Stable states and keystone species in soil microbiomes of agroecosystems

Hiroaki Fujita[1], Shigenobu Yoshida[2], Kenta Suzuki[3] and Hirokazu Toju[1†]

[1]Center for Ecological Research, Kyoto University, Otsu, Shiga 520-2133, Japan

[2]Institute for Plant Protection, National Agriculture and Food Research Organization, Tsukuba, Ibaraki 305-8666, Japan

[3]Integrated Bioresource Information Division, BioResource Research Center, RIKEN, Tsukuba, Ibaraki 305-0074, Japan

[†]**Correspondence:** Hirokazu Toju (toju.hirokazu.4c@kyoto-u.ac.jp).

This article includes 6 Figures, 7 Extended Data Figures, 2 Supplementary Tables and 1 Source Data.

Abstract

Archaea, bacteria, and fungi in the soil are increasingly recognized as major determinants of agricultural productivity and sustainability. Although our knowledge of soil biodiversity has been greatly enhanced for the last decade, defining key characteristics of stable microbiome structure and their ecosystem functions remains a major challenge. We here show that "stable states" of soil microbiomes can be defined with their disease-suppressive effects on crop plants. We integrated agricultural field metadata with microbial community analyses by targeting > 2,000 agricultural soil samples collected across the Japanese Archipelago, profiling microbiomes constituted, in total, by 579 archaeal, 26,640 bacterial, and 6,306 fungal species/strains. A DNA-metabarcoding-based evaluation of balance between prokaryote and fungal abundance indicated that fungus-dominated soil is more resistant to crop plant disease. A statistical physics approach for reconstructing the "energy landscape" of community structure indicated that disease-suppressive and disease-susceptible soils could be distinguished as alternative stable states of fungal communities. A network theoretical analysis further indicated that the entire microbiomes were partitioned into several substructures differing in potential impacts on crop disease level. We then highlighted potential keystone prokaryotes and fungi governing ecological processes within/across such substructures. These results suggest that agroecosystem performance and sustainability can be managed by identifying key structure and species within microbiomes.

The ongoing global-scale degradation of agroecosystems is threatening food production[1–3]. Maximizing the functions of microbial communities is a prerequisite for building bases of sustainable agriculture[4–9]. Archaea, bacteria, and fungi not only promote soil nutrient cycles but also protect crop plants from pests/pathogens and abiotic stresses[10–14]. Therefore, diagnosing soil microbiome states is of central interest when we try to design and manage resource-efficient and disease-tolerant agroecosystems.

Despite the promise that informatics of microbiomes renovates agricultural practices, defining favorable and unfavorable states of microbial communities[4,15,16] is notoriously difficult due to tremendous biodiversity in soil ecosystems. Microbiome compositions in the wild are consequences of multiple ecological processes[17–22]. Filtering by environmental conditions, for example, is a primary mechanism determining establishment of microbial populations[21,23,24]. Furthermore, competitive and facilitative interactions between microbial species are essential factors governing local community structure[25–27]. Continually influenced

52 by these mechanisms and factors, local microbial communities differ in their stability and
53 functions[16,28–30]. Consequently, mapping "alternative stable states[17,31–33]" within immensely
54 complex state space of microbiome structure is an essential step for managing agroecosystem
55 functions in light of possible community-level consequences with benign/adverse effects.

56 We here perform systematic profiling of cropland microbiome structure across the
57 Japanese Archipelago in order to define general characteristics of disease-suppressive and
58 disease-susceptible soil based on theoretical ecology. We integrated DNA metabarcoding
59 results of prokaryote and fungal communities with environmental meta-datasets representing
60 > 2,000 cropland soil samples of 19 plant species. With the bird's-eye view of possible
61 microbiome structure, we explore backbone characteristics associated with crop disease
62 conditions. Specifically, we reconstruct the "energy landscape" of microbial community
63 structure, drastically reducing dimensions for diagnosing disease-suppressive and disease-
64 susceptible states of agroecosystems. With a network theoretical approach, we further reveal
65 how soil microbiomes are partitioned into substructures potentially differing in agricultural
66 performance. Within the reconstructed microbe-to-microbe network, we identify candidates
67 of keystone species that potentially trigger switches between benign and adverse
68 agroecosystem states. Overall, this study provides frameworks and criteria for diagnosing
69 agroecosystem functions based on microbiome structure.

70

71 # Results

72 **Diversity of microbiome structure**

73 We compiled the field metadata of 2,903 soil samples collected in research projects of
74 National Agricultural and Food Research Organization (NARO). The rhizosphere/bulk soils
75 of farmlands were sampled from subtropical to cool-temperate regions (26.1–42.8 ºN) across
76 the Japanese Archipelago, targeting 19 crop plant species (Fig. 1a) from 2006 to 2014. Most
77 of the farmlands were managed with conventional agricultural practices (characterized by
78 intensive tillage and chemical fertilizer/pesticide application), while some were
79 experimentally controlled as organic agricultural fields (Source Data S1). The metadata
80 included the information of chemical [e.g., pH, electrical conductivity, carbon/nitrogen (C/N)
81 ratio, and available phosphorous concentration], physical (e.g., soil taxonomy), and biological
82 (e.g., crop disease level) properties, providing an invaluable platform for profiling and
83 categorizing ecosystem states of cropland soil.

84 To integrate the metadata with the information of microbial community structure, we
85 performed DNA metabarcoding analyses of both prokaryotes (archaea and bacteria) and
86 fungi. After a series of quality filtering, prokaryote and fungal community data were
87 successfully obtained from 2,534 and 2,384 samples, respectively. In total, 579 archaeal
88 amplicon sequence variants (ASVs) representing 14 genera (11families), 26,640 bacterial
89 ASVs representing 875 genera (332 families), and 6,306 fungal archaeal ASVs representing
90 573 genera (240 families) were detected (Fig. 1b; Extended Data Fig. 1).

91 The prokaryote communities did not have dominant taxa at the genus and family levels,
92 although the archaeal family Nitrososphaeraceae had the highest abundance at some research
93 sites (Fig. 1b). In contrast, the fungal communities were dominated by fungi in the families
94 Mortierellaceae, Chetomiaceae, and Nectriaceae (Fig. 1b). A reference database profiling of
95 fungal functional groups suggested that the fungal communities were dominated by plant
96 pathogenic and soil saprotrophic fungi (Fig. 1c) as characterized by the dominance of
97 *Fusarium* and *Mortierella* at the genus level (Extended Data Fig. 1). Meanwhile,
98 mycoparasitic fungi had exceptionally high proportions at some research sites, as represented
99 by the dominance of *Trichoderma* (Hypocreaceae) at those sites (Fig. 1b; Extended Data Fig.
100 1).

101 The prokaryote and fungal communities differed in basic biodiversity patterns. In terms
102 of α-diversity at the family level, prokaryote communities showed higher Shannon's entropy
103 than fungal communities in all but one of the research sites (Fig. 1d). In contrast, variability
104 of community structure among samples, i.e., Bray-Curtis β-diversity, was higher in fungi than
105 in prokaryotes in most study sites (Fig. 1e). Likewise, fungi showed higher levels of
106 community structural difference between research sites than prokaryotes (Welch's test; t = -
107 6.6, df = 168.9, $P < 0.0001$; Fig. 1f). These results indicate that diverse taxonomic groups of
108 archaea and bacteria are present without dominance by particular taxa, while fungal
109 community structure is spatially dynamic both within and across agroecosystems.

110 Compiling the metadata of edaphic factors, we found that variation in the community
111 structure of prokaryotes and fungi was significantly explained by crop plant identity and soil
112 taxonomy as well as by soil chemical properties such as pH, electrical conductivity, C/N ratio,
113 and available phosphorous concentration (Fig. 2a-b; Extended Data Fig. 2-3; Supplementary
114 Information). Meanwhile, both prokaryote and fungal community structure was significantly
115 associated with the severity of crop disease (Fig. 2c-d). Thus, the dataset as a whole offers an

opportunity for understanding mechanisms spanning physical, chemical, and biological processes in agroecosystems.

Crop disease and microbiome structure

We then focused on factors determining disease-suppressive and disease-susceptible properties of agroecosystem soil. In addition to chemical environmental factors (e.g., soil pH and electrical conductivity), whose associations with agricultural productivity have been routinely examined[34,35], we explored biological factors critical to crop-plant disease level. The α-diversity of prokaryotes and/or fungi, for example, is an index that can be estimated with DNA-based profiling of soil biological communities. In further reconstructing biological factors in a high-throughput manner, we focused on relative abundance of prokaryotes and fungi. As the classic chromatography-based method for quantifying phospholipid fatty acids specific to bacteria and fungi[36] could require impractically long time for processing > 2,000 samples, we developed an alternative approach for estimating the ratio of prokaryote abundance to fungal abundance (hereafter, prokaryote/fungus ratio) based on DNA metabarcoding. Specifically, the relative concentrations of prokaryote 16S rRNA gene sequences to fungal ITS gene sequences were calculated based on a quantitative amplicon sequencing analysis (Fig. 3a; see Methods for details). We then constructed statistical models comparing effects on crop-plant disease level among various types of chemical/biological factors.

A series of analysis indicated prokaryote/fungus ratio had strong impacts on crop-plant disease level. While soil chemistry (soil pH) had significant effects on the proportion of diseased crop plants, the prokaryote/fungus ratio showed even stronger associations with crop plant disease level (Model 1 in Fig. 3b). Specifically, crop plant disease was more likely to occur when prokaryotes were abundant relative to fungi (Fig. 3c). The effects of prokaryote/fungus ratio on plant disease level were conspicuous even when they were compared with effects of prokaryote and fungal α-diversity (Model 2 in Fig. 3b). The threshold values over which more than 50 % of plants showed disease levels 2–5 were estimated to be around 0.7 in a log-ratio unit (Fig. 3c), although the values potentially vary depending on various factors such as crop plants, geography, and climatic conditions.

Alternative stables states

148 We next reconstructed alternative stable states[17,31,33] of microbiomes in order to define
149 community structure of disease-suppressive and disease-susceptible soil. In an energy
150 landscape analysis of statistical physics[32], we modeled how microbial community
151 compositions (community states) depend on soil chemical factors and interactions between
152 microbial ASVs. The modeling approach allowed us to calculate the "energy" of each
153 community state. We then estimated the shape of an energy landscape, on which lower energy
154 represents more stable community state. Stable states were then defined as the community
155 compositions whose energy is lower than all possible adjacent community compositions[32].

156 The results of energy landscape reconstruction fundamentally differed between
157 prokaryotes and fungi. Even if community states were defined in a binary format at the family
158 or order level to reduce the complexity of the input data (i.e., 2^S possible community states,
159 where S denotes the number of families; see Methods), more than 10^3 stable states were
160 inferred to exist for the prokaryote dataset (Fig. 4). This result suggests that ecological
161 processes of soil prokaryote communities are described as shifts between many equally
162 possible compositions and that categorization of ecosystem functions based on prokaryote
163 community structure is challenging. In contrast, the energy landscape of fungal community
164 structure was represented by much smaller number of alternative stable states (Fig. 4). While
165 the estimated number of stables states varied depending on threshold values in the
166 binarization of community structural data, fungal communities consistently showed lower
167 number of stable states than prokaryote communities (Supplementary Information; Extended
168 Data Fig. 4).

169 We then filtered the (alternative) stable states of fungal communities based on the
170 "depth of basins" and the number of samples belonging to the stable states[32], obtaining 10
171 major stable states representing the energy landscape (Fig. 5a; see also Extended Data Fig. 5).
172 By integrating the information of energy landscape (Fig. 5b) with that of crop-plant disease
173 level (Fig. 5c), we ranked the stable states in terms of possible disease-suppressive functions
174 (Fig. 5d). We then found that all soil samples belonging to the top-three stable states had
175 associations with the lowest crop-disease level, while 92 % of samples belonging to the
176 lowest-rank stable state showed disease levels 2–5.

177 The 10 stable states were distinguished in light of balance among 22 fungal families
178 (Fig. 5c). Hypocreaceae, which embraced well-known mycoparasitic taxa (e.g.,
179 *Trichoderma*[37]), appeared at high proportions in stable states with potential disease-
180 suppressive functions. In contrast, Aspergillaceae, which included both plant pathogens and

181 saprotrophs[38], were relatively abundant in stable states with high disease potential. In addition
182 to these basic patterns, several families were keys to define disease-suppressive or disease-
183 susceptible soils. For example, high abundance of Didymellaceae, Lasiosphaeriaceae,
184 Mrakiaceae, Orbilliaceae, and Piptocephalidaceae characterized the two stable states with the
185 highest performance (Fig. 3d). Among the families, Orbiliaceae and Piptocephalidaceae were
186 known as nematophagous and mycoparasitic taxa, respectively[39,40], presumably working as
187 suppressors of plant-parasitic nematodes and fungi in the soil.

188

189 **Microbiome networks and keystones**

190 To understand potential mechanisms and factors driving shifts between disease-suppressive
191 and disease-susceptible microbiomes, we reconstructed the network architecture of microbe-
192 to-microbe associations. As the number of possible interactions inflates with that of examined
193 microbes (i.e., X^2 pairs, where X denotes the number of microbial species/ASVs), we targeted
194 microbial ASVs that appeared in ≥ 10 % of samples in the network statistical analysis. The
195 reconstructed network of coexistence was subdivided into several "modules", in which
196 archaeal, bacterial, and fungal ASVs sharing environmental preferences and/or those in
197 positive interactions were lined with each other (Fig. 6a). We found that the network modules
198 differed considerably in their associations with crop-plant disease level (Fig. 6a-b; Extended
199 Data Fig. 6). Modules 3 and 7, for example, were constituted by microbes with potentially
200 negative impacts on crop plants, while Modules 1, 2, 4, and 5 were characterized by microbes
201 positively associated with plant disease level (Fig. 6b).

202 Based on specificity to crop disease status (Extended Data Fig. 6) and topological
203 positions within the network (Extended Data Fig. 7), candidates of keystone microbes, which
204 impose profound impacts on ecosystem-level processes, were highlighted. We found that
205 some ASVs allied to a clade of ammonium-oxidizing archaea[41] (*Nitrocosmicus* and other
206 Nitrososphaeraceae ASVs) were associated with low disease level of plants and they were
207 placed at hub positions as indicated by the betweenness centrality index (Fig. 6c). Among
208 microbes associated with severely damaged plants, an *Chaetomium* fungus had the highest
209 network centrality (Fig. 6c). When the microbial ASVs were ranked in terms of potential roles
210 interlinking multiple modules, potentially ammonium-oxidizing archaea in the family
211 Nitrososphaeraceae[41] and nitrite-oxidizing bacteria in the genus *Nitrospira*[42] exhibited were
212 highlighted (Fig. 5d). Meanwhile, the above-mentioned *Chaetomium* fungus served as a hub
213 within Module 7 (Fig. 6d), which consisted of microbes associated with crop disease (Fig.

6b). These microbes occupying key topological positions are prioritized targets of research because control/manipulation of them can impose cascade effects throughout the network, triggering shifts between alternative stable states of microbiomes[26].

Discussion

By integrating large datasets of agricultural field properties and DNA metabarcoding, we found that seemingly diverse microbiome structure could be categorized into several stable states varying in ecosystem-level performance. Our analysis based on statistical physics suggested that prokaryote communities had numerous local compositional optima, presumably due to their rapid and continuous responses to changes in environmental conditions. In contrast, ecological processes of fungal communities could be interpreted as shifts between alternative stable states (Fig. 5). Basically, fungi can form complex hyphal structures and their individual lifespan is much longer than most archaea and bacteria. Given these biological properties, structure and functions of fungal communities are more likely to be influenced by legacy effects of past ecological processes[43,44] than those of prokaryotes. Consequently, agroecosystem functions can be managed by keeping fungal community structure at targeted stable states with positive impacts on crop production. From the perspective of pest/pathogen control, increasing the agroecosystem-level resistance to devastating nematodes and fungi is a pressing issue[45,46]. Our finding that specific taxa of nematophagous and mycoparasitic fungi are keys to distinguish disease-suppressive and disease-susceptible soils will help us reduce the use of chemical pesticides/fungicides by maximizing of ecosystem stability and functions.

The intensive profiling of microbiomes further allowed us to pinpoint potential keystone species/taxa, which could have disproportionate impacts on ecosystem-level processes through networks of microbe-to-microbe interactions. The fact that archaea and bacteria involved in ammonium/nitrite oxidization processes[41,42] were designated as candidate keystones (Fig. 6) is of particular interest. Control of these archaea/bacteria, for example, can alter the efficiency or pace of soil nitrogen cycles, resulting in increased/decreased N_2O emission from crop lands[47,48]. Thus, these potential keystone microbes deserve future prioritized research, in which we will explore the ground design of sustainable agroecosystems with efficient nutrient cycles, reduced greenhouse-gas emission, and enhanced defense against pests/pathogens.

While we herein profiled alternative stable states of microbiomes and potential keystone microbes by focusing mainly on conventional agricultural fields, soil microbiomes with higher functions and stability may be explored in the framework of conservation agriculture[49] or regenerative agriculture[50]. Minimum tillage[51], for example, can decrease damage to fungal hyphae, increasing plants' nutrient uptake through mycorrhizal fungi and suppressing pests/pathogens by the functions of diverse entomophagous, nematophagous, and mycoparasitic fungi. In this respect, the fact that crop disease decreased towards the range of dominance by fungi over prokaryotes (Fig. 3) is of particular interest. Given that the prokaryote/fungus ratio varied considerably among soil samples with similar community compositions (Fig. 3a), the byproduct index resulting from high-throughput DNA metabarcoding is expected to provide a diagnostic criterion of agroecosystem health in lines with stable-state analyses of microbiome structure.

Knowledge of ecosystem-level patterns and processes is indispensable for maximizing biodiversity functions for sustainable food production. Bird's-eye views of stable microbiome compositions and keystone species will provide platforms for integrating ecology, genomics, and microbiology towards the realization of "nature-positive" world[52].

References

1. Springmann, M. *et al.* Options for keeping the food system within environmental limits. *Nature* **562**, 519–525 (2018).
2. Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. Agricultural sustainability and intensive production practices. *Nature* **418**, 671–677 (2002).
3. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences* **108**, 20260–20264 (2011).
4. Toju, H. *et al.* Core microbiomes for sustainable agroecosystems. *Nat Plants* **4**, 247–257 (2018).
5. Busby, P. E. *et al.* Research priorities for harnessing plant microbiomes in sustainable agriculture. *PLoS Biol* **15**, e2001793 (2017).
6. Berendsen, R. L., Pieterse, C. M. J. & Bakker, P. A. H. M. The rhizosphere microbiome and plant health. *Trends Plant Sci* **17**, 478–486 (2012).
7. Bommarco, R., Kleijn, D. & Potts, S. G. Ecological intensification: Harnessing ecosystem services for food security. *Trends Ecol Evol* **28**, 230–238 (2013).

8. de Vries, F. T. & Wallenstein, M. D. Below-ground connections underlying above-ground food production: a framework for optimising ecological connections in the rhizosphere. *Journal of Ecology* **105**, 913–920 (2017).
9. Chaparro, J. M., Sheflin, A. M., Manter, D. K. & Vivanco, J. M. Manipulating the soil microbiome to increase soil health and plant fertility. *Biol Fertil Soils* **48**, 489–499 (2012).
10. Peay, K. G., Kennedy, P. G. & Talbot, J. M. Dimensions of biodiversity in the Earth mycobiome. *Nat Rev Microbiol* **14**, 434–447 (2016).
11. Baker, B. J. *et al.* Diversity, ecology and evolution of Archaea. *Nat Microbiol* **5**, 887–900 (2020).
12. Hayat, R., Ali, S., Amara, U., Khalid, R. & Ahmed, I. Soil beneficial bacteria and their role in plant growth promotion: A review. *Ann Microbiol* **60**, 579–598 (2010).
13. Newsham, K. K. A meta-analysis of plant responses to dark septate root endophytes. *New Phytol* **190**, 783–93 (2011).
14. van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
15. Schlatter, D., Kinkel, L., Thomashow, L., Weller, D. & Paulitz, T. Disease suppressive soils: New insights from the soil microbiome. *Phytopathology* **107**, 1284–1297 (2017).
16. Jansson, J. K. & Hofmockel, K. S. The soil microbiome — from metagenomics to metaphenomics. *Curr Opin Microbiol* **43**, 162–168 (2018).
17. Fukami, T. Historical contingency in community assembly: integrating niches, species pools, and priority effects. *Annu Rev Ecol Evol Syst* **46**, 1–23 (2015).
18. Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T. & Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. *Nat Rev Microbiol* **18**, 607–621 (2020).
19. Jiao, S., Yang, Y., Xu, Y., Zhang, J. & Lu, Y. Balance between community assembly processes mediates species coexistence in agricultural soil microbiomes across eastern China. *ISME Journal* **14**, 202–216 (2020).
20. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: Unraveling microbial community dynamics. *Curr Opin Microbiol* **25**, 56–66 (2015).
21. Tedersoo, L. *et al.* Global diversity and geography of soil fungi. *Science (1979)* **346**, 1256688 (2014).
22. Toju, H., Tanabe, A. S. & Sato, H. Network hubs in root-associated fungal metacommunities. *Microbiome* **6**, 116 (2018).

23. Talbot, J. M. *et al.* Endemism and functional convergence across the North American soil mycobiome. *Proceedings of the National Academy of Sciences* **111**, 6341–6346 (2014).
24. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
25. Faust, K. & Raes, J. Microbial interactions: From networks to models. *Nat Rev Microbiol* **10**, 538–550 (2012).
26. Toju, H., Yamamoto, S., Tanabe, A. S. A. S., Hayakawa, T. & Ishii, H. S. H. S. Network modules and hubs in plant-root fungal biomes. *J R Soc Interface* **13**, 20151097–20151097 (2016).
27. van der Heijden, M. G. A. & Hartmann, M. Networking in the plant microbiome. *PLoS Biol* **14**, 1–9 (2016).
28. Hartman, K. *et al.* Deciphering composition and function of the root microbiome of a legume plant. *Microbiome* **5**, 1–13 (2017).
29. Vorholt, J. A., Vogel, C., Carlström, C. I. & Müller, D. B. Establishing causality: opportunities of dynthetic vommunities for plant microbiome tesearch. *Cell Host Microbe* **22**, 142–155 (2017).
30. Mendes, L. W. *et al.* Soil-borne microbiome: linking diversity to function. *Microb Ecol* **70**, 255–265 (2015).
31. Beisner, B. E., Haydon, D. T. & Cuddington, K. Alternative stable states in ecology. *Front Ecol Environ* **1**, 376–382 (2003).
32. Suzuki, K., Nakaoka, S., Fukuda, S. & Masuya, H. Energy landscape analysis elucidates the multistability of ecological communities across environmental gradients. *Ecol Monogr* **91**, 1–21 (2021).
33. Scheffer, M., Carpenter, S., Foley, J. a, Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
34. Goulding, K. W. T. Soil acidification and the importance of liming agricultural soils with particular reference to the United Kingdom. *Soil Use Manag* **32**, 390–399 (2016).
35. Corwin, D. L. & Lesch, S. M. Apparent soil electrical conductivity measurements in agriculture. *Comput Electron Agric* **46**, 11–43 (2005).
36. Frostegård, A. & Bååth, E. The use of phospholipid fatty acid analysis to estimate bacterial and fungal biomass in soil. *Biol Fertil Soils* **22**, 59–65 (1996).
37. Poveda, J. Trichoderma as biocontrol agent against pests: New uses for a mycoparasite. *Biological Control* vol. 159 Preprint at https://doi.org/10.1016/j.biocontrol.2021.104634 (2021).

38. Tekaia, F. & Latgé, J. P. Aspergillus fumigatus: Saprophyte or pathogen? *Current Opinion in Microbiology* vol. 8 Preprint at https://doi.org/10.1016/j.mib.2005.06.017 (2005).

39. Yang, Y., Yang, E., An, Z. & Liu, X. Evolution of nematode-trapping cells of predatory fungi of the Orbiliaceae based on evidence from rRNA-encoding DNA and multiprotein sequences. *Proc Natl Acad Sci U S A* **104**, 8379–8384 (2007).

40. Lazarus, K. L., Benny, G. L., Ho, H. M. & Smith, M. E. Phylogenetic systematics of syncephalis (zoopagales, zoopagomycotina), a genus of ubiquitous mycoparasites. *Mycologia* **109**, 333–349 (2017).

41. Zhalnina, K. v. *et al.* Genome sequence of Candidatus nitrososphaera evergladensis from group I.1b enriched from everglades soil reveals novel genomic features of the ammonia-oxidizing archaea. *PLoS One* **9**, e101648 (2014).

42. Lücker, S. *et al.* A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci U S A* **107**, 13479–84 (2010).

43. Kennedy, P. G., Peay, K. G. & Bruns, T. D. Root tip competition among ectomycorrhizal fungi: are priority effects a rule or an exception? *Ecology* **90**, 2098–107 (2009).

44. Werner, G. D. A. & Kiers, E. T. Order of arrival structures arbuscular mycorrhizal colonization of plants. *New Phytologist* **205**, 1515–1524 (2015).

45. Callaway, E. Devastating wheat fungus appears in Asia for first time. *Nature* vol. 532 421–422 Preprint at https://doi.org/10.1038/532421a (2016).

46. Onkendi, E. M., Kariuki, G. M., Marais, M. & Moleleki, L. N. The threat of root-knot nematodes (Meloidogyne spp.) in Africa: A review. *Plant Pathology* vol. 63 727–737 Preprint at https://doi.org/10.1111/ppa.12202 (2014).

47. Tian, H. *et al.* A comprehensive quantification of global nitrous oxide sources and sinks. *Nature* **586**, 248–256 (2020).

48. Lugato, E., Leip, A. & Jones, A. Mitigation potential of soil carbon management overestimated by neglecting N2O emissions. *Nat Clim Chang* **8**, 219–223 (2018).

49. Kassam, A., Friedrich, T. & Derpsch, R. Global spread of conservation agriculture. *International Journal of Environmental Studies* **76**, 29–51 (2019).

50. Schreefel, L., Schulte, R. P. O., de Boer, I. J. M., Schrijver, A. P. & van Zanten, H. H. E. Regenerative agriculture – the soil is the base. *Glob Food Sec* **26**, https://doi.org/10.1016/j.gfs.2020.100404 (2020).

383 51. Li, Y., Zhang, Q., Cai, Y., Yang, Q. & Chang, S. X. Minimum tillage and residue
384 retention increase soil microbial population size and diversity: Implications for
385 conservation tillage. *Science of the Total Environment* **716**, 137164 (2020).
386 52. UNDP, SCBD & UNEP-WCMC. *Creating a nature-positive future: the contribution of*
387 *protected areas and other effective area-based conservation measures*. (2021).
388 53. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of
389 sequences per sample. *Proc Natl Acad Sci U S A* **108**, 4516–4522 (2011).
390 54. Apprill, A., Mcnally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU
391 rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton.
392 *Aquatic Microbial Ecology* **75**, 129–137 (2015).
393 55. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. Practical
394 innovations for high-throughput amplicon sequencing. *Nat Methods* **10**, 999–1002
395 (2013).
396 56. Toju, H., Kurokawa, H. & Kenta, T. Factors influencing leaf- and root-associated
397 communities of bacteria and fungi across 33 plant orders in a grassland. *Front*
398 *Microbiol* **10**, 1–14 (2019).
399 57. Stevens, J. L., Jackson, R. L. & Olson, J. B. Slowing PCR ramp speed reduces chimera
400 formation from environmental samples. *J Microbiol Methods* **93**, 203–205 (2013).
401 58. Ushio, M. Interaction capacity as a potential driver of community diversity.
402 *Proceedings of the Royal Society B: Biological Sciences* **289**, 20212690 (2022).
403 59. Ushio, M. *et al.* Quantitative monitoring of multispecies fish environmental DNA
404 using high-throughput sequencing. *Metabarcoding Metagenom* **2**, 1–15 (2018).
405 60. Fujita, H. *et al.* Alternative stable states, nonlinear behavior, and predictability of
406 microbiome dynamics. *bioRxiv*.
407 61. Toju, H., Tanabe, A. S. A., Yamamoto, S. & Sato, H. High-coverage ITS primers for
408 the DNA-based identification of ascomycetes and basidiomycetes in environmental
409 samples. *PLoS One* **7**, e40863 (2012).
410 62. Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. & Knight, R. Error-correcting
411 barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods*
412 **5**, 235–237 (2008).
413 63. Tanabe, A. S. & Toju, H. Two new computational methods for universal DNA
414 barcoding: a benchmark using barcode sequences of bacteria, archaea, animals, fungi,
415 and land plants. *PLoS One* **8**, e76910 (2013).
416 64. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina
417 amplicon data. *Nat Methods* **13**, 581–583 (2016).

65. R Core Team. R: A language and environment for statistical computing. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/* (2020).
66. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11**, e1004226 (2015).
67. Põlme, S. *et al.* FungalTraits: a user-friendly traits database of fungi and fungus-like stramenopiles. *Fungal Divers* **105**, 1–16 (2020).
68. Watanabe, T., Masuda, N., Megumi, F., Kanai, R. & Rees, G. Energy landscape and dynamics of brain activity during human bistable perception. *Nat Commun* **5**, 4765 (2014).
69. Wood, S. mgcv: Mixed GAM computation vehicle with automatic smoothness estimation available at https://cran.r-project.org/web/packages/mgcv/index.html. (2022).
70. Newman, M. & Girvan, M. Finding and evaluating community structure in networks. *Phys Rev E* **69**, 026113 (2004).
71. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* **1695**, 1–9 (2006).
72. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977).
73. Olesen, J. M., Bascompte, J., Dupont, Y. L. & Jordano, P. The modularity of pollination networks. *Proc Natl Acad Sci U S A* **104**, 19891–19869 (2007).
74. Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *J Math Sociol* **2**, 113–120 (1972).

443 # Methods

444 **Soil samples and metadata**

445 Over research projects of National Agricultural and Food Research Organization (NARO),
446 which were carried out through five national research programs funded by Ministry of
447 Agriculture, Forestry and Fisheries, 2,903 rhizosphere/bulk soil samples were collected from
448 conventional agricultural fields across the Japanese Archipelago from January 23, 2006, to
449 July 28, 2014 (Source Data S1). When the latitude and longitude of the sampling positions
450 were round to one decimal place, 42 research sites were identified. Across the metadata of the
451 2,903 samples, the information of 19 crop plants, 34 former crop plants (including "no crop"),
452 13 soil taxonomic groups (e.g., "Andosol"), 60 experimental/research purposes (e.g., "soil
453 comparison between organic and conventional management") was described. Likewise, the
454 metadata included the information of dry soil pH, electrical conductivity, carbon/nitrogen
455 (C/N) ratio, and available phosphorous concentration from, 2,830, 2,610, 2,346, and 2,249
456 samples, respectively. In addition, the information of the severity of crop plant disease was
457 available for 1,472 samples. The values of the proportion of diseased plants or disease
458 severity index was normalized within the ranges from 0 to 100, and they were then
459 categorized into five levels (level 1, 0-20; level 2, 20-40; level3; 40-60; level 4, 60-80; level
460 5, 80-100). The plant pathogens examined in the disease-level evaluation were *Colletotrichum*
461 *gloeosporioides* on the strawberry, *Fusarium oxysporum* on the celery, the lettuce, the
462 strawberry, and the tomato, *Phytopthora sojyae* on the soybean, *Plasmodiophora brassicae*
463 on Cruciferae plants, *Pyrenochaeta lycopersici* on the tomato, *Pythium myriotylum* on the
464 ginger, *Ralstonia solanacearum* on the eggplant and the tomato, and *Verticillium* spp. on
465 Chinese cabbage. For continuous variables within the metadata, emergent outliers (mean + 5
466 SD) were converted into "NA" in the data matrix used in the following statistical analyses as
467 potential measurement/recording errors. Unrealistic electrical conductivity records (> 20)
468 were converted into "NA" as well.

469 For the collection of each sample, ca. 500 g of soil was mixed and sieved with 2-mm
470 mesh in the field. The samples were stored at -20 ºC until DNA extraction. In laboratory
471 conditions, 0.4 g of soil (fresh weight) was subjected to DNA extraction with FastDNA SPIN
472 Kit for Soil (Q-BioGene).

473

474 **DNA amplification and sequencing**

475 Profiling of soil microbial biodiversity was performed by targeting archaea, bacteria, and
476 fungi. For the amplification of the 16S rRNA V4 region of archaea and bacteria (prokaryotes),
477 the set of the forward primer 515f[53] and the reverse primer 806rB[54] were used. The primers
478 were fused with 3–6-mer Ns for improved Illumina sequencing quality[55] and Illumina
479 sequencing primers as detailed in a previous study[56]. PCR was performed using KOD ONE
480 PCR Master Mix (TOYOBO, Osaka) with the temperature profile of 35 cycles at 98 ºC for 10
481 seconds (denaturation), 55 ºC for 5 seconds (annealing of primers), and 68 ºC for 30 seconds
482 (extension), and a final extension at 68 ºC for 2 minutes. The ramp rate through the thermal
483 cycles was set to 1 ºC/sec to prevent the generation of chimeric sequences[57]. In the PCR, we
484 added five artificial DNA sequence variants with different concentrations (i.e., standard DNA
485 gradients; 1.0×10^{-4}, 5.0×10^{-5}, 2.0×10^{-5}, 1.0×10^{-5}, and 5.0×10^{-6} nM) to the PCR master mix
486 solution in order to calibrate 16S rRNA gene copy numbers in the template DNA samples as
487 detailed elseshere[58–60]. Likewise, the ITS1 region of fungi was amplified using the set of the
488 forward primer ITS1F_KYO1 and the reverse primer ITS2_KYO2[61]. PCR was performed
489 using the Illumina-sequencing fusion primer design mentioned above with the temperature
490 profile of 35 cycles at 98 ºC for 10 seconds, 53 ºC for 5 s seconds, and 68 ºC for 5 seconds,
491 and a final extension at 68 ºC for 2 minutes (ramp rate = 1 ºC/sec). Newly designed artificial
492 sequence variants (1.0×10^{-5}, 7.0×10^{-6}, 5.0×10^{-6}, 2.0×10^{-6}, and 1.0×10^{-6} nM; Supplementary
493 Information) were added to the PCR master mix as standard DNA gradients for the
494 calibration of the ITS sequence concentrations in the template DNA samples.

495 The PCR products of the prokaryiote 16S rRNA and fungal ITS1 regions were
496 respectively subjected to the additional PCR step for linking Illumina sequencing adaptors
497 and 8-mer sample identifier indexes[62] with the amplicons as detailed elsewhere[56]. The
498 temperature profile in the PCR was 8 cycles at 98 ºC for 10 seconds, 55 ºC for 5 seconds, and
499 68 ºC for 5 seconds, and a final extension at 68 ºC for 2 minutes. The PCR products were then
500 pooled for each of the 16S rRNA and fungal ITS1 regions after a purification/equalization
501 process with the AMPureXP Kit (Beckman Coulter, Inc., Brea). Primer dimers, which were
502 shorter than 200 bp, were removed from the pooled library by supplemental purification with
503 AMpureXP: the ratio of AMPureXP reagent to the pooled library was set to 0.8 (v/v) in this
504 process. The sequencing libraries of the two regions were processed in an Illumina MiSeq
505 sequencer (10% PhiX spike-in). Because the quality of forward sequences is generally higher
506 than that of reverse sequences in Illumina sequencing, we optimized the MiSeq run setting in
507 order to use only forward sequences. Specifically, the run length was set 271 forward (R1)
508 and 31 reverse (R4) cycles to enhance forward sequencing data: the reverse sequences were

509 used only for discriminating between prokaryote 16S and fungal ITS1 sequences in the
510 following bioinformatic pipeline.

511

512 **Bioinformatics**

513 In total, 23,573,405 sequencing reads were obtained in the Illumina sequencing (16S rRNA,
514 11,647,166 sequencing reads; ITS, 11,926,239 sequencing reads). The raw sequencing data
515 were converted into FASTQ files using the program bcl2fastq 1.8.4 distributed by Illumina.
516 For each of the 16S rRNA and fungal ITS1 regions, the output FASTQ files were
517 demultiplexed using Claident v0.9.2022.01.26[63]. The sequencing data were deposited to DNA
518 Data Bank of Japan (DDBJ) (Bioproject accession no.: PSUB018361). The removal of low-
519 quality sequences and ASV inferences were done using DADA2[64] v1.17.5 of R v3.6.3[65]. The
520 mean number of filtered sequencing reads obtained per sample was 3,949 and 4,075 for the
521 prokaryote and fungal, datasets, respectively. Taxonomic annotation of prokaryotes was
522 conducted based on the SILVA 138 SSU (Quast et al., 2013) the assignTaxonomy function of
523 DADA2. For the taxonomic annotation of fungi, the QCauto method[63] was applied to the
524 NCBI nucleotide sequence database bundled with Claident v0.9.2022.01.26. ASVs that were
525 not assigned to the domain Archaea/Bacteria and the kingdom Fungi were removed from the
526 16S rRNA and ITS1 datasets, respectively. For each target organismal group (prokaryotes and
527 fungi), we then obtained a sample × ASV matrix, in which a cell entry depicted the number of
528 sequencing reads of an ASV in a sample. The samples with less than 1,000 reads were
529 discarded from the matrices. The matrix was rarefied to 1,000 reads per sample for the
530 downstream statistical pipeline except for the network analysis, which had an original
531 standardization protocol of input data[66]. In total, prokaryote and fungal community data were
532 obtained for 2,534 and 2,384 samples, respectively. The number of ASVs included in the
533 rarefied matrices were 27,219 and 6,306 for prokaryotes and fungi, respectively. For fungal
534 ASVs, putative functional groups (e.g., "plant pathogen") were inferred using the program
535 FungalTraits[67]. The estimation of DNA concentrations of the prokaryote 16S rRNA and
536 fungal ITS regions was performed, respectively, based on the calibration with the standard
537 DNA gradients (artificial DNA variants introduced to the PCR master mix solutions) as
538 detailed elsewhere[60].

539

540 **Diversity of microbiome structure**

541 For each sample, Shannon's α-diversity estimates of prokaryotes and fungi were respectively
542 calculated at the family level. Likewise, mean Bray-Curtis dissimilarity (β-diversity) in
543 family-level compositions within each of the 42 research sites was calculated as a measure of
544 within-site variability of prokaryote/fungal community structure. In addition, the degree of
545 among-site community differentiation was evaluated for prokaryotes and fungi. Specifically,
546 after calculating mean family-level compositions at each research site, Bray-Curtis β-diversity
547 was calculated in all pairs of research sites. Those α-diversity, within-site β-diversity, and
548 among-site β-diversity estimates were respectively compared between prokaryotes and fungi.

549 For each of the prokaryote and fungal datasets, permutational analyses of variance
550 (PERMANOVA) were performed to examine associations between family-level community
551 compositions and variables in the metadata. Prior to the PEMANOVA, Bray-Curtis β-
552 diversity was calculated based on the rarefied sample × ASV matrices. Two types of
553 PEMANOVA models were constructed (10,000 iterations). Specifically, one is constituted by
554 categorical explanatory variables (crop plant, former crop plant, soil taxonomy, research site,
555 and sampling date), while the other included continuous explanatory variables (soil pH,
556 electrical conductivity, C/N ratio, and available phosphorous concentration, latitude, and
557 longitude).

558 To reduce the dimensions of the community composition data, a principal coordinate
559 analysis (PCoA) was performed based on the Bray-Curtis β-diversity calculated with the
560 rarefied sample × ASV matrix of prokaryotes or fungi. For each PCoA axis (axes 1 to 5) in
561 each of the prokaryote and fungal analyses, Pearson's correlation with each chemical
562 environmental factor (soil pH, electrical conductivity, C/N ratio, and available phosphorous
563 concentration) was calculated.

564

565 **Crop disease and microbiome structure**

566 We then evaluated how community structure of prokaryotes and fungi are associated with
567 crop disease. For each of the prokaryote and fungal data sets, a generalized linear mixed
568 model (GLMM) of crop-disease level was constructed by included the PCoA axes 1 to 5 as
569 fixed effects and experimental/research purposes in the metadata as random effects. A logit-
570 link function and binomial error was assumed in the GLMM after converting the response
571 variable into a binary format [disease level 1 (= 1) vs. disease levels 2–5 (= 0)].

572 We also examined how balance between prokaryote and fungal abundance could be
573 associated with crop disease levels. Based on the estimated concentrations of prokaryote 16S
574 rRNA and fungal ITS sequences in template DNA solutions, we calculated the ratio of
575 prokaryote DNA concentrations to fungal DNA concentrations in respective samples
576 (prokaryote/fungus ratio) as follows:

577 $$log\frac{\text{prokaryote 16S rRNA gene concentration(DNA copies/}\mu\text{L)}}{\text{fungal ITS gene concentration (DNA copies/}\mu\text{L)}}.$$

578 Although potential variation in DNA extraction skills of researchers might affect absolute
579 DNA concentrations in the template DNA solutions, balance between prokaryote and fungal
580 DNA in each template DNA sample could be used as a reliable measure.

581 Potential effects of the prokaryote/fungus ratio on crop-disease level were examined in
582 a GLMM with a logit-link function and binomial errors. In the GLMM, soil pH and electrical
583 conductivity were included as fixed effects in addition to the prokaryote/fungus ratio. In the
584 GLMM, C/N ratio and available phosphorous concentration were not included due to the
585 small number of samples with both crop-disease information and the chemical variables. The
586 experimental/research purpose information in the metadata was included as random effects in
587 the GLMM. To examine the effects of other aspects of biological factors, we constructed
588 another GLMM including the ASV-level Shannon diversity of prokaryotes, that of fungi, and
589 the prokaryote/fungus ratio. All the explanatory variables were *z*-standardized (zero-mean and
590 unit-variance) before the GLMM analysis. To visualize how the proportion of least-diseased
591 (disease-level 1) samples varied along the axis of the prokaryote/fungus ratio, we performed a
592 supplementary analysis with generalized linear models (GLMs). Specifically, for the five
593 experimental/research purpose categories from which 10 or more disease-level 1 (least-
594 diseased) samples and 10 or more disease-level 2–5 (severely diseased) samples were
595 obtained, a GLM of disease level was constructed by including the prokaryote/fungus ratio as
596 an explanatory variable.

597

598 **Alternative stables states**

599 We estimated the alternative stable states of microbial community structure based on the
600 framework of an energy landscape analysis[32,68] (tutorials of energy landscape analyses are
601 available at https://community.wolfram.com/groups/-/m/t/2358581). Probabilities of
602 community states $\mathrm{p}\left(\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})}\right)$ are given by

$$P\left(\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})}|\boldsymbol{\varepsilon}\right) = \boldsymbol{e}^{-\boldsymbol{H}(\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})},\boldsymbol{\varepsilon})}/Z$$

$$Z = \sum_{\boldsymbol{i}=\mathbf{1}}^{\mathbf{2}^{\boldsymbol{S}}} \boldsymbol{e}^{-\boldsymbol{H}(\vec{\boldsymbol{\sigma}}^{(\boldsymbol{i})},\boldsymbol{\varepsilon})},$$

where $\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})} = (\sigma_1^{(k)}, \sigma_2^{(k)}, \dots, \sigma_S^{(k)})$ is a community state vector of *k*-th community state and *S* is the total number of taxa (e.g., ASVs, species, genera, or families) examined. $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M)$ is an array of continuous values representing environmental factors (e.g., soil pH and electrical conductivity) and *M* is the total number of environmental parameters. $\sigma_{\mathrm{i}}^{(k)}$ is a binary variable that represents presence (1) or absence (0) of taxon *i*: i.e., there are a total of 2^S community states. As the exploration of the 2^S community states were computationally intensive, we coded community states at the family level taxonomic compositions. Specifically, families that represented 10 % or more of the total sequencing reads of a target sample were coded as 1, while those representing less than 10 % of total reads were coded as 0. Then, the energy of the community state $\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})}$ is given by the extended pairwise maximum entropy model:

$$H(\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})}, \boldsymbol{\varepsilon}) = -\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{S}} h_{\boldsymbol{i}} \vec{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{(\boldsymbol{k})} - \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{S}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{M}} \boldsymbol{g}_{\boldsymbol{ij}}\, \boldsymbol{\varepsilon}_{\boldsymbol{i}}^{(\boldsymbol{k})} \boldsymbol{\sigma}_{\boldsymbol{j}}^{(\boldsymbol{k})} - \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{S}} \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{S}} \boldsymbol{J}_{\boldsymbol{ij}} \vec{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{(\boldsymbol{k})} \vec{\boldsymbol{\sigma}}_{\boldsymbol{j}}^{(\boldsymbol{k})}/\mathbf{2},$$

where h_i represents the net effect of implicit abiotic factors, by which *i*-th taxon is more likely to present ($h_i > 0$) or not ($h_i < 0$), g_{ij} represents the effect of the *i*-th observed environmental factor, and J_{ij} represents a co-occurrence pattern of *i*-th and *j*-th taxa. Since the logarithm of the probability of a community state is inversely proportional to $H(\vec{\boldsymbol{\sigma}}^{(\boldsymbol{k})})$, a community state having lower *H* is more frequently observed. To consider dynamics on an assembly graph defined as a network whose 2^S nodes represent possible community states and the edges represents transition path between them (two community states are adjacent only if they have the opposite presence/absence status for just one species), we assigned energy to nodes with the above equation, and so imposed the directionality in state transitions. Then, by using the steepest descent algorithm[32] we identified the stable state communities as the energy minima of the weighted network (nodes having the lowest energy compared to all its neighbors), and determined their basins of attraction based on a steepest descent procedure starting from each node. As environmental variables, soil pH, electrical conductivity, C/N ratio, and available phosphorous concentration were included after normalization within the ranges from 0 to 1.

The number of stable states within energy landscapes of community structure was inferred, respectively, for three types of datasets, namely, the prokaryote community matrix,

634 the fungal matrix, and the matrix including both prokaryotes and fungi. In addition to the
635 energy landscape analyses based on the abovementioned family-level delineation of
636 community states, analyses based on community-state delineation at the order-level were
637 performed. Likewise, additional analyses based on another threshold value (5 %) in the
638 binarization of community states were conducted. Consequently, we performed energy
639 landscape analyses in each of the combinations of three types of data (prokaryotes, fungi, and
640 prokaryotes + fungi), two taxonomic levels (family or order), and two threshold levels (5 %
641 and 10 %). For each combination (setting), the parameters of the extended pairwise maximum
642 entropy model were adjusted to fit the model to the empirical data. The number of stable
643 states was then estimated by using the gradient descent algorithm for 20,000 randomly
644 generated community states within the 2^X possible community states. In the heuristic search,
645 sets of values of the environmental variables (soil pH, electrical conductivity, C/N ratio, and
646 available phosphorous concentration) were drawn from those of sampling points.

647 We then reconstructed energy landscapes for the settings with which community
648 structure could be characterized by relatively small numbers of stable states (i.e., the family-
649 level fungal data matrix defined at the 10-% threshold and the order-level fungal data matrix
650 defined at the 10-% threshold; Fig. 4; Extended Data Fig. 4). Minor stable states whose
651 shallowest "energy barrier" (i.e., deviation of energy from adjacent "tipping points") was less
652 than 1/10 of that of the deepest alternative stables states were incorporated into the basins of
653 deeper stable states (i.e., pruning of alternative stable states[32]). The "energy" estimates were
654 then plotted against the PCoA axes mentioned above. Spline smoothing of the landscape was
655 performed with optimized penalty scores using the mgcv v.1.8-40 package[69] of R.

656 We then showed the energy of each sample on the two-dimensional surface of the state
657 space inferred with the above-mentioned PCoA: the horizontal axes indicated PCoA 1 and 2
658 scores and the vertical axis indicated the energy of target samples' community states. On the
659 energy landscape, samples belonging to the same alternative stable states were indicated by
660 colors. For the alternative stable states that represented 10 or more samples, the proportions of
661 disease-level-1 samples were calculated.

662

663 **Microbiome network modules**

664 To infer potential interactions between microbial ASVs, the algorithm of sparse inverse
665 covariance estimation for ecological association inference (SPIEC-EASI) was applied based

666 on the Meinshausen-Bühlmann (MB) method as implemented in the SpiecEasi package[66] of
667 R. The network inference based on co-occurrence patterns allowed us to detect pairs of
668 microbial ASVs that potentially interact with each other in facilitative ways and/or those that
669 might share ecological niches (e.g., preference for edaphic factors). Because estimation of co-
670 occurrence patterns was not feasible for rare nodes, the prokaryote and fungal ASVs that
671 appeared in more than 10 % of the sequenced samples were included in the input matrix of
672 the network analysis. Network modules, within which closely associated ASVs were
673 interlinked with each other, were identified with the algorithm based on edge betweenness[70]
674 using the igraph package[71] or R.

675 We then characterized the inferred network modules based on their potential impacts on
676 crop-plant disease level. For each microbial ASV constituting the modules, we evaluated
677 specificity of occurrences in samples differing in crop disease levels based on a
678 randomization analysis. For the calculation, the rarefied matrices of prokaryotes and fungi
679 were combined. Within the combined sample × ASV matrix, samples were categorized into
680 the five disease levels. Mean read counts across samples displaying each of the five disease
681 levels were then calculated for each ASV. Meanwhile, mean read counts for respective
682 disease levels were calculated as well for randomized matrices, in which disease labels of the
683 samples were shuffled (10,000 permutations). For *i*-th ASV, standardized specificity to
684 disease level *j* (s_{ij}) was obtained as follows:

685 $$s_{ij} = \frac{O_{ij} - Mean(R_{ij})}{SD(R_{ij})},$$

686 where O_{ij} and R_{ij} is the mean read counts of *i*-th ASV across disease-level-*j* samples in the
687 observed and randomized matrices, respectively, and $Mean(R_{ij})$ and $SD(R_{ij})$ indicate mean
688 and standard deviation across the randomized matrices. An s_{ij} value higher than 4 or that
689 lower than -4 indicates statistically strong specificity of the *i*-th ASV to the disease level *j*
690 (Extended Data Fig. 6). For each module in the inferred co-occurrence network, we then
691 calculated mean standardized specificity to disease level 1 across constituent ASVs.

692

693 **Network hubs**

694 To explore potential keystone microbes within the network, we scored respective ASVs on
695 the basis of their topological positions. Among the indices used for evaluating ASVs,
696 betweenness centrality[72], which measures the extent to which a given nodes (ASV) is located
697 within the shortest paths connecting pairs of other nodes in a network, is commonly used to

find hubs mediating flow of effects in a network. The network centrality scores were normalized as implemented in the igraph packages of R[71]. In addition, by focusing on the above-mentioned network modules, we ranked ASVs based on their within-module degree and among-module connectivity[73]. The former index is obtained as the number of nodes linked with a target node within a target network module, suggesting the topological importance of a node within the module it belongs to. The latter index represents the extent to which a node interlinks nodes belonging to different network modules. Within-module degree was *z*-standardized (i.e., zero-mean and unit-variance) within each module, while among-module connectivity was defined to vary between 0 to 1. In addition to those indices for evaluating topological roles within a network, eigenvector centrality[74] was calculated for respective nodes.

Data availability

The 16S rRNA and ITS sequencing data are available from the DNA Data Bank of Japan (DDBJ: Bioproject: PSUB018361). The microbial community data are deposited at our GitHub repository (https://github.com/hiro-toju/Soil_Microbiome_NARO3000).

Code availability

All the R scripts used to analyze the data are available at the GitHub repository (https://github.com/hiro-toju/Soil_Microbiome_NARO3000).

Acknowledgements. We thank the SuperComputer System, Institute for Chemical Research, Kyoto University for the use of super computers. This work was financially supported by JST PRESTO (JPMJPR16Q6), JST FOREST (JPMJFR2048), Human Frontier Science Program (RGP0029/2019), JSPS Grant-in-Aid for Scientific Research (20K20586), NEDO Moonshot Research and Development Program (JPNP18016), and JST FOREST (JPMJFR2048) to H.T., JSPS Grant-in-Aid for Scientific Research (20K06820 and 20H03010) to K.S., and JSPS Fellowship to H.F..

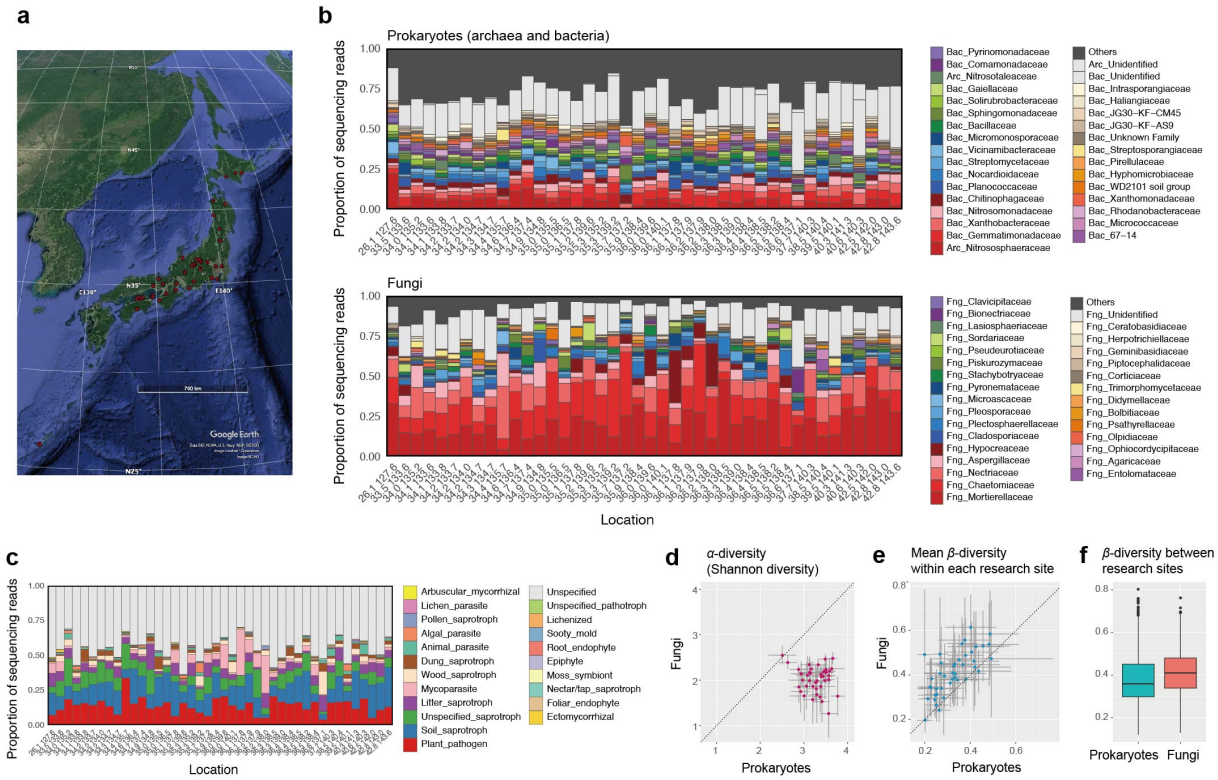
Author Contributions. H.F., S.Y. and H.T. designed the work. H.F. performed molecular experiments. H.F. and H.T. analyzed the data. H.T. wrote the paper with H.F., S.Y., and K.S..

Competing Interests. The authors declare no competing interests.

Additional information

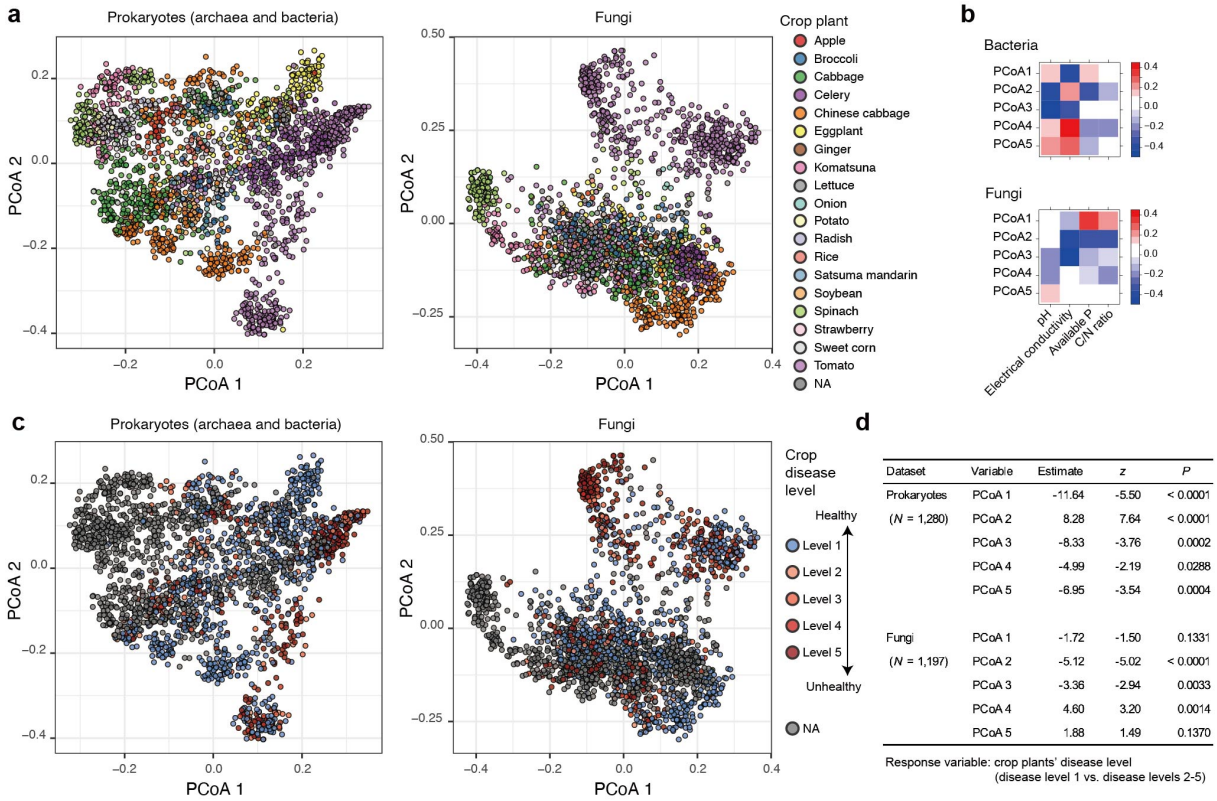
Supplementary information is available for this paper at xxxx [URL to be supplied by the publisher].

Correspondence and requests for materials should be addressed to toju.hirokazu.4c@kyoto-u.ac.jp.



739

740 **Fig. 1 | Diversity of agroecosystem microbiomes**. **a**, Map of research sites across the
741 Japanese Archipelago. The 2,903 soil samples were grouped into 42 research sits were when
742 their latitude and longitude profiles were round to one decimal place. **b**, Taxonomic
743 compositions of prokaryotes (archaea and bacteria; top) and fungi (bottom) at the family
744 level. See Extended Data Fig. 1 for results at the genus, order, and class levels. **c**,
745 Compositions of functional groups of fungi. **d**, Comparison of α-diversity between
746 prokaryotes and fungi. Mean Shannon diversity at the family level was shown for each
747 research site. Bars represent standard deviations. **e**, Mean community dissimilarity (β-
748 diversity) between pairs of samples within each research site. **f**, Community dissimilarity (β-
749 diversity) in mean community compositions between pairs of research sites.
750



Dataset	Variable	Estimate	z	P
Prokaryotes	PCoA 1	-11.64	-5.50	< 0.0001
(*N* = 1,280)	PCoA 2	8.28	7.64	< 0.0001
	PCoA 3	-8.33	-3.76	0.0002
	PCoA 4	-4.99	-2.19	0.0288
	PCoA 5	-6.95	-3.54	0.0004
Fungi	PCoA 1	-1.72	-1.50	0.1331
(*N* = 1,197)	PCoA 2	-5.12	-5.02	< 0.0001
	PCoA 3	-3.36	-2.94	0.0033
	PCoA 4	4.60	3.20	0.0014
	PCoA 5	1.88	1.49	0.1370

Response variable: crop plants' disease level (disease level 1 vs. disease levels 2-5)

Fig. 2 | Dimensions of soil microbiome structure. **a**, Principal co-ordinate analysis (PCoA) of prokaryote and fungal community structure. The identify of crop plants is shown by colors. See Extended Data Figs. 2-3 for relationship between community structure and environmental factors. **b**, Correlation between PCoA scores and soil environmental factors. **c**, Microbiome structure and crop disease levels. **d**, Relationship between prokaryote/fungal community structure on the disease level of crop plants. A GLMM of crop plants' disease level (disease level 1 vs. disease levels 2-5) with a logit-link function and binomial errors was constructed by setting principal coordinates of prokaryote/fungal community structure as explanatory variables (fixed effects). The identity of research projects was included as random effects in the GLMM.

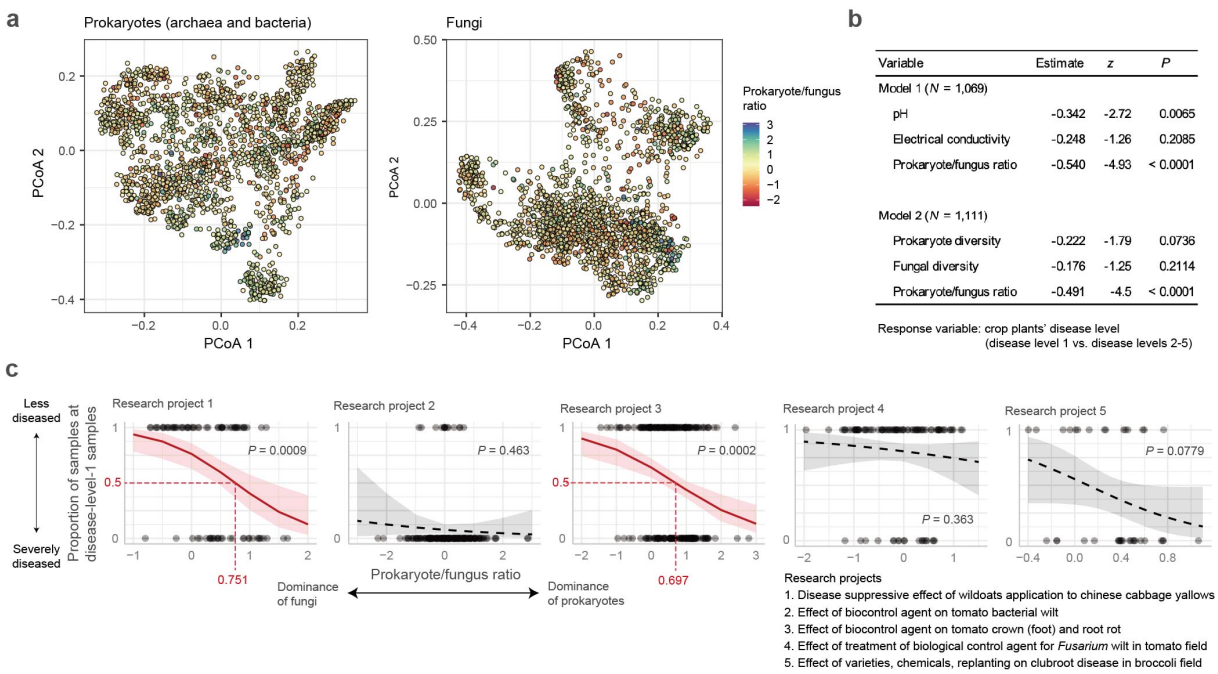


Fig. 3 | Balance between prokaryote and fungal abundance. **a**, Prokaryote/fungus ratio. The relative abundance of prokaryotes and fungi was evaluated based on estimated DNA concentrations ($log\,\frac{\text{prokaryote 16S rRNA gene concentration(DNA copies/}\mu\text{L)}}{\text{fungal ITS gene concentration (DNA copies/}\mu\text{L)}}$). The prokaryote/fungus ratio is shown for each sample on the PCoA surfaces of prokaryote and fungus community compositions. **b**, Effects of prokaryote/fungus ratio on the disease level of crop plants. In each GLMM with a logit-link function and binomial errors, crop plants disease level [least-diseased (disease level 1) = 1, severely-diseased (disease level 2-5) = 0] was included as response variable. In addition to bacterium/fungal abundance ratio, soil environmental variables (pH and electrical conductivity; Model 1) and α-diversity estimates of prokaryotes and fungi (Model 2) were included as explanatory variables (fixed effects). The identity of research projects was included as random effects in the GLMM. A negative estimate indicates that the proportion of least-diseased plants decreased with increasing values of a target explanatory variable. **c**, Representative relationship between prokaryote/fungus ratio on crop disease. For each of the five research projects with enough number of disese-level-1 and disease-level-2-5 samples, a GLM was constructed. For the two research projects with significant effects of prokaryote/fungus ratio on crop disease level, the prokaryote/fungus ratios at which 50 % of samples display disease-level-1 symptoms are shown.

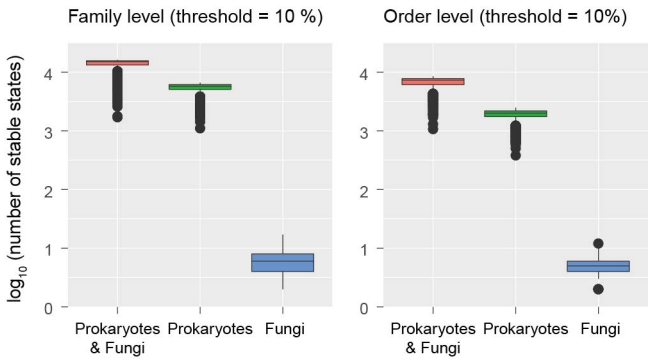


Fig. 4 | Number of the stable states of community structure. The parameters of the extended pairwise maximum entropy model were adjusted to fit the model to the empirical data (see Methods). The number of stable states in the energy landscape was then estimated by using the gradient descent algorithm for 20,000 randomly generated community states within the 2^S possible community states, where S denotes the number of families or orders. In the heuristic search, sets of values of the environmental variables (soil pH, electrical conductivity, C/N ratio, and available phosphorous concentration) were drawn from those of sampling points. The boxplots show variation in the number of stable states across different sets of environmental variables. A box represents the first quartile, median and the third quartile, while bars and circles indicate 1.5 × inter-quartile range and outliers, respectively. The analysis was performed, respectively, for the dataset representing both prokaryotes and fungi, the prokaryote dataset, and the fungus dataset. The threshold value in the binarization of the community data was set at 10 %: see Extended Data Figure 4 for results based on another threshold value (5 %).

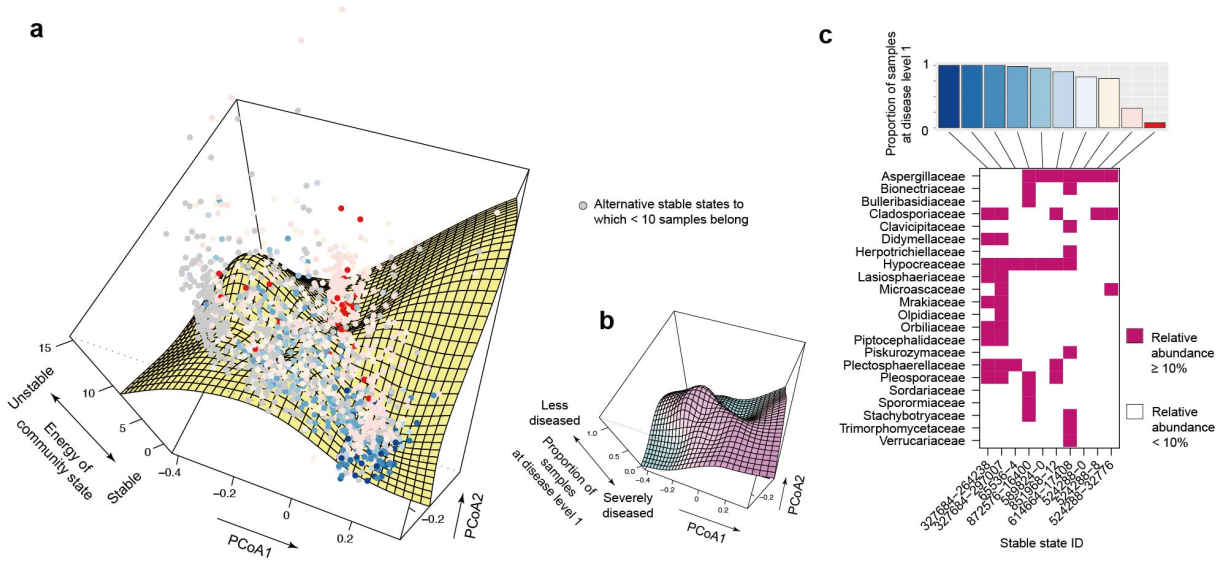


Fig. 5 | Alternative stable states of microbiomes. **a**, Energy landscape of fungal community structure. The surface of energy levels was reconstructed on the PCoA dimensions (Fig. 2) of fungal community structure based on spline smoothing. Community states with lower energy are inferred to be more stable. Data points (samples) indicated by the same color belong to the basins of the same alternative stable states, whose energy is lower than any adjacent community structure. See panel **c** for details about the alternative stable states. **b**, Surface of crop disease conditions. The area of community structure with less diseased states and lower energy (i.e., community stability; panel **a**) are expected to represent stably-manageable microbiomes with potential disease-suppressive functions. **c**, Definitions of the stable states. To reduce the complexity of input data to a computationally feasible range, alternative stable states were explored based on family-level community compositions (threshold for binarization = 10 %). The 10 alternative stable states representing ≥ 10 samples were compared in terms of their associations with crop disease conditions.

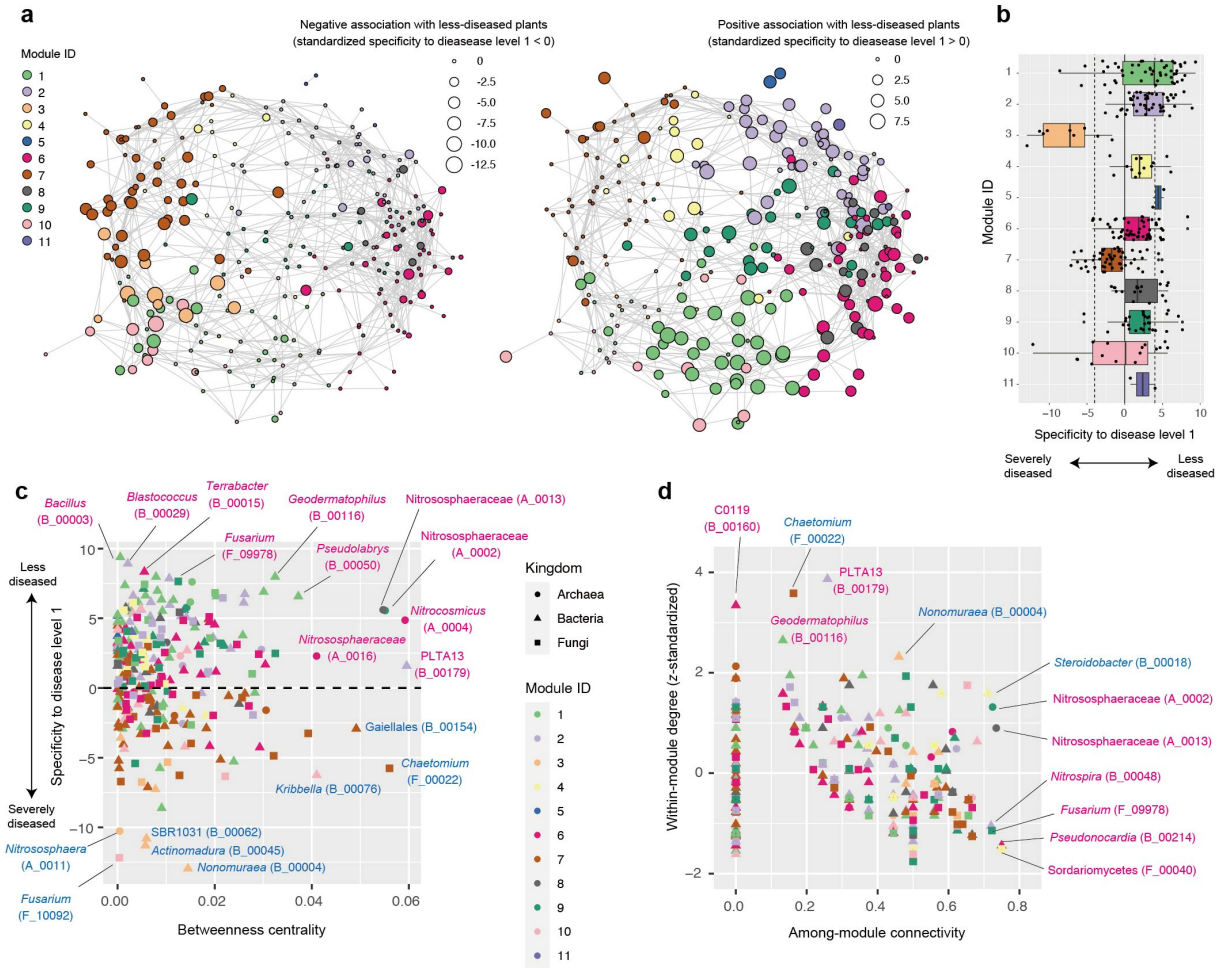
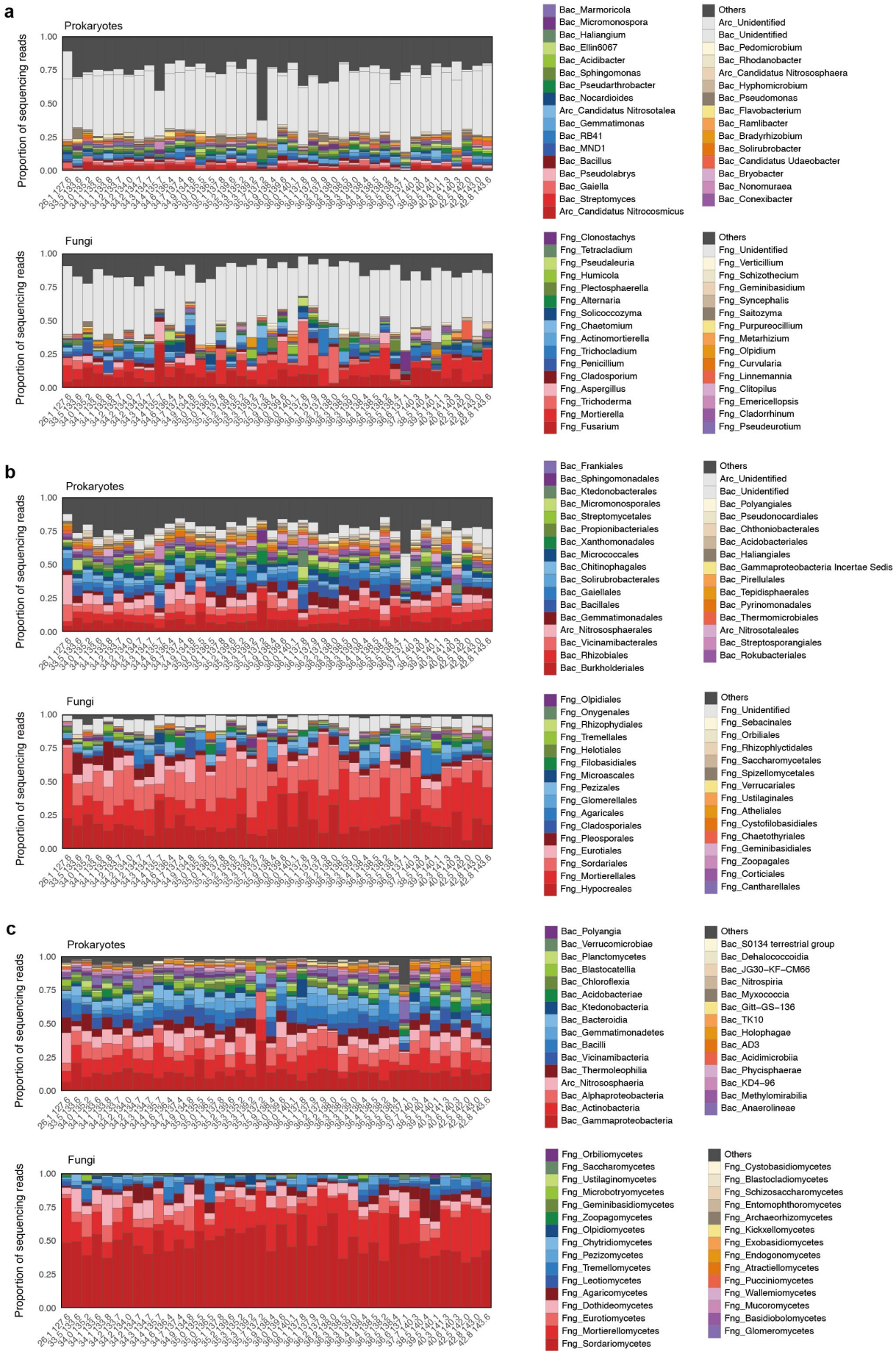


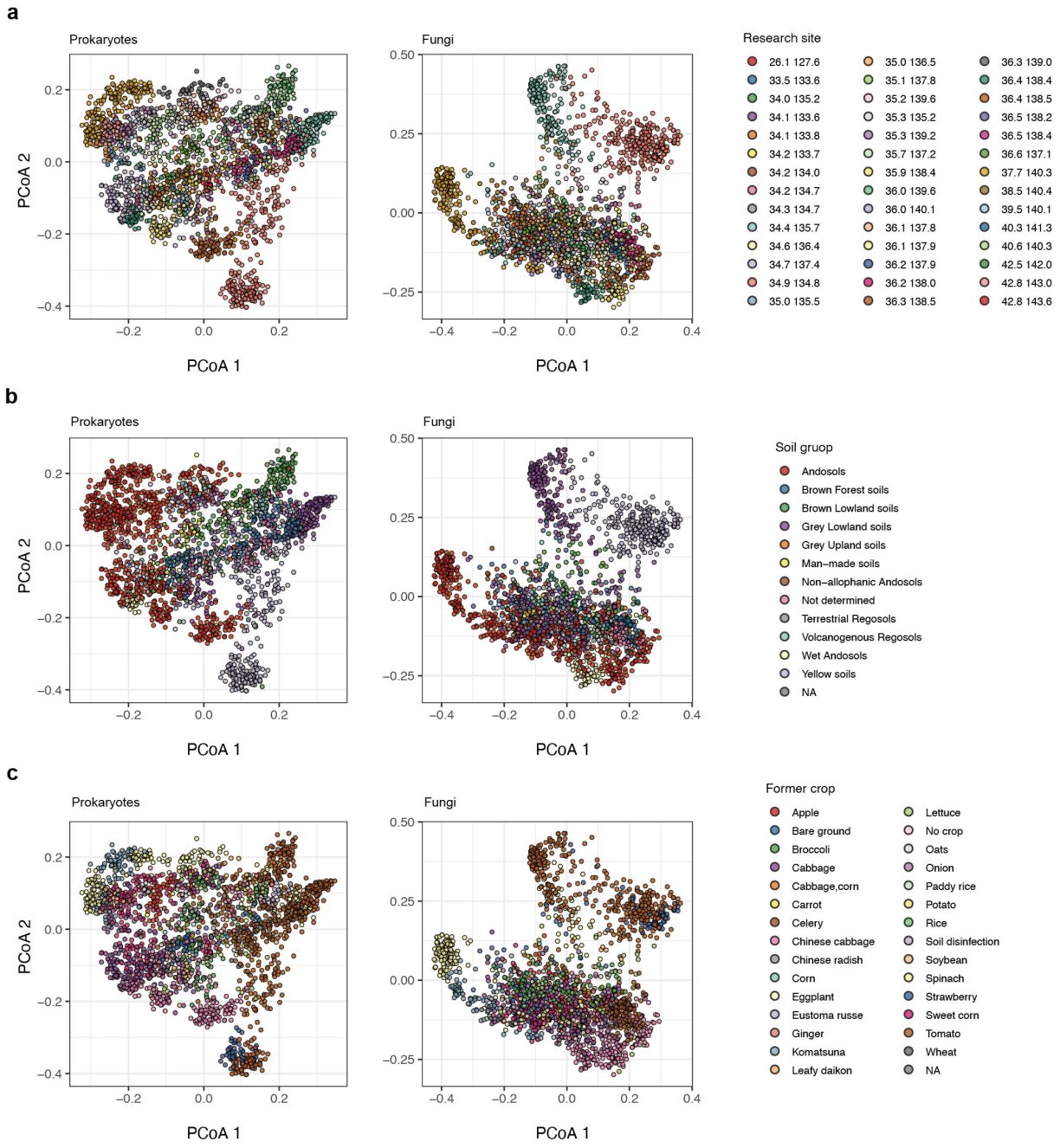
Fig. 6 | Network modules and hubs within the microbiomes. **a**, Co-occurrence networks of archaea, bacteria, and fungi. Specificity of occurrences to disease-level-1 (the lowest disease level) samples is shown for each ASV within the network. The specificity is shown as node size separately for negative (left) and positive (right) associations with least-diseased states of crop plants. Colors indicate network modules, in which microbial ASVs in commensalistic/mutualistic interactions and/or those sharing environmental preferences are densely linked with each other. See Extended Data Figure 7 for taxonomy (archaea, bacteria, or fungi) of respective nodes. **b**, Characteristics of network modules. Mean specificity to disease-level-1 samples is shown for each network module. **c**, Betweenness centrality of ASVs. For each ASV, betweenness centrality within the co-occurrence network is shown along the horizontal axis. Specificity to disease-level-1 samples is shown along the vertical axis. **d**, Topological roles of ASVs within and across network modules. For each ASV, within-module degree represents the number of the ASVs linked with the target ASV within a module (*z*-standardized). Among-module connectivity represents the extent to which an ASV interlinks ASVs belonging to different network modules.

Extended Data Figures

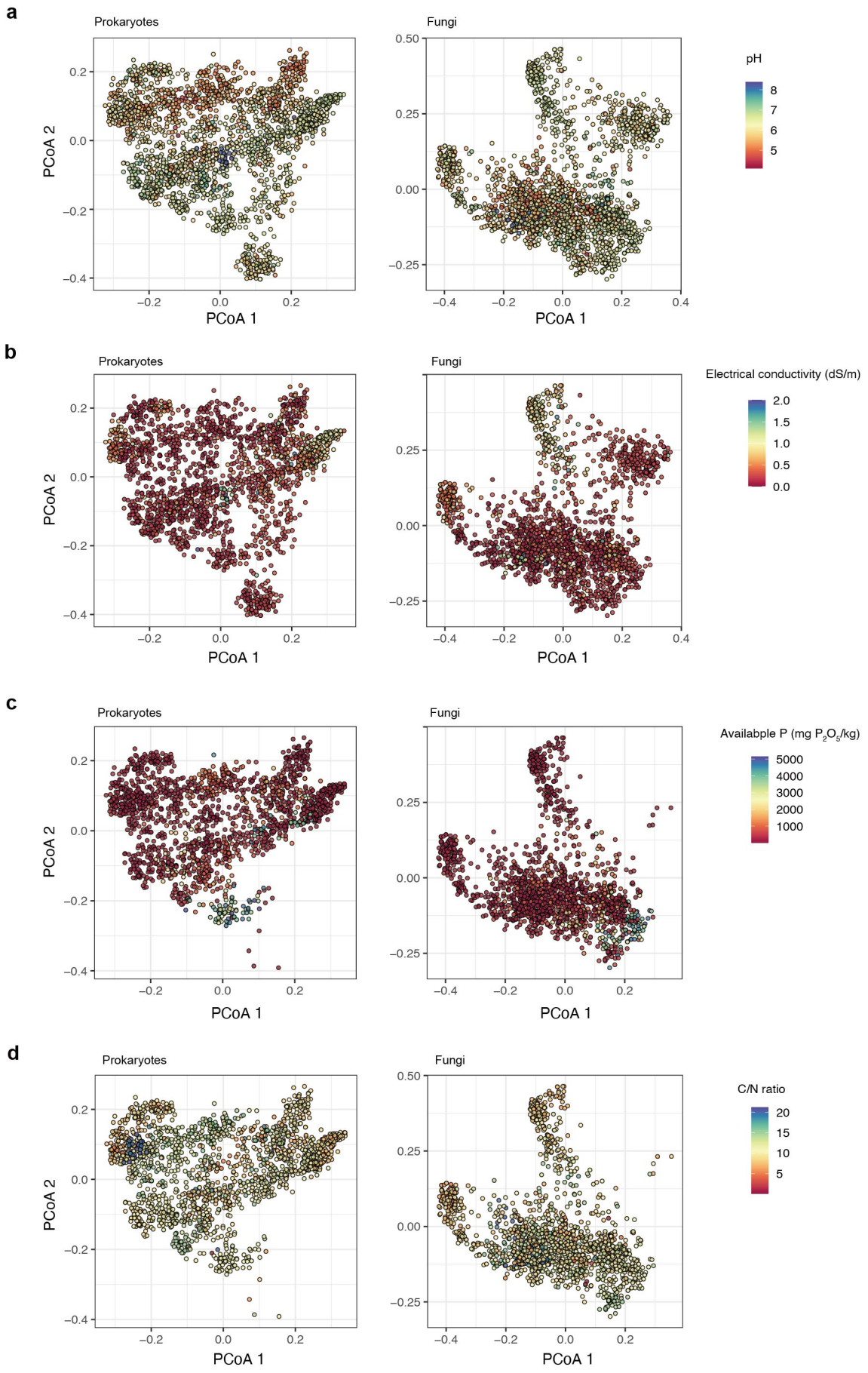

a
Prokaryotes
Proportion of sequencing reads
1.00
0.75
0.50
0.25
0.00
Bac_Marmoricola
Bac_Micromonospora
Bac_Haliangium
Bac_Ellin6067
Bac_Acidibacter
Bac_Sphingomonas
Bac_Pseudarthrobacter
Bac_Nocardioides
Arc_Candidatus Nitrosotalea
Bac_Gemmatimonas
Bac_RB41
Bac_MND1
Bac_Bacillus
Bac_Pseudolabrys
Bac_Gaiella
Bac_Streptomyces
Arc_Candidatus Nitrocosmicus
Others
Arc_Unidentified
Bac_Unidentified
Bac_Pedomicrobium
Bac_Rhodanobacter
Arc_Candidatus Nitrososphaera
Bac_Hyphomicrobium
Bac_Pseudomonas
Bac_Flavobacterium
Bac_Ramlibacter
Bac_Bradyrhizobium
Bac_Solirubrobacter
Bac_Candidatus Udaeobacter
Bac_Bryobacter
Bac_Nonomuraea
Bac_Conexibacter
Fungi
Fng_Clonostachys
Fng_Tetracladium
Fng_Pseudaleuria
Fng_Humicola
Fng_Plectosphaerella
Fng_Alternaria
Fng_Solicoccozyma
Fng_Chaetomium
Fng_Actinomortierella
Fng_Trichocladium
Fng_Penicillium
Fng_Cladosporium
Fng_Aspergillus
Fng_Trichoderma
Fng_Mortierella
Fng_Fusarium
Others
Fng_Unidentified
Fng_Verticillium
Fng_Schizothecium
Fng_Geminibasidium
Fng_Syncephalis
Fng_Saitozyma
Fng_Purpureocillium
Fng_Metarhizium
Fng_Olpidium
Fng_Curvularia
Fng_Linnemannia
Fng_Clitopilus
Fng_Emericellopsis
Fng_Cladorrhinum
Fng_Pseudeurotium
b
Prokaryotes
Bac_Frankiales
Bac_Sphingomonadales
Bac_Ktedonobacterales
Bac_Micromonosporales
Bac_Streptomycetales
Bac_Propionibacteriales
Bac_Xanthomonadales
Bac_Micrococcales
Bac_Chitinophagales
Bac_Solirubrobacterales
Bac_Gaiellales
Bac_Bacillales
Bac_Gemmatimonadales
Arc_Nitrososphaerales
Bac_Vicinamibacterales
Bac_Rhizobiales
Bac_Burkholderiales
Others
Arc_Unidentified
Bac_Unidentified
Bac_Polyangiales
Bac_Pseudonocardiales
Bac_Chthoniobacterales
Bac_Acidobacteriales
Bac_Haliangiales
Bac_Gammaproteobacteria Incertae Sedis
Bac_Pirellulales
Bac_Tepidisphaerales
Bac_Pyrinomonadales
Bac_Thermomicrobiales
Arc_Nitrosotaleales
Bac_Streptosporangiales
Bac_Rokubacteriales
Fungi
Fng_Olpidiales
Fng_Onygenales
Fng_Rhizophydiales
Fng_Tremellales
Fng_Helotiales
Fng_Filobasidiales
Fng_Microascales
Fng_Pezizales
Fng_Glomerellales
Fng_Agaricales
Fng_Cladosporiales
Fng_Pleosporales
Fng_Eurotiales
Fng_Sordariales
Fng_Mortierellales
Fng_Hypocreales
Others
Fng_Unidentified
Fng_Sebacinales
Fng_Orbiliales
Fng_Rhizophlyctidales
Fng_Saccharomycetales
Fng_Spizellomycetales
Fng_Verrucariales
Fng_Ustilaginales
Fng_Atheliales
Fng_Cystofilobasidiales
Fng_Chaetothyriales
Fng_Geminibasidiales
Fng_Zoopagales
Fng_Corticiales
Fng_Cantharellales
c
Prokaryotes
Bac_Polyangia
Bac_Verrucomicrobiae
Bac_Planctomycetes
Bac_Blastocatellia
Bac_Chloroflexia
Bac_Acidobacteriae
Bac_Ktedonobacteria
Bac_Bacteroidia
Bac_Gemmatimonadetes
Bac_Bacilli
Bac_Vicinamibacteria
Bac_Thermoleophilia
Arc_Nitrososphaeria
Bac_Alphaproteobacteria
Bac_Actinobacteria
Bac_Gammaproteobacteria
Others
Bac_S0134 terrestrial group
Bac_Dehalococcoidia
Bac_JG30–KF–CM66
Bac_Nitrospiria
Bac_Myxococcia
Bac_Gitt–GS–136
Bac_TK10
Bac_Holophagae
Bac_AD3
Bac_Acidimicrobiia
Bac_Phycisphaerae
Bac_KD4–96
Bac_Methylomirabilia
Bac_Anaerolineae
Fungi
Fng_Orbiliomycetes
Fng_Saccharomycetes
Fng_Ustilaginomycetes
Fng_Microbotryomycetes
Fng_Geminibasidiomycetes
Fng_Zoopagomycetes
Fng_Olpidiomycetes
Fng_Chytridiomycetes
Fng_Pezizomycetes
Fng_Tremellomycetes
Fng_Leotiomycetes
Fng_Agaricomycetes
Fng_Dothideomycetes
Fng_Eurotiomycetes
Fng_Mortierellomycetes
Fng_Sordariomycetes
Others
Fng_Cystobasidiomycetes
Fng_Blastocladiomycetes
Fng_Schizosaccharomycetes
Fng_Entomophthoromycetes
Fng_Archaeorhizomycetes
Fng_Kickxellomycetes
Fng_Exobasidiomycetes
Fng_Endogonomycetes
Fng_Atractiellomycetes
Fng_Pucciniomycetes
Fng_Wallemiomycetes
Fng_Mucoromycetes
Fng_Basidiobolomycetes
Fng_Glomeromycetes

831

832 **Extended Data Fig. 1 | Microbial taxonomic compositions at the genus, order, and class**
833 **levels**. **a**, Genus-level taxonomic compositions of prokaryotes and fungi. See Figure 1a for the
834 map of the research sites. **b**, Order-level taxonomic compositions. **c**, Class-level taxonomic
835 compositions.
836

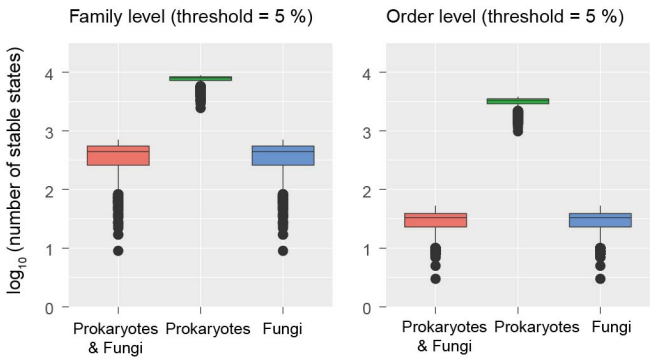


Extended Data Fig. 2 | Community structure and metadata properties of the samples. **a**, Prokaryote/fungal community structure and research-site profiles. Research sites are indicated by colors on the PCoA surface of prokaryote/fungal community structure. **b**, Soil taxonomy profile. **c**, Former crop plant.

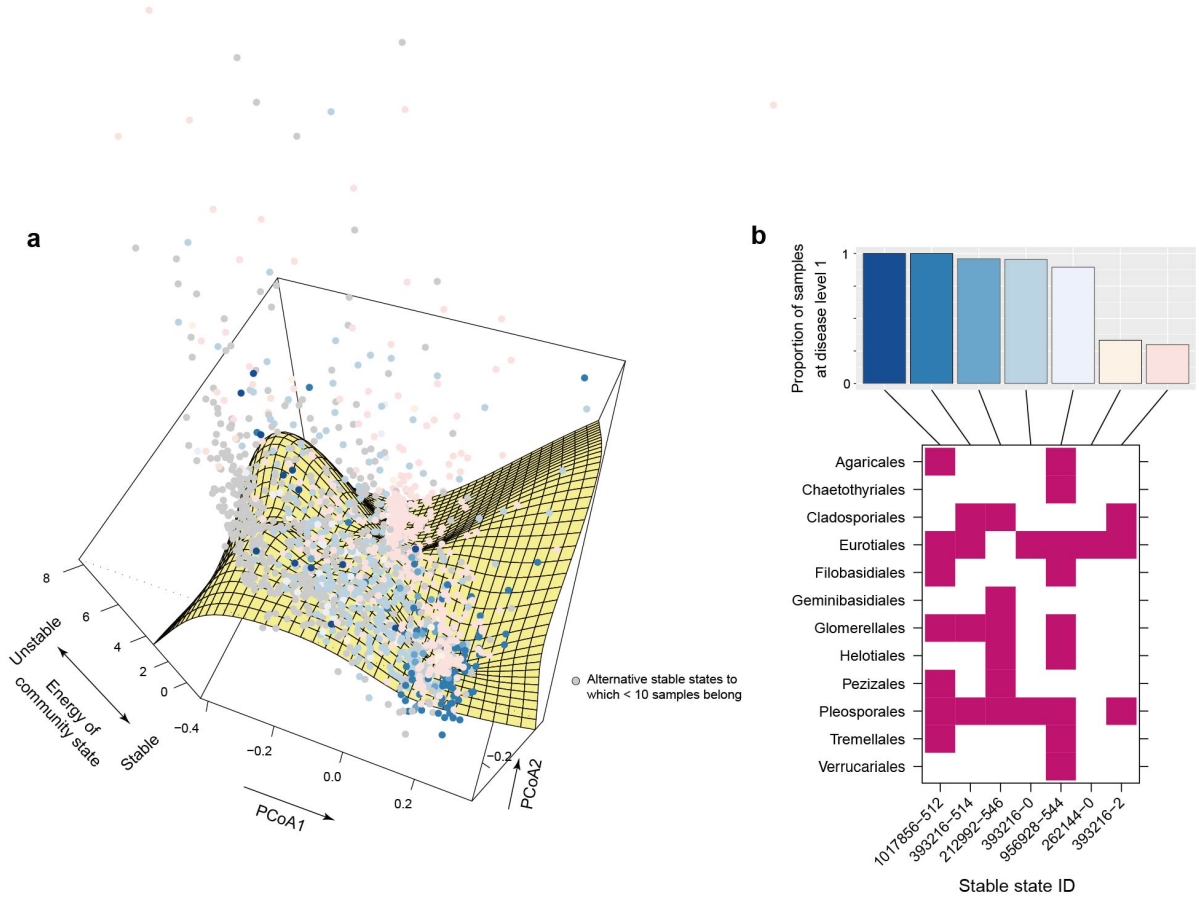

a
Prokaryotes
Fungi
pH
8
7
6
5
PCoA 2
PCoA 1
b
Prokaryotes
Fungi
Electrical conductivity (dS/m)
2.0
1.5
1.0
0.5
0.0
PCoA 2
PCoA 1
c
Prokaryotes
Fungi
Availabple P (mg P_2O_5/kg)
5000
4000
3000
2000
1000
PCoA 2
PCoA 1
d
Prokaryotes
Fungi
C/N ratio
20
15
10
5
PCoA 2
PCoA 1

843

Extended Data Fig. 3 | Community structure and soil chemical properties. **a**, Prokaryote/fungal community structure and soil pH of the samples. **b**, Electrical conductivity. **c**, Available phosphorous concentration. **d**, Carbon to nitrogen ratio.

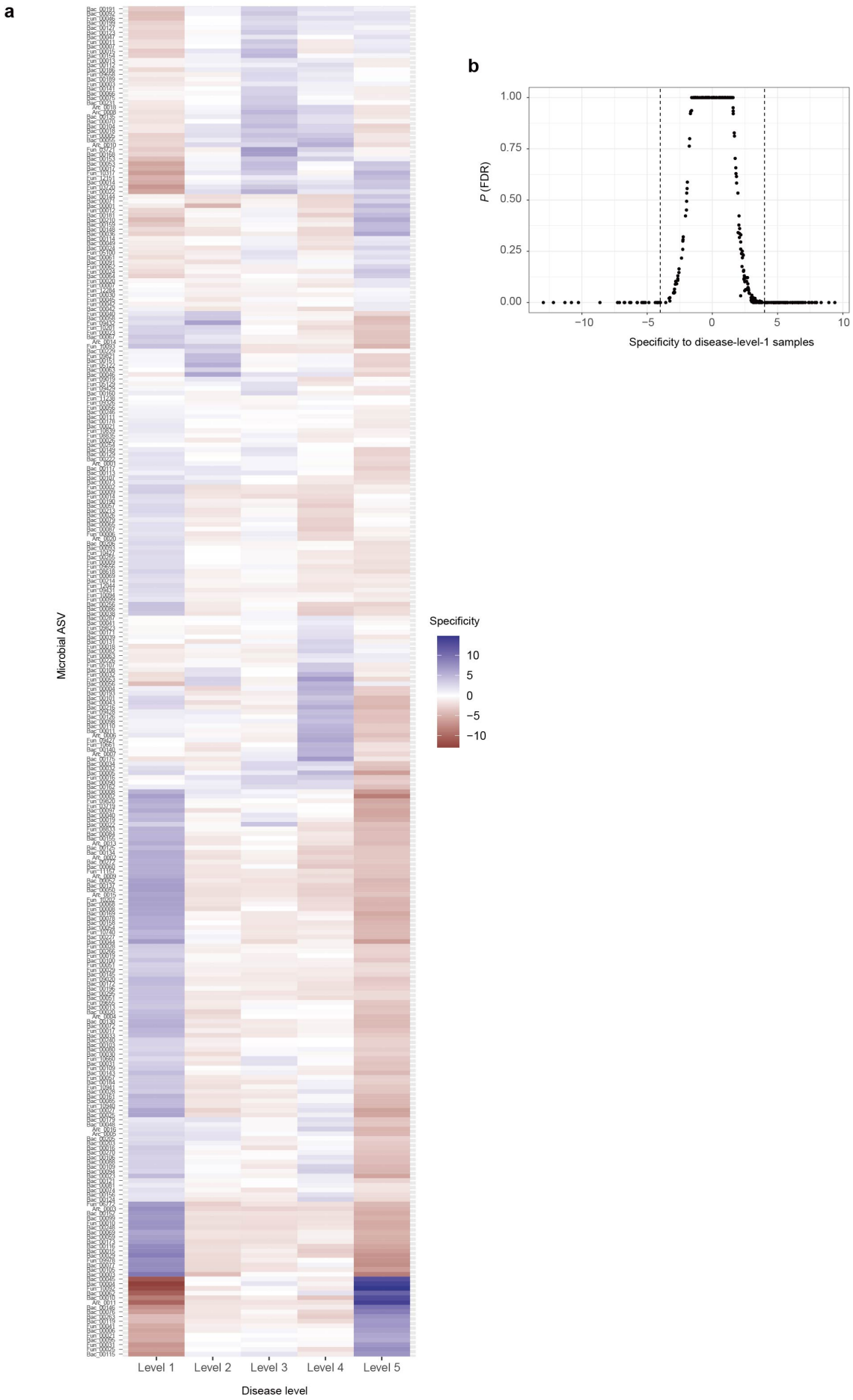


Extended Data Fig. 4 | Number of the stable states of community structure (threshold for binarization = 5 %). The parameters of the extended pairwise maximum entropy model were adjusted to fit the model to the empirical data (see Methods). The number of stable states in the energy landscape was then estimated by using the gradient descent algorithm for 20,000 randomly generated community states within the 2^S possible community states, where S denotes the number of families or orders. In the heuristic search, sets of values of the environmental variables (soil pH, electrical conductivity, C/N ratio, and available phosphorous concentration) were drawn from those of sampling points. The boxplots show variation in the number of stable states across different sets of environmental variables. A box represents the first quartile, median and the third quartile, while bars and circles indicate 1.5 × inter-quartile range and outliers, respectively. The analysis was performed, respectively, for the dataset representing both prokaryotes and fungi, the prokaryote dataset, and the fungus dataset. The threshold value in the binarization of the community data was set at 10 %: see Extended Data Figure 4 for results based on another threshold value (5 %).



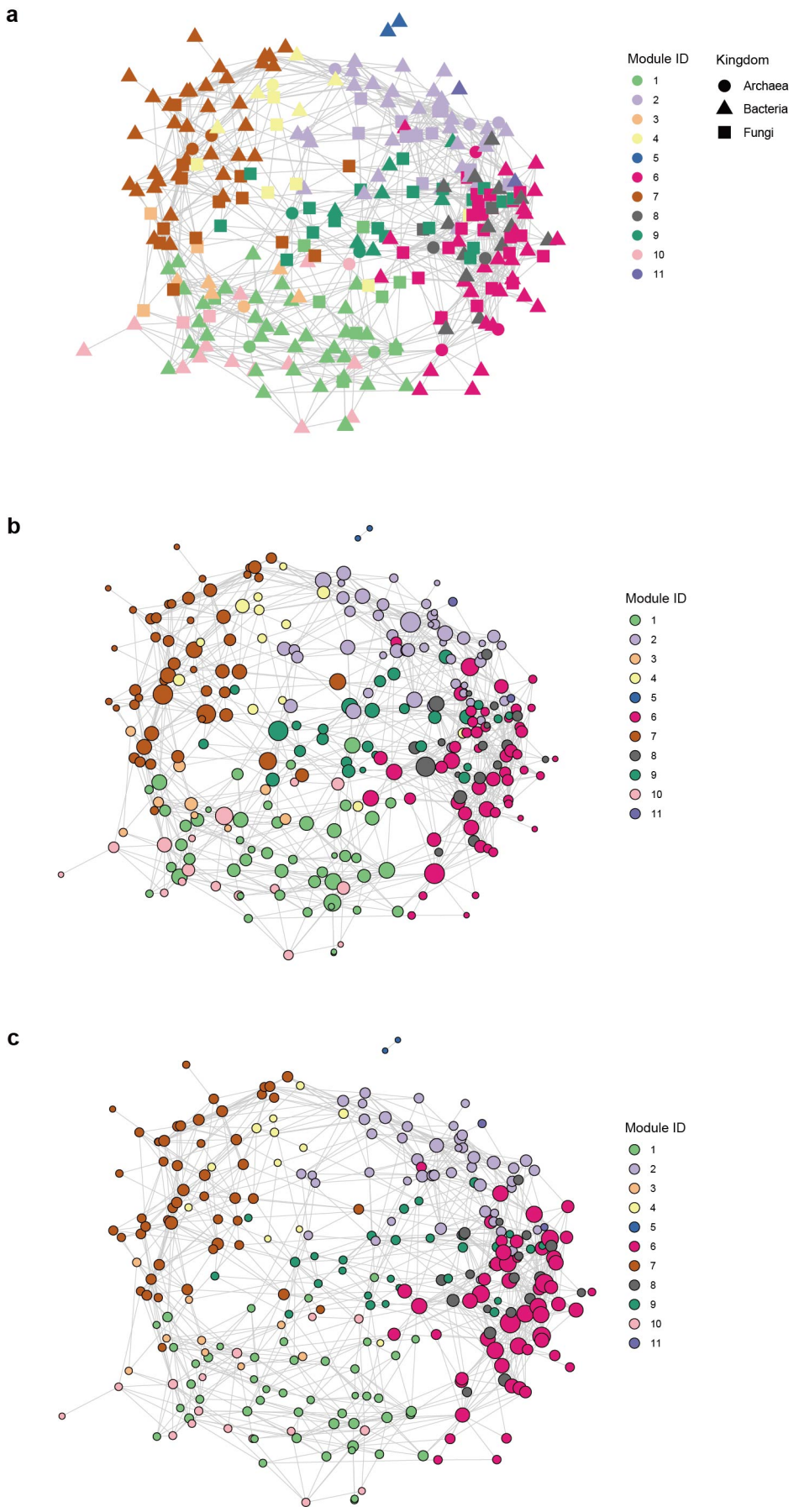
Extended Data Fig. 5 | Alternative stable states defined for order-level fungal data (threshold for binarization = 5 %). **a**, Energy landscape of fungal community structure. The surface of energy levels was reconstructed on the PCoA dimensions (Fig. 2) of fungal community structure based on spline smoothing. Community states with lower energy are inferred to be more stable. Data points (samples) indicated by the same color belong to the basins of the same alternative stable states, whose energy is lower than any adjacent community structure. See panel **c** for details about the alternative stable states. **b**, Definitions of the stable states. To reduce the complexity of input data to a computationally feasible

873 range, alternative stable states were explored based on order-level community compositions
874 (threshold for binarization = 5 %). The 7 alternative stable states representing ≥ 10 samples
875 were compared in terms of their associations with crop disease conditions.
876


a
Microbial ASV
Level 1
Level 2
Level 3
Level 4
Level 5
Disease level
Specificity
10
5
0
−5
−10
b
P (FDR)
1.00
0.75
0.50
0.25
0.00
−10
−5
0
5
10
Specificity to disease-level-1 samples

877

878 **Extended Data Fig. 6 | Each ASV's specificity to disease levels**. **a**, Standardized specificity
879 to disease level (s_{ij}). Based on a randomization approach, each ASV's specificity to samples
880 differing in crop disease levels was evaluated. A higher value of the specificity index
881 indicates that a microbial ASV displayed higher abundance in samples of the target disease
882 level than that expected by chance. **b**, Specificity scores and statistical significance. An s_{ij}
883 value higher than 4 or that lower than -4 indicates statistically strong specificity of an ASV
884 to a target disease level.
885


a
Module ID
1
2
3
4
5
6
7
8
9
10
11
Kingdom
Archaea
Bacteria
Fungi
b
Module ID
1
2
3
4
5
6
7
8
9
10
11
c
Module ID
1
2
3
4
5
6
7
8
9
10
11

886

887 **Extended Data Fig. 7 | Network centrality scores**. **a**, Distribution of archaea, bacteria, and
888 fungi within the co-occurrence network. **b**, Each ASV's betweenness centrality score within
889 the network. The node size represents relative betweenness centrality scores. **c**, Each ASV's
890 eigenvector centrality score within the network. The node size represents relative eigenvector
891 centrality scores.