bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A global survey of host, aquatic, and soil microbiomes reveals shared abundance and genomic features between bacterial and fungal generalists

Daniel Loos**[1] Ailton Pereira da Costa Filho**[2] Bas E. Dutilh[3,4]
Amelia E. Barber[2,*] Gianni Panagiotou[1,5,*]

[1] Department of Microbiome Dynamics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute. Jena, Germany.

[2] Junior Research Group Fungal Informatics, Institute of Microbiology, Friedrich Schiller University. Jena, Germany.

[3] Institute of Biodiversity, Friedrich Schiller University. Jena, Germany.

[4] Theoretical Biology and Bioinformatics, Utrecht University. Utrecht, the Netherlands.

[5] The University of Hong Kong. Hong Kong SAR, China.

[*] Correspondence: Amelia E. Barber <amelia.barber@uni-jena.de>, Gianni Panagiotou <gianni.panagiotou@leibniz-hki.de>

** Equal contribution

Abstract

Environmental change coupled with alteration in human lifestyles are profoundly impacting the microbial communities that play critical roles in the health of the earth and its inhabitants. To identify bacteria and fungi that are resistant and susceptible to habitat changes

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respectively, we retrieved paired 16S and ITS rRNA amplicon sequence data from 1,580 host, soil, and aquatic samples and explored the ecological patterns of the thousands of detected bacterial and fungal genera. Through this large-scale analysis, we identified 48 bacterial and 4 fungal genera that were prevalent and abundant across the three biomes, demonstrating their resilience in diverse environmental conditions. These generalists comprised a substantial fraction of the taxonomic diversity of their respective kingdom. Their distribution across samples explained a large percentage of the variation in the cross-kingdom community structure. We also found that the genomes of these generalists were larger and encoded more secondary metabolism and antimicrobial resistance genes, illuminating how they can dominate diverse microbial communities. Conversely, 30 bacterial and 19 fungal genera were only found in a single habitat, suggesting they cannot readily adapt to different and changing environments. These findings can contribute to designing microbiome-mediated strategies for pressing global changes.

Significance

Humans, plants, and aquatic and soil environments are home to a collection of microorganisms, known as their microbiome. The microbial communities are becoming increasingly perturbed by environmental change and changes in their associated hosts. Through the analysis of 1,580 microbiomes, we identify a small number of bacteria and fungi that can achieve high abundance in diverse host, aquatic, and soil environments, demonstrating their resilience to variable environments. These microbes contribute to microbiome diversity and more frequently engage in positive interactions with other microbes. We also identify a subset of bacteria and fungi whose environmental distribution was extremely limited, suggesting that they are vulnerable to change and may not be able to survive large shifts in their habitat.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Environments, plants, and animals are colonized with communities of microbial organisms, termed the microbiome that play critical roles in the function and health of their hosts and habitats. However, environmental change and alterations in host lifestyle are profoundly affecting these microbial consortia. Westernized diets low in fiber and rich in saturated fats and sugars, have decreased the abundance of beneficial microbes and been linked with myriad health conditions, including obesity, type 2 diabetes, and inflammatory bowel disease (1–4). Changes in marine environments due to climate change have induced major shifts in marine food webs, primary productivity, and carbon export (5–8). Additionally, anthropogenic climate change is resulting in net carbon loss in soil and changes in microbial community composition (9).

Ecological theory predicts that generalists, or organisms that are fit across a wider range of conditions, will be more resilient to changing environmental conditions (10, 11). Conversely, specialists, or organisms that are adapted to thrive in very specific environments, will be less able to withstand perturbations to their habitat. We are currently lacking a comprehensive understanding of the capacity of individual bacterial and fungal taxa to adapt to changing environmental conditions. However, this is crucial as those unable to change are susceptible to biodiversity loss, while those that can grow in a wider range of conditions may survive and flourish with unknown consequences. To this end, we performed a large-scale analysis of community sequencing data sets from host, soil, and aquatic environments with paired bacterial and fungal characterization to shed light on the ecological properties of the genera present and their putative resilience to change. We focused on three aspects: (i) the identification of bacteria and fungi that occurred in diverse environments capable of adapting to diverse environments (generalists) or were limited to highly specific environments (specialists); (ii) the relative abundance of bacterial and fungal generalists and specialists as a marker for their fitness and competitive colonization potential; and (iii) whether their presence in a habitat might trigger global changes in inter- and cross-kingdom population structure.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Environmental specificity of bacterial and fungal communities

For a global survey of bacteria and fungi across microbial communities, we analyzed paired 16S and ITS rRNA amplicon sequence data from 1,580 samples deposited in public databases. Samples were collected from Europe, Asia, and the Americas between 2010 and 2018 (Figure 1A). For cross-biome comparisons, samples were classified as aquatic, host, or soil environments based on the habitat they were collected from. This broad grouping is supported by principle coordinate analysis (PCoA) based on Bray-Curtis dissimilarity showing that samples from each environment largely cluster with each other and distinct from the other environments (Figure 1B)—a finding mirrored by a recent study of 22,700 bacterial microbiomes (12).

Of the 1,580 samples that we analyzed, 871 originated from soils, 494 from hosts (both mammalian and non-mammalian), and 215 from aquatic environments. The habitats that contributed largest number of samples for each environment were temperate (N=498) and conifer forests (N=147) for the soil, gut (N=287) and skin (N=68) for the hosts, and large lakes (N=87) and other freshwater (N=71) for the aquatic environments (see Supp. File 1 for details of all projects). Taxonomic profiling of bacterial and fungal communities was performed using the SILVA and UNITE databases, respectively. Rarefaction curves of each habitat indicated that most projects adequately captured the diversity of both the bacterial and fungal communities (Figure 1C). In total 2,977 bacteria and 1,740 fungal genera were detected across all samples (Figure 1D). We next examined the overlap of genera between environments, where a genus was considered shared if it was detected in at least one habitat in each of the three different environments (host, aquatic, and soil). For bacteria, soil and aquatic environments had the highest number of shared genera (N=1,662), followed by host-soil (N=1,483) and host-aquatic (1,226). The pattern was different for fungi, with host-soil sharing the most (N=884), followed by aquatic-soil (N=205) and host-aquatic (N=189). These trends remained after controlling for the different number of samples across the three

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

environments in 842 and 998 out of 1,000 random down sampled subsets for bacteria and fungi, respectively. Finally, we also confirmed that a similar degree of overlap between the environments was observed for different 16S and ITS amplicons, as well as significant correlations in the abundances of individual genera (Supp. Figure 1A-C).

While 40% of the total bacterial genera were found in all three environments, the percentage dropped to only 11% for fungal genera, indicating a higher degree of environmental specificity (Figure 1E & F). The most prevalent higher order taxonomic ranks that were detected in all three environments were *Proteobacteria* followed closely by *Firmicutes* for bacteria, and *Ascomycota* for fungi. For both bacteria and fungi, soil was the environment with the highest percentage of uniquely detected genera (i.e. genera not detected in any sample from host or aquatic origin), with 23% and 38%, respectively for each two kingdom. While aquatic-specific bacteria accounted for 7% of the total number of detected genera, the percentage of unique fungi in aquatic samples was only 2% (Figure 1 E & F). The opposite trend was observed for host-associated microbes, with only 3% and 8% of unique bacteria and fungi, respectively in this environment.

We subsequently compared the relative abundance of genera that were found in all environments or were uniquely detected in soil-, host- or aquatic-associated environments. Bacterial genera detected in all three environments were significantly more abundant (Wilcoxon Rank Sum test, $p < 0.001$) than genera uniquely detected in one of the environments (Figure 1G). A similar pattern was observed with fungi. However, a notable exception was the relatively high abundance of fungi that were uniquely detected in aquatic samples. Genera of aquatic fungi were more abundant than either common genera and uniquely detected in soil- or host-associated environments (Figure 1H). This observation was also robust across the different 16S and ITS regions used in the dataset (Supp. Figure 1D). Taken together, we find that soil bacteria and fungi show a higher degree of biome specificity that microbes in other biomes and that genera commonly detected in all environments were also more abundant in these microbial communities.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bacterial and fungal generalists are more abundant than specialists and have distinct genomic features

Generalists and specialists play important, yet distinct roles in ecosystems. However, objectively identifying them has proven challenging. To define multi-kingdom generalists and specialists, we set the following criteria: generalists are genera found with high prevalence (>40%) in at least one habitat from each of the three environments (host, aquatic, soil). Conversely, specialists are genera with a high prevalence (>40%) in one habitat and low prevalence (<5%) in every other. A relative abundance (RA) cut-off of 0.01% was used for determining whether a genus was present in a sample. Using this approach, we detected 48 bacterial generalists and 30 specialists (Figure 2A; Supp. Table 1).

To confirm our definition of generalists and specialists, we calculated Levins' niche breadth indices (B_n), which measures taxon distribution across environments and where higher values indicate even distribution across environments (13). Generalists showed significantly higher B_n values than specialists (Wilcoxon Rank Sum test, $p < 0.001$; Supp. Figure 2A). All specialists and all generalists, with the exception of the *Christensenellaceae R7 genus*, were above the detection limit and had a significant Levins' niche breadth signal after Benjamini-Hochberg adjustment (13). As our criteria for defining generalists and specialists was reliant on human-defined biome annotations, we further validated our approach by comparing it to the recently developed social niche breadth (SNB) score (12). By comparing the similarity or diversity of microbial communities where a given genus occurs, SNB provides a data-driven score independent of biome annotations based on an independent dataset of over 22,500 bacterial microbiomes (12). Indeed, the generalists identified in our study had significantly higher SNB scores than the bacterial specialists we identified (Wilcoxon Rank Sum test, $p < 0.001$; Supp. Figure 2B).

We observed multiple phylogenetic origins for both generalists or specialists (Chi squared test, $p > 0.05$), indicating their roles as generalists and specialists evolved independently (Supp. Figure 3). Each of the top five bacterial generalists were detected in more than 50% of

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the 1,580 samples. Among them, the most prevalent was *Pseudomonas* which was detected in 52%, 70% and 89% of host, soil and aquatic samples respectively, followed by *Bacillus* (33%, 71%, 35%) and *Bradyrhizobium* (17%, 73%, 35%). The most extreme bacterial specialists came from the soil. While *Gryllotalpicola* and *Anaerovibrio* were found in >91% of biochar samples, the prevalence dropped to 0.1% on average for non-soil environments (Supp. Table 1). Specialists were also found in host- and aquatic-associated environments. For example, *Acetatifactor* was found in 80% of samples from the murine gut, but had a prevalence of <3% in all other habitats. The genus *Leptospira* was found in 83% of samples of the Cuyahoga River, but had a prevalence less than 2% in all other habitats. Interestingly, when comparing the relative abundance of generalists and specialists, we observed that both bacterial and fungal generalists had a significantly higher abundance (Figure 2B, Wilcoxon Rank Sum test, $p < 0.001$). This pattern remained when we used stricter and looser thresholds to define generalists and specialists (Supp. Figure 4). This finding confirms the pattern observed above (Figure 1G, H), suggesting that independently of how groups are defined, genera that can colonize diverse environments are usually able to outcompete niche-specific genera.

When looking at the fungal kingdom, the number of generalists was much lower and only *Aspergillus*, *Malassezia*, *Aureobasidium*, and *Cortinarius* satisfied the criteria of a generalist (Supp. Table 1). Among these, *Aspergillus* had the highest overall prevalence among all samples with 38%, 52% and 12% in the host, soil, and aquatic samples, respectively. From the 19 fungal specialists, *Chrysanthotrichum* and *Mycocentrospora* were the most habitat-specific, with prevalences of 68% and 48% in temperate and conifer forests, respectively, but a mean prevalence of 0.1% in all other habitats. Only two of the 19 fungal specialists (11%) originated from outside soil environments (*Vuilleminia* and *Seimatosporium* from plants). As with bacterial genera, the relative abundance of fungal generalists was significantly higher than that of fungal specialists (Figure 2A, Wilcoxon Rank Sum test, $p < 0.001$).

To gain insight into how generalists predominate microbial communities in abundance, even across diverse environments, we analyzed the genomes belonging to generalists and specialist genera available on NCBI (see Methods for details on genome selection). For bacteria, when analyzing the genomes of 2,328 generalists and 471 specialists, the generalists had

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significantly larger genomes, as measured by the number of coding sequences (CDS) with a mean of 4,671 CDS for generalists and 3,189 for specialists. (Figure 2A, Wilcoxon Rank Sum test, $p < 0.001$). As secondary metabolism genes are often used by microbes during competition for resources and as chemical warfare in crowded environments, we examined the genomes of generalists and specialists for the presence of biosynthetic gene clusters (BCGs). Strikingly, the genomes of bacterial generalists encoded almost double the number of BCGs with an average of 5.0 compared to 2.4 for specialists (Figure 2C). Further differentiating bacterial generalists, they also contained significantly more antimicrobial (AMR) and stress-resistance genes, with an average of 4.9 AMR and 4.8 stress genes compared to 1.9 and 1.5 for specialists, respectively (Figure 2D; Wilcoxon Rank Sum test, $p < 0.001$). For fungi, no significant differences in either the number of genes or the number of BCGs was observed (Figure 2B, C), likely due to the severe underrepresentation of publicly available fungal specialist genomes (N=5).

To explore intra and inter-kingdom interaction patterns and to gain further insight into the downstream effects of the observed differences between generalists and specialists, we constructed individual co-abundance networks for soil, host, and aquatic environments (see Methods for details). Despite only considering the 1,188 bacteria and 184 fungal genera commonly detected in all three environments, the topological characteristics of the networks for each environment were highly distinct, as measured by significant differences in betweenness and Kleinberg's hub node centrality scores (Wilcoxon test, $p < 0.001$; Supp. Figure 5). In spite of the differences in topology, we could still compile subnetworks of inter- and intra-kingdom correlations found jointly among host-soil, host-aquatic and/or soil-aquatic environments. Strikingly, 45 of the 48 bacterial generalists and all 4 fungal generalists were part of those subnetworks, which are characterized by a higher number of positive than negative edges (Figure 2E). The ratio of positive to negative edges was higher in correlations involving a generalist (2.5) compared to all other edges (2.2). When we looked for interactions between genera found in all three environments, we identified 43 such edges that all represented positive interactions between bacteria and included 21 generalists. Together these findings suggest that the success of generalists in colonizing diverse environments and achieving high abundances may be attributable to their ability carve out a niche for them-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selves using secondary metabolism and AMR genes and by eliciting positive interactions with other highly prevalent genera.

Bacterial generalists exert a strong influence on the intra- and inter-kingdom community structure

We subsequently explored whether the presence of generalists and specialists had an impact on the diversity of a community. Interestingly, alpha diversity, as measured as Chao 1 and Shannon, was significantly lower in samples where no generalist was detected compared to samples with generalists present for both bacterial (Figure 3A) and fungal (Figure 3B) communities (Permutation test of samples lacking any of the N generalists compared with samples lacking any N random taxa, $1*10^4$ permutations, $p < 0.03$). Conversely, the impact of specialists on alpha diversity in their specific habitat was much less profound and varied by habitat without a clear trend (Supp. Figure 6).

We subsequently shifted our focus to inter-kingdom interactions which are often overlooked in microbial ecology studies and examined bacterial generalists for a role in shaping the mycobiome community structure and vice versa. As expected, we observed a significant separation between the soil, host, and aquatic micro- and mycobiome beta diversity by Bray-Curtis dissimilarity (Figure 3C, PERMANOVA, $p < 0.001$ for both bacteria and fungi). Constrained ordination revealed a significant, linear relationship between bacterial Bray-Curtis dissimilarity and fungal community composition and vice versa (Distance-based redundancy analysis, abbreviated dbRDA, ANOVA $p < 0.04$ for all explanatory genera of the other kingdom in a multivariate model). Bacteria genera could explain an extensive part of the mycobiome variation observed in the three environments with a partial R2 of 25% by dbRDA. Of the bacterial genera, *Conexibacter*, *Bacillus*, and *Lysobacter* had the highest explanatory power on mycobiome variation (Fig 3D). Interestingly, six out of the top ten explanatory bacteria genera in the dbRDA were generalists. Similarly, fungal genera explained 26% of the microbiome variation between host, soil, and aquatic samples, with *Mortierella*, *Trichocladium*, and *Candida* having the highest explanatory power (Fig 3D). Among the top ten explana-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tory genera was one of the four fungal generalists - *Malassezia*. Altogether, our analysis indicates that bacterial and fungal generalists profoundly impact microbial communities by contributing positively to the taxonomic (alpha) diversity of their kingdom- an ecological characteristic often associated with healthy environments, and they can also contribute to shaping cross-kingdom microbial structures.

Discussion

Recent grant global changes are profoundly affecting the health of the earth we live on and its inhabitants (14–16). As environmental and host-associated microbial communities become increasingly exposed to our changing world, we are still lacking knowledge regarding the capacity for millions of bacterial and fungal species to cope with these shifts. With this in mind, we performed a large-scale global survey of host, aquatic, and soil microbiomes to reveal ecological and genomic properties of bacterial and fungal genera that may promote or limit their establishment in new environments and how they contribute to the richness and diversity of an environment. The metagenomic analysis of 1,580 paired host, soil, and aquatic microbiomes and mycobiomes identified approximately 3,000 bacterial and 1,700 fungal genera. Using cutoffs selected by data-driven approaches, we identified ~70 specialist genera whose limited distribution suggests they may struggle in different or changing habitats and identify ~50 widely abundant genera with a clear ability to thrive in many environments.

While the concept of generalists and specialists in ecology is not new, it has mostly been applied in specific habitats (17–21) and not on a global scale. Although some studies on generalist and specialist microbes have appeared over the years (12, 18, 22–26), these have rarely included eukaryotic microorganisms such as fungi. Moreover, these studied have not investigated cross-kingdom biotic interactions shaping microbial communities. We demonstrate that both bacterial and fungal generalists share ecological features, including the ability to reach significantly higher abundances than specialists and contributing positively to the richness and diversity of their respective kingdom. Moreover, six bacterial generalists, including *Bacillus*, *Lysobacter*, *Escherichia* and *Gemmatimonas* and one fungal generalist,

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Malassezia, harbor additional ecological properties and appear to play a significant role in shaping cross-kingdom microbial composition (Figure 3D).

Our global survey of bacterial-fungal communities has generated a valuable list of genera containing organisms that may be susceptible to biodiversity decline and even extinction under changing environmental threats (27, 28). Conversely, the identified generalist bacteria and fungi are highly resilient against environmental perturbations and may even be considered as targets for microbiome engineering, where their ability to flourish in highly diverse environments and contribute to richness is a desirable trait. Their beneficial ability to thrive in diverse communities may explain the fact that they carry an enhanced arsenal of antimicrobial resistance genes (Figure 2D). One challenge ahead will be moving the analysis of generalists and specialists to the species and strain level to understand the functional characteristics that differentiate generalists from other microbes. Currently, taxonomic classification of bacteria and fungi to the species and strain level is inadccurate using amplicon metagenomics (29–31), so this was not addressed in our study. Moreover, the species- and strain-level diversity of the microbial world is enormous so many more samples would be required to gain a comprehensive overview of its generalists and specialists. One way forward that can be likely explored is by using deep functional characterization at the pathway and enzyme level using shotgun metagenomics datasets, especially for bacteria. The functional characterization of fungal generalists may prove to be a much greater challenge, as the tools for functional prediction based on metagenomic data lag behind prokaryotic microorganisms. Nevertheless, we believe that large-scale computational analyses combined with laboratory experiments in cross-disciplinary approaches will be able to overcome these challenges and address the many open questions about microbial niche range and its consequences for microbial extinction and global biodiversity loss.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and methods

Sample selection

Included studies were retrieved by querying NCBI BioProject with the terms 'bacteria' and 'fungi' in any field. Only Biosamples with both 16S rRNA and ITS amplicon sequencing data were considered for the concurrent analysis of both kingdoms. We used both the identifier and attributes of the biosample, such as aliases and library names, to map fungal and bacterial read files to a sample using a custom script. Samples were associated to an environment (aquatic, host, or soil) using manual curation of associated publications and biosample attributes provided by the depositor. The three environments were further subdivided into 17 habitat groups based on the body part and/or the ecoregion of the sampling location for host and other samples, respectively (32). Habitats with less than five samples were pooled together.

Generation of genus-level abundance profiles

Genus-level abundance profiles were calculated using a custom nextflow pipeline (33). Briefly, reads were downloaded from NCBI SRA using grabseqs, except for the American Gut Project, which was downloaded from Qiita (34, 35). Paired-end reads were merged using NGmerge (36). Quality Control (QC) and adapter removal was performed using trimmomatic with a minimum Phread quality of 20 and a minimal read length of 100 (37). Quality was assessed using FastQC and MultiQC (38). Subsequent steps were performed using QIIME2 (39). Reads were dereplicated following closed-reference OTU picking for both kingdoms separately using VSEARCH with a 97% identity threshold (40). For taxonomic annotation, SILVA 132 97% consensus and UNITE 8.2 dynamic databases were used for bacteria and fungi, respectively (41, 42). Following quality control, a total of 1,580 samples were selected for downstream analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discovery of sample rRNA amplified region

Multiple rRNA regions were used to characterize microbial diversity as the study dataset is composed of many sequencing projects. When available, the specific rRNA region ampliefied was obtained from deposited metadata or linked publication. For BioProjects where this information was not available, the following was performed. As the SILVA database (v138.1) contains full length bacterial rRNA sequence, the hypervariable regions (e.g. V1-V3, V4-V5) from each taxa was extracted using the in silico pcr tool (https://github.com/egonozer/in_silico_pcr) with primers described in (43).Amplicon sequence data from each project was then aligned to each variable region using BWA-MEM v.0.7 and contig coverage quantified using BBTools v.39.01. The 16S variable region with the highest percent coverage was taken as the region amplified in the study. For the ITS amplicon data, ITSx1.1.13 (44) was used to extract the ITS1 and/or ITS2 consensus from sequence reads. The BioProject primers identified through this analysis, as well as those retrieved from association publications is listed in Supp. File 1.

Abundance correlation between varying rRNA amplions

To calculate the correlation in genus abundances between the differing rRNA regions amplified, genera that were detected in all three environments were considered and samples aggregated into whether they included sequence from the V1-V4 regions or V4-V5 regions for bacteria and ITS1 or ITS2 for fungi. For each rRNA category, Pearson's correlation coefficients were calculated for genus abundance in each environment. The similarity between the correlation matrices (V1-V4 and V4-V5 for bacteria and ITS1 and ITS2 for fungi) was then calculated by transforming the upper triangle of each correlation matrix into a vector and calculating the correlation coefficient between the two.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Workflow and statistical analysis

Analyses were performed using a custom drake pipeline (45) built using the programming language R 4.0.2. Briefly, abundances obtained from OTU profiling were total-sum-scaled (TSS) and pooled at genus rank. All tools were used with default parameters if not explicitly specified.

Diversity

Alpha diversity was estimated using Shannon and Chao1 metrics with the phyloseq and vegan packages (46, 47). To quantify the contributions of a bacterial community profile with the fungal one and *vice versa*, we used linear and unsupervised Canonical Correlation Analysis, as implemented in the function CCorA of the vegan R package (46). P-values were obtained using blocked permutations to control for the habitat and to reduce assumptions of the test. Supervised constrained ordination was performed using stepwise Distance-based Redundancy Analysis (dbRDA) adapted from (48). This analysis shows linear relationships between bacterial dissimilarities and abundances of selected explanatory fungal genera (and *vice versa*). An optimal subset of up to 50 explanatory genera of the other kingdom was computed using a stepwise feed-forward approach, as implemented in the ordistep function of the vegan R package (46).

Co-abundance networks

SparCC, as implemented in FastSpar, was used to assess correlation between taxa pairs for each environment separately (49, 50). Both kingdoms were pooled together, allowing for the identification of interkingdom correlations. Only genera found in all three environments were considered for pairwise correlation. Node topology metrics were calculated using the R package igraph.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Generalists and specialists

Genera were defined as generalists if they were found in at least 40% of samples in at least one habitat from each environment (host, soil, aquatic) with a relative abundance of at least 0.01%. Complementary, genera were defined as specialists if they were found in at least 40% of samples in one habitat and less than 5% of samples in all other habitats using the same abundance threshold as for generalists. Levins' niche breadth index was calculated as implemented in the R package MicroNiche (51).

Genome features of generalists and specialists

As amplicon sequence data is based on maker genes, deposited genomes were used to characterize functional traits associated with the genomes of generalists and specialists. The generalists and specialist genera were queried in NCBI. Of the resulting genome list, all genomes or up to 60 randomly selected genomes if more were available were selected for each genus. This resulted in genomes for 2,328 bacterial generalists, 117 fungal generalists, 471 bacterial specialists, and 5 fungal specialists. Genome size and number of coding regions was obtained from the NCBI metadata. For the calculation of the number and type of biosynthetic gene clusters in each genome, AntiSMASH v6.1.1 was used (52). Antimicrobial and stress resistance genes were predicted in bacterial genomes using AMRFinderPlus (53).

Code and data availability

Scripts created for data processing and statistical analysis are available at https://github.com/bioinformatics-leibniz-hki/its-16s. Raw sequence data can be downloaded from any International Nucleotide Sequence Database Collaboration (INSDC) server using accessions as provided in Supplemental Table 1-2 and the git repository.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 20151 – Project-ID 390813860. BED is supported by the European Research Council (ERC) Consolidator grant 865694 and the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship.

Figures

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

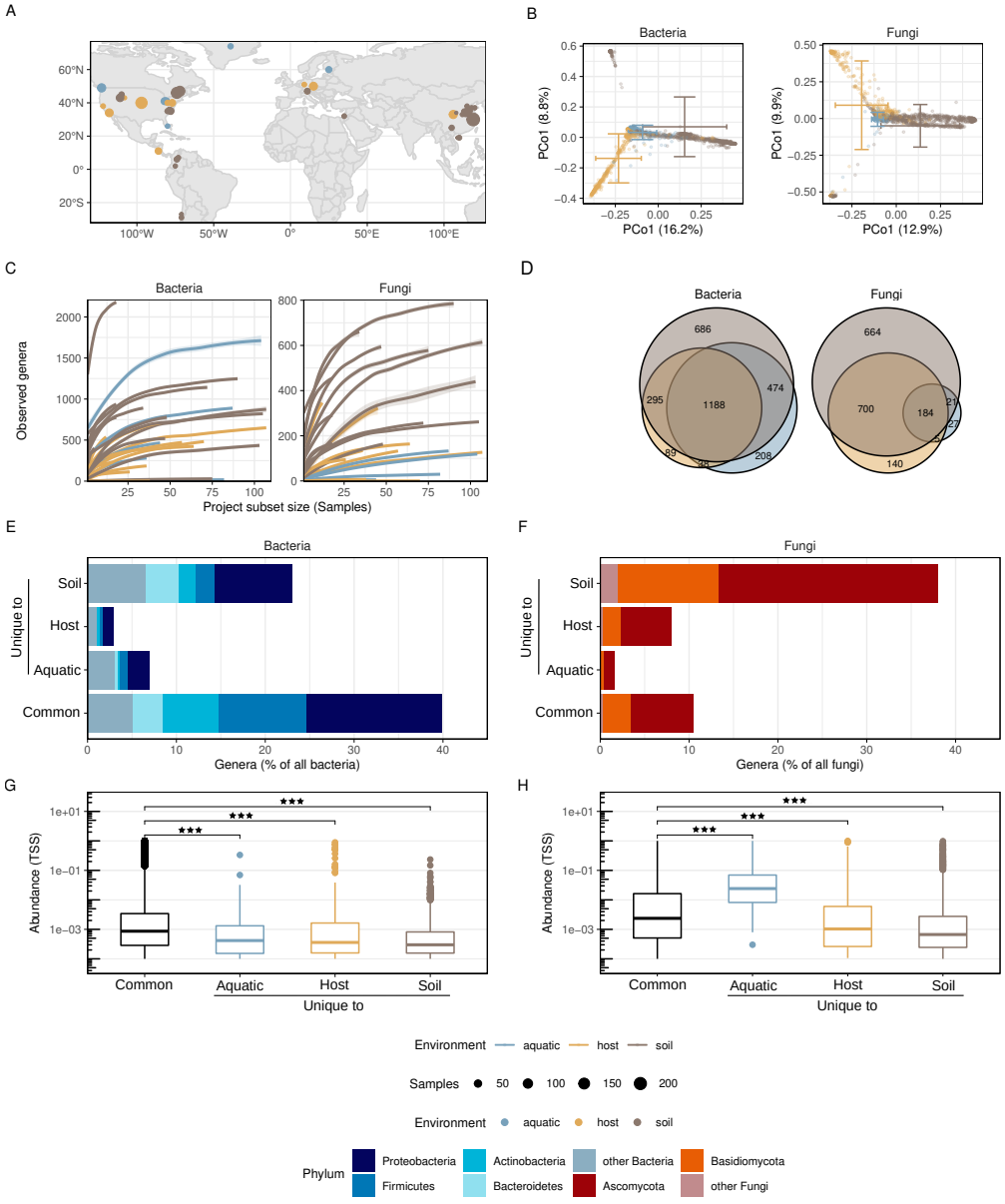

Figure 1: **A global analysis of microbial communities reveals differences in environmental specificities between bacteria and fungi.** (A) Distribution of samples used in this study (N=1,580) by geographic location. (B) Bray-Curtis dissimilarity between samples, colored by environment. Crosshatches represent the mean ± SD for each environment. (C) Rarefaction curves of Shannon alpha diversity for each study demonstrate sufficient sampling depth. Curves are shown as LOESS regressions from 10 independent sampling trials at 10 given sampling subset sizes. Lines are colored by environment and are surrounded by ribbons indicating the 95% confidence interval across the trails. (D) Intersection of bacterial and fungal genera found in at least one sample in each environment as Venn diagrams. (E,F) Percentage of genera found in all three or only one environment. (G,H) Abundance comparisons of common and unique genera by total sum scaling (TSS). A genus was considered present in a sample using a threshold of abundance > .01%. Significance determined by Wilcoxon rank sum test; *** denotes $p < 0.001$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

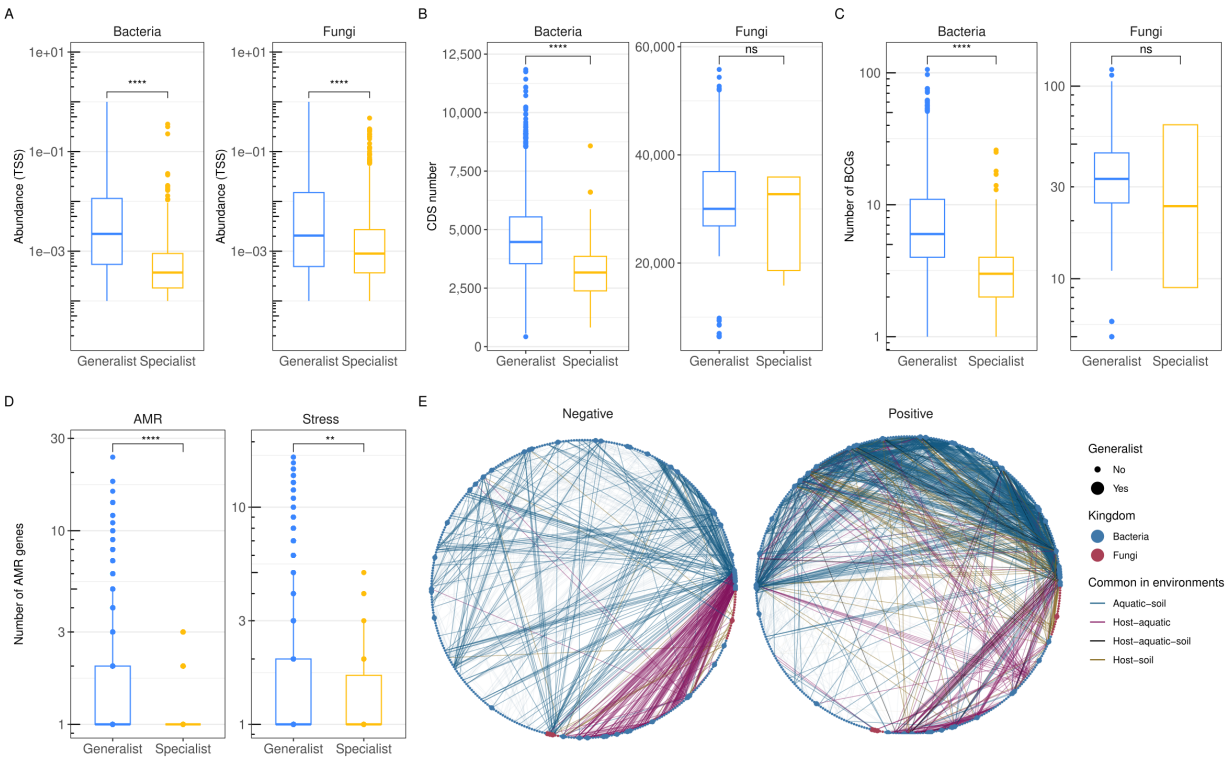

Figure 2: **Generalists are more abundant and bacterial generalists have larger genomes with more biosynthetic gene clusters and antimicrobial resistance genes.** (A) Relative abundances of bacterial and fungal generalists and specialists. Values were averaged by project to account for different cohort sizes. Statistical significance calculated using Wilcox rank-sum test (*** denotes $p<0.001$). (B-C) Number of coding sequences (CDS) (B) and biosynthetic gene clusters (BCGs) (C), in the genomes of generalists and specialists. Data from the genomes of 2,328 bacterial generalists, 117 fungal generalists, 471 bacterial specialists, and 5 fungal specialists. Statistical significance calculated by Wilcox rank-sum test (**** denotes $p<0.0001$). (D) Number of antimicrobial resistance (AMR) and stress genes in the genomes of bacterial specialists. (E) Networks of genera found in all three environments and significantly co-abundant in the majority of environments (SparCC FDR $p < 0.05$, $|r| > 0.2$).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

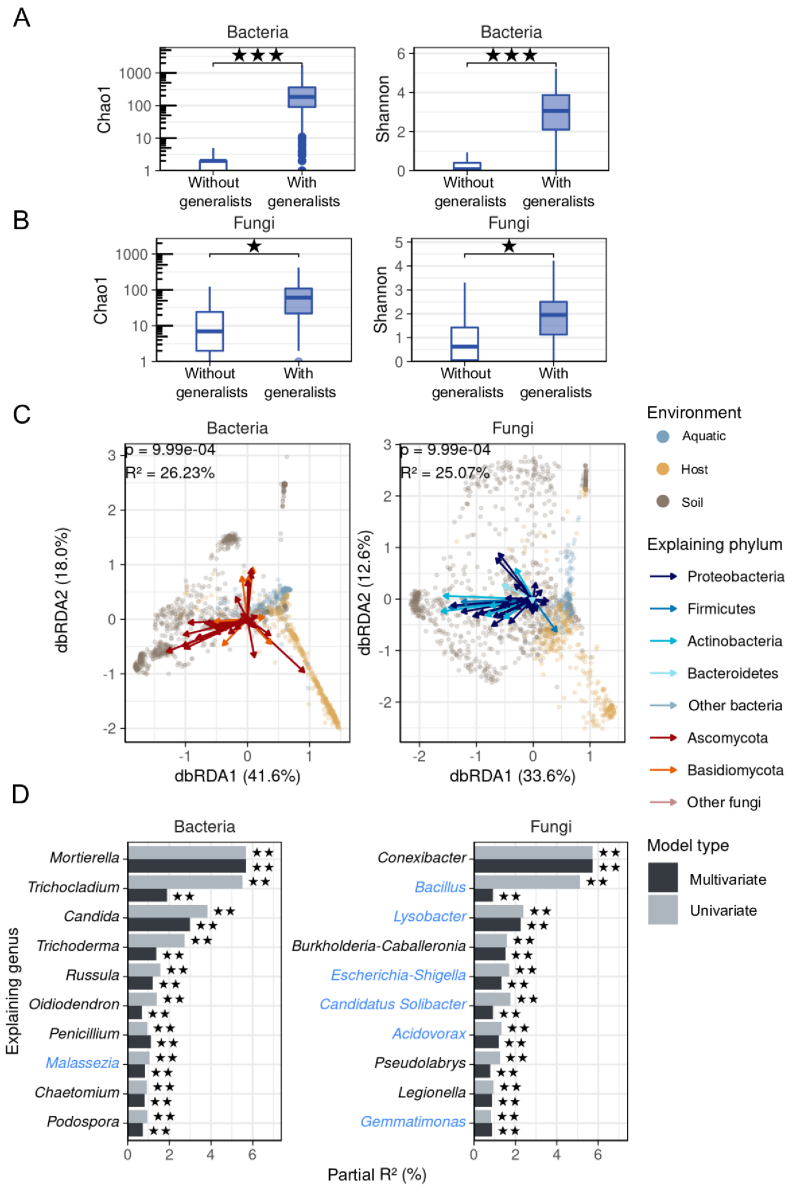

Figure 3: **Generalists significantly impact diversity and cross-kingdom variation.** A-B: Shannon and Chao1 alpha diversity were calculated for bacteria (A) and fungi (B). Samples were grouped by whether they contained any generalist (genera with >40% prevalence in at least one habitat from every environment; abundance > 0.01%), or not. Significance bars indicate permutation test compared to samples without random taxa instead of generalists (* q<0.05, ** q<0.01, *** q<0.001). C-D: Bacterial and fungal Bray-Curtis dissimilarities constrained by explanatory genus abundances of the other kingdom using distance-based Redundancy Analysis (dbRDA). C: Explaining genera were selected using a feedforward approach. Effect size of most explanatory taxa is shown in D by multivariate model (as displayed in C) or univariate model containing only the taxon of interest. Generalists are indicated with blue text. Stars indicate significance by ANOVA (** p<0.01).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. P. D. Cani, *et al.*, Changes in Gut Microbiota Control Metabolic Endotoxemia-Induced Inflammation in High-Fat Diet–Induced Obesity and Diabetes in Mice. *Diabetes* **57**, 1470–1481 (2008).

2. K. Sugihara, T. L. Morhardt, N. Kamada, The Role of Dietary Nutrients in Inflammatory Bowel Disease. *Frontiers in Immunology* **9**, 3183 (2019).

3. W. A. L. Kim Kyung-Ah AND Gu, High fat diet-induced gut microbiota exacerbates inflammation and obesity in mice via the TLR4 signaling pathway. *PLOS ONE* **7**, 1–11 (2012).

4. W.-Z. Li, K. Stirling, J.-J. Yang, L. Zhang, Gut microbiota and diabetes: From correlation to causality and mechanism. *World Journal of Diabetes* **11**, 293–308 (2020).

5. X. Guo, *et al.*, Threat by marine heatwaves to adaptive large marine ecosystems in an eddy-resolving model. *Nature Climate Change* **12**, 179–186 (2022).

6. I. A. G. Ullah Hadayet AND Nagelkerken, Climate change could drive marine food web collapse through altered trophic flows and cyanobacterial proliferation. *PLOS Biology* **16**, 1–21 (2018).

7. P. Brun, *et al.*, Climate change has altered zooplankton-fuelled carbon export in the North Atlantic. *Nature Ecology & Evolution* **3**, 416–423 (2019).

8. R. Cavicchioli, *et al.*, Scientists' warning to humanity: Microorganisms and climate change. *Nature Reviews Microbiology* **17**, 569–586 (2019).

9. J. M. Melillo, *et al.*, Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science* **358**, 101–105 (2017).

10. L. Wang Shenshen AND Dai, Evolving generalists in switching rugged landscapes. *PLOS Computational Biology* **15**, 1–21 (2019).

11. G. W. Gilchrist, Specialists and generalists in changing environments. I. Fitness landscapes of thermal sensitivity. *The American Naturalist* **146**, 252–270 (1995).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

12. F. A. B. von Meijenfeldt, P. Hogeweg, B. E. Dutilh, A social niche breadth score reveals niche range strategies of generalists and specialists. *Nature Ecology & Evolution* (2023) https:/doi.org/10.1038/s41559-023-02027-7 (April 17, 2023).

13. P. Feinsinger, E. E. Spears, R. W. Poole, A simple measure of niche breadth. *Ecology* **62**, 27–32 (1981).

14. S. S. Myers, *et al.*, Human health impacts of ecosystem alteration. *Proceedings of the National Academy of Sciences* **110**, 18753–18760 (2013).

15. L. Tedersoo, *et al.*, Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).

16. T. Newbold, *et al.*, Global patterns of terrestrial assemblage turnover within and among land uses. *Ecography* **39**, 1151–1163 (2016).

17. J. Liao, *et al.*, The importance of neutral and niche processes for bacterial community assembly differs between habitat generalists and specialists. *FEMS Microbiol Ecol* **92** (2016).

18. L. Liu, S. Wang, J. Chen, Transformations from specialists to generalists cause bacterial communities are more stable than micro-eukaryotic communities under anthropogenic activity disturbance. *Sci Total Environ* **790**, 148141 (2021).

19. C. E. Garrison, E. K. Field, Introducing a "core steel microbiome" and community functional analysis associated with microbially influenced corrosion. *FEMS Microbiol Ecol* **97** (2020).

20. J. Walter, R. Ley, The human gut microbiome: ecology and recent evolutionary changes. *Annu Rev Microbiol* **65**, 411–429 (2011).

21. Y. J. Chen, *et al.*, Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J* **15**, 2986–3004 (2021).

22. E. E. L. Muller, Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. *mSystems* **4**, e00080–19 (2019).

23. T. Thomas, *et al.*, Diversity, structure and convergent evolution of the global sponge microbiome. *Nature Communications* **7**, 11870 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24. L. A. Malard, M. Z. Anwar, C. S. Jacobsen, D. A. Pearce, Biogeographical patterns in soil bacterial communities across the arctic region. *bioRxiv* (2019) https:/doi.org/10.1101/655431.

25. M. Cobo-Simón, J. Tamames, Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC Genomics* **18**, 499 (2017).

26. M. Garcia-Garcera, E. P. C. Rocha, Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nature Communications* **11**, 758 (2020).

27. M. G. Weinbauer, F. Rassoulzadegan, Extinction of microbes: Evidence and potential consequences. *Endangered Species Research* **3**, 205–215 (2007).

28. S. Louca, *et al.*, Function and functional redundancy in microbial systems. *Nature Ecology & Evolution* **2**, 936–943 (2018).

29. J. S. Johnson, *et al.*, Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* **10**, 5029 (2019).

30. L. M. A. L. Poretsky Rachel AND Rodriguez-R, Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLOS ONE* **9**, 1–12 (2014).

31. L. Tedersoo, *et al.*, Best practices in metabarcoding of fungi: From experimental design to results. *Molecular Ecology* **31**, 2769–2795 (2022).

32. D. M. Olson, *et al.*, Terrestrial ecoregions of the world: A new map of life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience* **51**, 933–938 (2001).

33. P. Di Tommaso, *et al.*, Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).

34. A. Gonzalez, *et al.*, Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* **15**, 796–798 (2018).

35. L. J. Taylor, A. Abbas, F. D. Bushman, grabseqs: simple downloading of reads and metadata from multiple next-generation sequencing data repositories. *Bioinformatics* **36**, 3607–3609 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

36. J. M. Gaspar, NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 536 (2018).

37. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

38. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

39. E. Bolyen, *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852–857 (2019).

40. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

41. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590–596 (2013).

42. R. H. Nilsson, *et al.*, The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic acids research* **47**, D259–D264 (2019).

43. A. Ebou, D. Koua, A. Zeze, HyperEx: A tool to extract hypervariable regions from 16S rRNA sequencing data. *bioRxiv* (2021) https:/doi.org/10.1101/2021.09.03.455391.

44. J. Bengtsson-Palme, *et al.*, Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution* **4**, 914–919 (2013).

45. W. M. Landau, The drake r package: A pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software* **3**, 550 (2018).

46. P. Dixon, VEGAN, a package of r functions for community ecology. *Journal of Vegetation Science* **14**, 927–930 (2003).

47. P. J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).

48. S. Vieira-Silva, *et al.*, Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**, 310–315 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.15.515575; this version posted May 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

49. J. Friedman, E. J. Alm, Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687 (2012).

50. S. C. Watts, S. C. Ritchie, M. Inouye, K. E. Holt, FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2019).

51. D. Finn, *MicroNiche: Microbial niche measurements* (2020).

52. K. Blin, *et al.*, antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**, W29–W35 (2021).

53. M. Feldgarden, *et al.*, AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports* **11**, 12728 (2021).