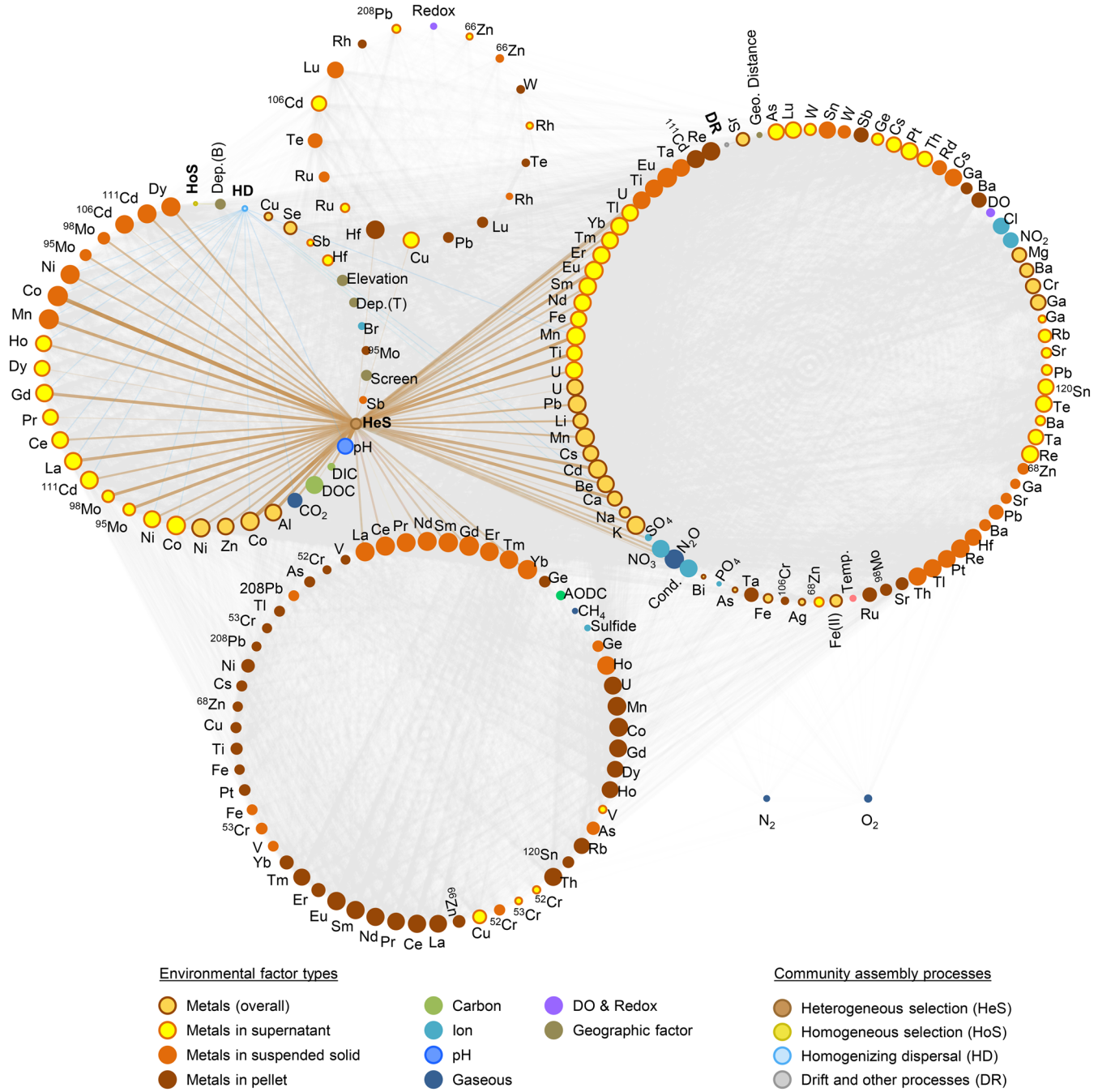
Extended Data Fig. 4 | Determinism and stochasticity of groundwater bacterial assembly at different stress levels. a, Abundance-weighted percentage of neutral taxa (NP) based on neutral theory model. **b**, Stochastic turnover ratio, that is, percentage of community turnovers governed by stochastic assembly processes based on a framework of entire-community null model analysis (QPEN). **c**, Relative importance of stochastic assembly processes based on a framework of phylogenetic-bin null model analysis (iCAMP). MSI, maximum stress index. The violin and box plots are based on bootstrapping results at each stress level (n = 13 in each level; bootstrapping 1000 times). Colors of violin plots indicate the stress levels. In box plots, center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; dots, outliers; triangle, mean value at each stress level. Black line, gray shadow, R^2, and *P* are the trendline, 95% confidence interval, coefficient of determination, and significance based on linear regression of the mean values as a function of log-transformed MSI.

https://doi.org/10.1038/s41564-023-01573-x



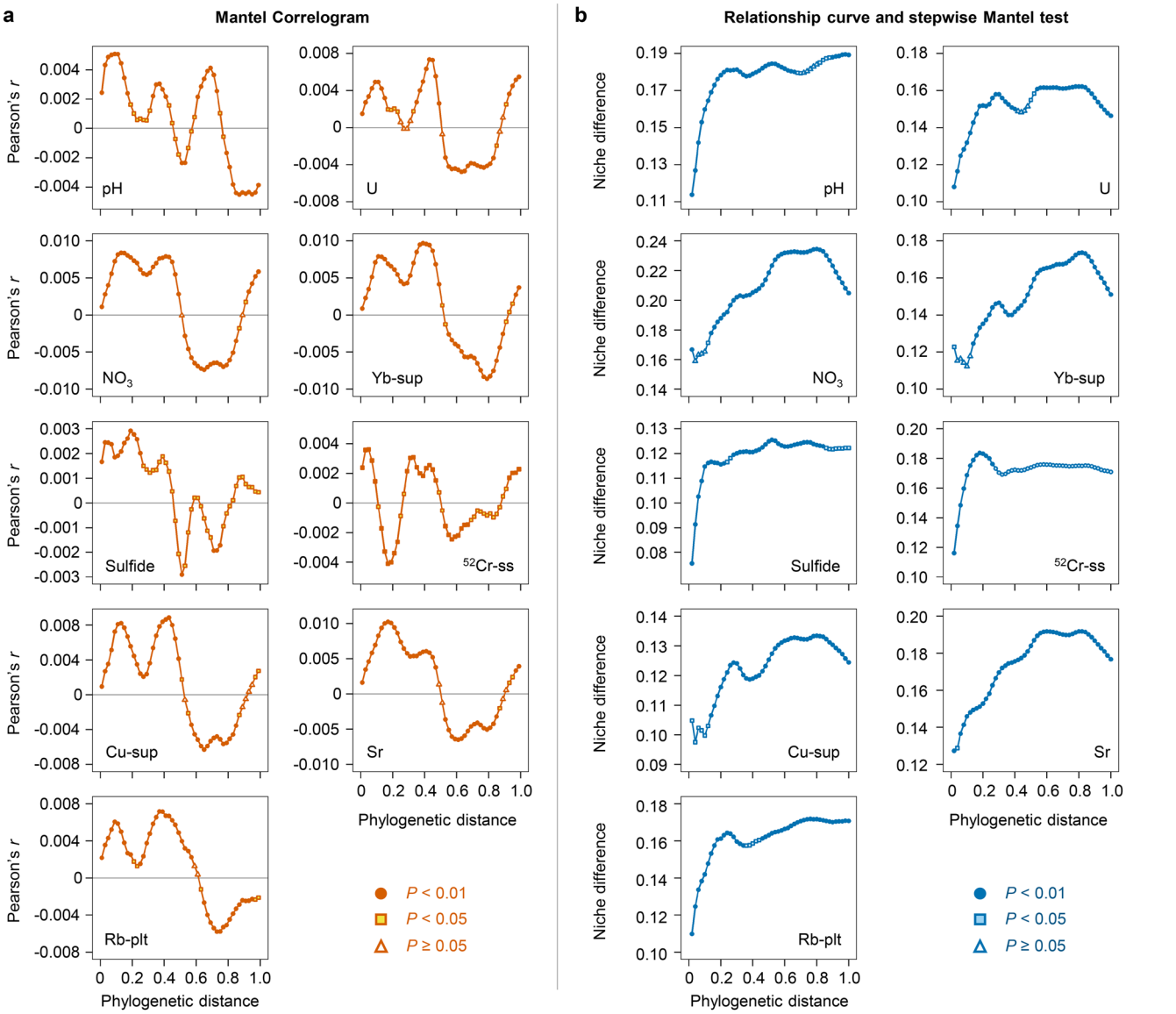
Extended Data Fig. 5 | Variation of groundwater bacterial assembly processes along the stress gradient. a, Homogeneous selection (HoS). **b**, Homogenizing dispersal (HD). The violin and box plots are based on bootstrapping results at each stress level (n = 13 in each level; bootstrapping 1000 times). Colors of violin/box plots indicate the stress levels. In box plots, center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; dots, outliers; triangle, mean value at each stress level. Black line, gray shadow, R^2, and *P* values are the trendline, 95% confidence interval, coefficient of determination, and significance based on linear regression of the mean values as a function of log-transformed MSI.

https://doi.org/10.1038/s41564-023-01573-x



Extended Data Fig. 6 | Association network of environmental variables and relative importance of community assembly processes. Links show significant correlation (cross-validated $R^2_{CV} > 0.01$ and $P < 0.05$); P values (one-sided) were calculated based on permutational test (1000 times) for correlations among environmental factors and Mantel test (permutated 1000 times) for correlations between assembly processes and environmental factors, then, adjusted by false discovery rate (FDR) method; line width reflecting the R^2_{CV}. Node area is related to its degree (associated node number). Nodes are grouped by modules based on greedy optimization. Dep. (T) and Dep. (B), depth of the screen top and bottom, respectively; geo.distance, geographic distance; cond., conductivity. Heterogeneous selection (HeS) showed much more strong associations with environmental variables and mainly with pH and metals in water phase (supernatant and suspended solid). The numerous gray links demonstrated the high density of strong associations among environmental variables.

https://doi.org/10.1038/s41564-023-01573-x



Extended Data Fig. 7 | Relationship between bacterial phylogenetic distance and niche preference difference (phylogenetic signal). a, Mantel correlogram. **b**, Relationship curve and stepwise Mantel test. Mantel correlogram was performed as previously described. Stepwise Mantel test was used to evaluate the correlation between niche preference difference and phylogenetic distance within the phylogenetic distance from 0 to a certain value. On the relationship curves, each niche difference value is the mean value in each phylogenetic distance 'class' within a distance interval of 0.02. The niche preference difference was estimated based on nine representative environmental variables. In the nine variables, pH, U, and nitrate (NO_3) are major stressors in the contaminated area; Yb-sup, Sulfide, ^{52}Cr-ss, Cu-sup, Sr, and Rb-plt were identified as 'centroid' variables which showed the nearest distance to the centroids of six environmental variable clusters, to represent the variation of each cluster. The clusters were identified by hierarchical clustering based on pairwise correlation among environmental variables. The distances to centroid were calculated by multivariate homogeneity of groups dispersions using function 'betadisper' in R package 'vegan'. *P* values (one-sided) are based on Mantel test (permutated 1000 times) and adjusted by false discovery rate (FDR) method. The Mantel correlograms and stepwise Mantel tests showed generally significant ($P < 0.05$) phylogenetic signal, validating the use of phylogenetic diversity to infer the influence of environmental selection. However, the trends in the relationship curves and the change of correlation coefficients and P values revealed the tipping points of phylogenetic signal at short phylogenetic distances around 0.2 to 0.6, supporting the necessity to use the phylogenetic-bin-based null model approach (iCAMP) which better exploits phylogenetic signal within relatively short phylogenetic distance.

nature portfolio

Corresponding author(s): Jizhong Zhou

Last updated by author(s): Nov 20, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing raw data were collected by MiSeq Platform (Illumina, San Diego, CA, USA). Other physical and chemical data were collected using corresponding equipment detailed in Method part and stored using Excel in Microsoft 365.
Data analysis	Sequencing data were analyzed by USEARCH (version 11.0.667), PyNAST (version 1.2.2), FastTree (version 2.1.11), QIIME2 (version 2021.2), sklearn-based taxonomy classifier (version 0.23.1), and the custom scripts at https://github.com/spacocha/16S_pre-processing_scripts/. All statistical analyses were implemented by R (version 4.2.2) and packages 'iCAMP' (version 1.6.1), 'NST' (version 3.1.10), 'vegan' (version 2.5-7), 'akima'(version 0.5-1.2), 'bigmemory' (version 4.6.1), and 'ropls' (version 1.30.0), as detailed in the method section. All custom scripts and the latest version of R package 'iCAMP' are available from GitHub (https://github.com/DaliangNing/iCAMP1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data are accessible from a KBase narrative (https://narrative.kbase.us/narrative/145709) of which the current static version is https://kbase.us/n/145709/10/. The 16S rRNA sequencing data are from the previously published paper (Smith et al. mBio 2015. doi:10.1128/mBio.00326-15) and available in MG-RAST with accession code mgp8190 (http://metagenomics.anl.gov/mgmain.html?mgpage=project&project=mgp8190). The taxonomy classifier trained on Silva SSU 138 is from QIIME 2 (version 2021.2), available at https://data.qiime2.org/2021.2/common/silva-138-99-515-806-nb-classifier.qza.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study is to develop a theoretical framework to conceptualize the relationships between community assembly processes and environmental stresses, and test the framework with groundwater microbial communities in a site under extremely wide ranges of stress conditions.
Research sample	Groundwater samples were collected from 97 representative wells, and 91 wells with more completed data were used in this study. The sampling well selection was based on a careful design to reflect the large geochemical gradient across the site.
Sampling strategy	All groundwater and filtered-groundwater samples were collected from the mid-screen level and analyzed to determine geochemistry and to characterize the microbial community structure. Prior to collection of samples, groundwater was pumped until pH, conductivity, and oxidation-reduction (redox) values were stabilized. This was done to purge the well and the line of standing water. According to recommendation in standard methods or our experience in analyzing samples from this site, approximately 2 to 20 liters of groundwater was purged from each well. For all wells, water was collected with either a peristaltic or a bladder pump using low flow in order to minimize draw down in the well. More details are reported in Smith et al (mBio 2015, doi:10.1128/mBio.00326-15).
Data collection	The ENIGMA team (Ecosystems and Networks Integrated with Genes and Molecular Assemblies, http://enigma.lbl.gov) collaborated together to collect the physical, chemical, and microbiome data, as detailed in the published paper (Smith et al. mBio 2015, doi:10.1128/mBio.00326-15)
Timing and spatial scale	For the 91 samples used in this study, only one sample was taken from each well. The first well was sampled on Nov. 13, 2012, and the last well was sampled on Feb. 21, 2013. All wells are from the same site, and the largest distance is around 10 km.

Data exclusions	This study requires relatively completed dataset to explore the environmental effect. Therefore, in the 97 wells sampled, we excluded 6 wells with too many missing data (>6 variables not available).
Reproducibility	Four different ways were tested to define stress levels, in which each stress level has at least 13, 9, 11, 8 samples, respectively. Bootstrapping test and cross validation are applied to improve the reliability of the conclusions.
Randomization	Since the sampling location selection is based on representativeness of geochemical properties rather than location (Smith et al. mBio 2015, doi:10.1128/mBio.00326-15), the distribution of the wells is relatively random. Besides, the coordinations or distance, as well as other geographic factors, are included in correlation and regression analysis. In addition, all samples were randomly reordered before measurements as possible, in order to rule out as many biases by the following sampling handling, e.g., DNA extraction, PCR amplification and sequencing.
Blinding	Samples were in random order during processing. People who processed the samples to determine 16S rRNA sequences and physical chemical properties did not know the actual meaning of the sample IDs.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The Department of Energy's (DOE) Oak Ridge Field Research Center (FRC) consists of 243 acres of contaminated area and 402 acres of an uncontaminated background area used for comparison located within the Bear Creek Valley watershed in Oak Ridge, Tennessee. The mean annual temperature is around 14.9 C, and the mean annual precipitation is around 1360 mm. Contamination at this site includes radionuclides (e.g., uranium and technetium), nitrate, sulfide, and volatile organic compounds. The main source of contamination has been traced back to the former S-3 waste disposal ponds located within the Y-12 national security complex. During the Cold War era, these unlined ponds were the primary accumulation site for organic solvents, nitric acid, and radionuclides generated from nuclear weapon development and processing. In 1988, the S-3 ponds were closed and capped; however, contaminants from these ponds leached out, creating a groundwater contaminant plume across the field site. These source plumes are continuously monitored and have been the subject of a number of studies over the years. Further information regarding the plume and sources of contamination can be found at http://www.esd.ornl.gov/orifrc/.
Location	The site is located at Oak Ridge, Tennessee, USA. The sampled wells are at latitude from 35.941 to 35.995 N, longitude from 84.235 to 84.336 W.
Access & import/export	Sampling was governed by the Y-12 Groundwater Protection Program Groundwater and Surface Water Sampling and Analysis Plan (Y/SUB/17-104804/2). This plan provides a description of the program's water quality monitoring activities planned at the U.S. Department of Energy (DOE), Y-12 National Security Complex (Y-12). The monitoring was managed by the Y-12 Groundwater Protection Program to meet requirements of U.S. DOE Orders 436.1 and 458.1. Radiological sample handling is governed by Oak Ridge National Laboratory Research Safety Summary No 10967 , RSS No 15251, RSS No 8685 and also 10CFR835 appendix D AND DOE Radiological Control Standard (DOE Standard 1098-2008).
Disturbance	All sampling wells have already been available before this study. We estimated the minimal volume of each sample we needed and performed one-time-point sampling from each well to minimize the disturbance.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging