Abundant fungi dominate the coexistence of microbiota in soil of contaminated site: High-precision community analysis by full-length sequencing

Kang Yan, Jiahang Zhou, Cong Feng, Suyuan Wang, Bart Haegeman, Weirong Zhang, Jian Chen, Shouqing Zhao, Jiangmin Zhou, Jianming Xu, et al.

Abundant fungi dominate the coexistence of microbiota in soil of contaminated site: High-precision community analysis by full-length sequencing

Kang Yan[1], Jiahang Zhou[1], Cong Feng[2], Suyuan Wang[1], Bart Haegeman[3], Weirong Zhang[1], Jian Chen[4], Shouqing Zhao[4], Jiangmin Zhou[5], Jianming Xu[1], Haizhen Wang[1, *]

[1] Institute of Soil and Water Resources and Environmental Science, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

[2] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China

[3] Centre for Biodiversity Theory and Modelling, Experimental Ecology Station, Centre National de Recherche Scientifique, Moulis, France

[4] Plant Protection, Fertilizer and Rural Energy Agency of Wenling, Zhejiang Province, Wenling 317500, China

[5]College of Life and Environmental Sciences, Wenzhou University, Wenzhou 325035, Zhejiang, China

Corresponding author's e-mail address: wanghz@zju.edu.cn

Abstract

In the past decade, the characterization of microbial community in soil of contaminated sites was primarily done by high-throughput short-read amplicon sequencing. However, the short-read approach often limits the microbial composition analysis at the species level due to the high similarity of 16S rRNA and ITS genes amplicon sequences. Here, we simultaneously performed full-length (PacBio platform) and short-read (Illumina platform) amplicon sequencing to clarify the adaptation mechanisms of different microbial taxa to soil pollution from a high-resolution perspective. We found that (1) full-length 16S rRNA gene sequencing from PacBio platform gave better resolution for bacterial identification at all levels (especially at the level of genus and species), while there was no significant difference between the two platforms for fungal identification in some samples. (2) abundant taxa dominated the microbial communities, and abundant fungal species such as *Mortierella alpine*, *Fusarium solani*, *Mrakia frigida*, and *Chaetomium homopilatum* served as the keystone species. (3) heavy metal and soil texture affected microbial community structure significantly, and abundant taxa preferred deterministic processes, whereas rare taxa randomly formed due to weak selection. Importantly, our study for the first time characterized soil microbiota in contaminated sites with a superior resolution at the species level, emphasizing that abundant taxa, especially abundant fungi, played the keystone role in co-occurrence networks. Overall, these findings expand current understanding of the ecological mechanisms and microbial interactions in contaminated site ecosystems and demonstrate that full-length sequencing has the potential to provide more details of microbial community.

Keywords: Contaminated site; PacBio Sequel; Full-length sequencing; Species-level analysis; Abundant taxa

1. Introduction

Contaminated sites, defined as the land that are actually or potentially hazardous to the environment or human health, have become an extensive and serious environmental problem in China [1-6]. Heavy metals, polycyclic aromatic hydrocarbons, and polychlorinated biphenyls are common pollutants at contaminated sites [3-5, 7, 8], and soil is the primary environmental recipient of these pollutants [2]. Long-term exposure of soil to these pollutants may damage soil health and ecosystem function by altering the biodiversity [7]. For example, compared with clean plots, old creosote-contaminated sites had lower bacterial diversity, and the relative abundance of *Actinobacteria* and *Planctomycetes* was also reduced [7]. Restoration of contaminated sites ecosystem needs a comprehensive understanding of microbial responses to contaminants, both in terms of community composition and function [7-9]. In addition, clarifying native indigenous and their activities in the soil of contaminated sites will help implement remediation measures [10]. There have been many studies using high-throughput sequencing methods to characterize the microbial communities in contaminated soil [7-9, 11, 12], and results showed that some microorganisms may be used for bioremediation [11].

Unfortunately, the use of the 16S rRNA or internal transcribed spacer (ITS) gene as a taxonomic marker has been constrained by the short-read sequencing from Illumina platform, which is the most commonly used second-generation sequencing for microbial community profiling [13-15]. The majority of contemporary 16S rRNA or ITS gene sequence information from short-read sequencing platforms limited the classification of microbiota below the genus level [15]. Species-level identification and functional assignment of microbiota are core objectives of microbial and soil ecology [16-17]. Third-generation sequencing platforms, such as Pacific Biosciences (PacBio)

and Oxford Nanopore Technologies, have developed new technologies with full-length sequencing [14,18]. Full-length sequencing can dramatically increase the accuracy of taxonomic assignments on previously unobtainable scales, and has the potential to facilitate microbial community studies [13, 15]. However, a major shortcoming of full-length sequencing is its higher error rate (~ 10%) compared to short-read sequencing (~ 0.5%) [14, 19]. For PacBio platform, this issue can be solved through construction of a circular consensus sequence (CCS), in which individual amplicon molecules are sequenced multiple times using circularized library templates that provide consensus-sequence error correction [19-20]. In this way, the error rate of a CCS-generated amplicon can be reduced to a level comparable to that of short-read platform [19-20].

In the natural ecosystem, soil microbial community typically shows a skewed taxa abundance distribution, with relatively few abundant taxa and a high number of rare taxa [21-23]. Clarifying the factors that influence the assembly of different subcommunities (abundant taxa and rare taxa) in the soil of the contaminated site is of great significance for ecosystem restoration [23]. In many studies, second-generation sequencing has been used for direct identification of rare and abundant microbes with the relative abundance and the microbial diversity in these subcommunities [23-28]. For example, the relative abundance of < 0.01% has been often used to define rare taxa [22-23]. However, the heterogeneity of soil might result in large differences of microbial biomass among the various samples [7-9]. Namely, the absolute abundances of taxa related to the relative abundance of 0.01% were different among the various samples. Moreover, the studies reviewed that the trends of the relative and absolute abundances of some taxa in community were inconsistent by using the integrated high-throughput absolute abundance quantification (iHAAQ) method [29]. A potential bias existed in the use of relative abundances to facilitate the comparison across

communities [29-33], which largely influenced the identification of rare taxa [21]. Furthermore, as mentioned earlier, short-read sequencing failed to describe the microbial composition of different sub-communities at species level [15-16]. Therefore, it is necessary to define properly rare taxa with full-length sequencing by both the absolute and relative abundances than by only the relative abundance.

The aim of this study was to characterize the dynamics of microbial community in soil at the contaminated site from full-length sequencing perspective. Specifically, for clarifying the high resolution of full-length sequencing in the classification of microbiota, we simultaneously performed full-length 16S rRNA/ITS gene amplicon sequencing using PacBio Sequel and short-read amplicon sequencing (targeting the V3-V4 region of the 16S rRNA gene and internal transcribed spacer ITS1) using the Illumina platform to investigate the structure of the microbiota in contaminated soil. And then, we explore the community composition, functional differences, ecological status, and community assembly mechanisms of the abundant/rare taxa in the soil of the contaminated site using full-length sequencing combined with iHAAQ method.

2. Material and Methods

2.1 Soil sampling and physicochemical analysis

Soil samples were collected from two typical contaminated sites, Anshan, Liaoning Province (AS, 41° 9′ N, 122°0′ E) with a long history of heavy industry activates, and Taizhou, Zhejiang Province (TZ, 28° 28′ N, 121°20′ E) with intensive electronic dismantling activities in China (Fig. S1). The surface soil samples (0-20cm) were collected around the contaminated sites, and each soil sample was composited by five soil cores. All samples were stored at low temperature in ice bags and transported to the laboratory. After homogenization, passed through a 2-mm sieve, and a portion of

each soil sample was stored at - 80 °C until DNA extraction, while the remainder of the sample was immediately processed for physicochemical analysis.

Soil physicochemical properties were determined according to previous methods [34]. Specifically, soil pH was measured by pH meter (pHSJ-3F, Leici, China) using a 1:2.5 soil/water mixture. The soil total carbon (TC) and total nitrogen (TN) were determined using an elemental analyser (Elementar Analysensysteme Gmb H, Germany). The concentrations of dissolved organic carbon (DOC) and dissolved organic nitrogen (DON) were extracted by 0.5 mol/L K_2SO_4 for 30 min and filtered, and the extracts were analyzed using high-temperature combustion (Multi N/C 3100, Analytik Jena AG, Jena, Germany) [35]. Available phosphorus (AP) was extracted by NH_4F-HCl or $NaHCO_3$ (based on soil pH), and was determined colorimetrically at 880 nm. Available potassium (AK) was extracted using 1 mol/L NH_4OAc and analyzed by flame atomic-absorption spectrophotometry. Soil texture composition was measured using the pipette method. For the total heavy metals (Cd, Cu, Ni, Zn, Pb, and As), a 0.2-g soil sample was digested with an acid mixture of HNO_3, HF, and H_2O_2 (volume ratio = 4:2:2) in a microwave digester (MARS6, CEM Microwave Technology Ltd., USA), then determined through ICP-MS (ICP-MS NEXION300XX, PerkinElmer, Inc., USA) [36]. The extraction of polycyclic aromatic hydrocarbons (PAHs) in soil samples using the ultrasonic agitation for 30 min with 10 mL of 1:1 acetone/n-hexane (v/v) [37], and were analyzed by GC-MS (N6890/5975B, Agilent, USA). The results of physicochemical properties and each pollutant were shown in Table S1-S2

2.2 Soil microbial DNA extraction and sequencing

Soil DNA was extracted from soil samples using the TGuide S96 Magnetic Soil/Stool DNA Kit (DP812, TIANGEN BIOTECH, China), following the manufacturer's instructions. The DNA was quantified using a Synergy HTX multi-

144 mode reader (Gene Company Limited) for nucleic acid quantification. Then, full-
145 length 16S rRNA/ITS gene amplicon sequencing, short-read amplicon sequencing
146 (targeting the V3-V4 region of the 16S rRNA gene and ITS1) and quantitative PCR
147 (qPCR) determination were both performed using the same soil DNA sample.
148 For full-length 16S rRNA gene amplicon sequencing, the reaction volume for the
149 PCR was 20 μL, using primers 27F (5'-AGRGTTTGATYNTGGCTCAG- 3') and
150 1492R (5'-TASGGHTACCTTGTTASGACTT -3'). And the full-length ITS region was
151 amplified using primers ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3') and ITS4R
152 (5'-TCCTCCGCTTATTGATATGC-3') with 30μL reaction volume. The PCR
153 amplification conditions for full length of bacterial 16S rRNA were as follows: initial
154 denaturation at 95 °C for 2 min, 25 cycles at 98 °C for 10 s, 55 °C for 30 s, 72 °C for
155 90 s, and finally at 72 °C for 2 min. And for the full length of fungal ITS region, the
156 PCR amplification conditions were: initial denaturation at 95 °C for 5 min, 8 cycles at
157 95 °C for 30 s, 55 °C for 30 s, 72 °C for 45 s, 24 cycles at 95 °C for 30 s, 60 °C for 30
158 s, 72 °C for 45 s, and finally at 72 °C for 5 min. The SMRTbell libraries were
159 constructed using a PacBio SMRTbell template prep kit (PacBio, USA) following the
160 manufacturer's manual. The PCR products were purified (AMPure PB), and library
161 size distributions were analyzed using an Agilent 2100 bioanalyzer. Finally, qualified
162 and adequate library ligated with primer and polymerase using PacBio Binding kit
163 (PacBio, USA). The final library was sequenced on PacBio Sequel Ⅱ (PacBio, USA).
164 The hypervariable V3-V4 regions of bacterial 16S rRNA gene were amplified using
165 universal primer pairs 338F (5′- ACTCCTACGGGAGGCAGCA -3′) and 806R (5′-
166 GGACTACHVGGGTWTCTAAT -3′), and the fungal ITS1 region was amplified
167 using a common primer pairs ITS1-F (5'- CTTGGTCATTTAGAGGAAGTAA -3') and
168 ITS2 (5'- GCTGCGTTCTTCATCGATGC -3'). The Solexa PCR reaction mixture (20

μL) contained 5 μL of targeted PCR product, 2.5 μL of MPPI-a (2 μM), 2.5 μL of MPPI-b (2 μM), and 10 μL of 2 × Q5 HF MM. The PCR protocol was as follows: an initial denaturation at 98 °C for 30 s; 10 cycles of 10 s at 98 °C, 30 s at 65 °C, and 30 s at 72 °C; and a final extension at 72 °C for 5 min. The amplified DNA for each samples were sequenced by Illumina NovaSeq6000 platform (Illumina, USA).

2.3 qPCR determination

Each DNA sample with three analytical replicates (Roche LightCycler 480 Real-Time PCR Machine) was performed to determine the copy numbers of 16S rDNA and ITS (internal transcribed spacer) region of rDNA [38]. Primers and the qPCR conditions were listed in Table S3. Amplifications were performed in a final volume of 20 μL containing 10 μL of SYBR Green qPCR SuperMix UDG (Invitrogen, USA), 1 μL DNA, 0.4 μL of each primer, and 8.2 μL nucleic acid-free water. Plasmids carrying the corresponding gene were used to establish qPCR standard curves (additional details in Table S1). All runs had standard efficiency curves of $R^2 > 0.999$ and the efficiencies reached 90-110% (result were in Table S4).

2.4 Analysis of sequencing outputs

For the full-length sequencing data, circular consensus sequence was clustered into amplicon sequence variants (ASVs) with the software packages DADA2 (version 1.18.0) in R v4.0.3 [39-41]. Briefly, primers were detected and removed from reads using the removePrimers function. The remaining sequences were filtered using the filterAndTrim function [39]. Then, the ASVs were inferred after learning error rates. The learnErrors method learns this error model from the data, by alternating estimation of the error rates and inference of sample composition until they converge on a jointly consistent solution [40]. The learnErrors function with parameter: errorEstimationFunction=PacBioErrfun, BAND_SIZE=32 [39-40]. And the dada

function was set with parameters of OMEGA_A=1e-10 and DETECT_SINGLETONS=TRUE [41]. Chimera sequences were removed using the consensus method in the removeBimeraDenovo command. Taxonomy up to the species-level was assigned using the assignTaxonomy function based on the silva_nr99_v138_wSpecies_train_set.fa.gz (recommend for full-length 16S data from PacBio HiFi sequencing) for full-length16S data and UNITE general FASTA release (Version 01.12.2017) for full-length ITS data [39, 42].

Short-read sequences were also processed using the DADA2 pipeline (https://benjjneb.github.io/dada2/tutorial.html), which allowed inference of exact ASVs. After assessing forward and reverse read quality (plotQualityProfile command), we trimmed the sequences and removed primers with the command 'filterAndTrim' using the following parameters: maxN = 0, truncQ = 2, rm.phix=TRUE, maxEE = c(1,1), minQ=3. The filtered reads were then fed to the error rate learning, dereplication, and ASV inference steps using the functions learnErrors, derepFastq, and dada [40]. Afterward, the forward and reverse reads were merged using the mergePairs function. Chimeric sequences were removed using the removeBimeraDenovo function with the "consensus" method parameter [43]. Finally, the taxonomic assignment of Chimera-free sequences for short-read 16S data was completed using assignTaxonomy and addSpecies function with silva_nr99_v138_train_set.fa.gz and silva_species_assignment_v138.fa.gz and for short-read ITS data using assignTaxonomy function with UNITE general FASTA release (Version 01.12.2017) [40, 44].

2.5 Data analysis and statistics

All statistical analyses were performed using R software (version 4.0.3) [45]. Alpha diversity indices, including Shannon diversity index and Pielou evenness index,

were calculated for each sample in vegan (version 2.5-7) [46]. The principal coordinates analysis (PCoA) with Bray-Curtis distance and nonmetric multidimensional scaling with Bray-Curtis dissimilarities were analyzed by vegan (version 2.5-7) and ggplot2 (version 3.3.2) packages in R to illustrate the differences between the two sequencing platforms in revealing community structure [47]. After comparing the two sequencing platforms, the full-length sequencing was used to analyze taxonomy and functional composition, co-occurrence network, and community assembly, as it could provide high-resolution phylogenetic microbial community profiling. Only ASVs that occurred in a minimum of 2 samples were kept for these analysis [48].

Moreover, the absolute abundance of taxon i in sample S ($AA_{i,s}$) was calculated by the following equation (1) according to the previous studies [29,31-32].

$$AA_{i,s} = RA_{i,s} \times Q_s \quad (1)$$

where $RA_{i,s}$represents the relative abundance of taxon i in sample S obtained by full-length sequencing, and Qs stands for the copy numbers of bacterial/fungi of sample S detected by qPCR.

To compensate for potential bias lies in the use of relative abundance alone to define rarity, the absolute and relative abundances were considered simultaneously to distinguish rare/abundant taxon in this study. The total absolute abundance of taxon i ($\overline{AA_i}$), total copy number detected by qPCR ($\overline{Q}$) across all 12 soil samples were denoted by the equations 2 and 3, respectively.

$$\overline{AA_i} = \sum_{s=1}^{12} AA_{i,s} \quad (2)$$

$$\overline{Q} = \sum_{s=1}^{12} Q_s \quad (3)$$

And if $\overline{AA_i} > 0.1\% \overline{Q}$ or $\overline{AA_i} < 0.01\% \overline{Q}$, then the taxon i defined as abundant or rare taxon. The taxon i was subjected to the intermediate taxa when its copy number between the thresholds of rare and abundant taxa.

Each ASV across the dataset was analyzed by using the PICRUSt2 pipeline to get a deeper understanding of the metabolic profiles of abundant and rare bacterial subcommunities [49], and the program FUNGuild was used to predict fungal functional guilds [50]. A network analysis was performed to explore the co-occurrence patterns between different subcommunities. The ASVs with low frequency were removed from network analysis to avoid possible biases, as a large number of zeros values may introduce spurious correlations [27,51]. The correlation matrix between two ASVs was calculated in "psych (version 2.0.12)" [52]. Spearman's correlation coefficient was >0.6 and the *p* -value was <0.01 were integrated into network analysis, and the several network properties (degree, betweenness centrality, closeness centrality and eigencentrality) were calculated with the interactive platform Gephi [53]. The ASVs with the highest betweenness centrality scores were considered keystone [54-55]. The normalized stochasticity ratio (NST) was utilized to estimate the determinacy and stochasticity of the different subcommunities assembly processes with high accuracy and precision [56]. The NST values were used to evaluate the deterministic (<50%) or more stochastic (>50%) assembly processes of soil microbiota. We calculated NST variations based on Ruzicka (abundance-based) dissimilarity metrics in the "NST" package (version 3.0.6) [56].

3. Results

3.1 Taxonomic analysis of different platforms

After quality filtering, a total of 155,022 full-length (average 12,919 sequence reads per soil sample) and 474,506 short-read (average 39,542 sequence reads per soil sample) 16S rRNA gene sequence reads were produced. In total, 103,963 and 7453 ASVs were identified from PacBio and Illumina sequencing, respectively. For the ITS gene sequencing, a total of 116,434 full-length (average 9703 sequence reads per soil sample) and 426,931 short-length (average 35,578 sequence reads per soil sample) sequence reads were produced after quality filtering. The total of 4831 and 3367 ASVs were identified from PacBio and Illumina sequencing, respectively.

Full-length 16S rRNA gene sequencing from PacBio platform gave better resolution than the short-read from Illumina platform for bacterial identification at all levels, especially at the level of genus and species ($p < 0.01$) (Fig. 1, Fig. S2-S3). In the 12 test soil samples, the PacBio platform detected 59 phyla, 147 classes, 337 orders, 425 families, 889 genera and 788 species of bacteria, while 49 phyla, 126 classes, 258 orders, 314 families, 523genera, and 86 species were detected by the Illumina platform (Fig. 1). For fungi, the PacBio platform detected 14 phyla, 42 classes, 110 orders, 228 families, 438 genera and 584 species, while 14 phyla, 36 classes, 95 orders, 208 families, 425 genera, and 563 species were detected by the Illumina platform (Fig. 1). Full-length ITS sequencing had identified more taxa than Illumina ITS amplicon sequencing at all taxonomic levels, but there had no significant difference ($p > 0.05$).

3.2 Microbial community composition revealed by two platform

Although a large number of taxa at the genus and species level of the bacteria were only identified by full-length sequencing (Fig. 1), the relative abundance of the genus and species level still has a very significant correlation between the results of the two sequencing platforms ($p < 0.001$) (Fig. 2). Spearman's rank correlation illustrated that the two sequencing platforms revealed similar bacterial compositions at the level

of phylum (rho-value 0.665-0.909, $p < 0.001$), class (rho-value 0.511-0.755, $p < 0.001$), order (rho-value 0.365-0.725, $p < 0.001$) and family (rho-value 0.345-0.744, $p < 0.001$) (Fig. S4-S7). However, the fungal compositions of some soil samples at the levels of phylum to family obtained from the two sequencing platforms were different ($p > 0.05$) (Fig. S8-S11).

The relative abundance of the 15 most abundant bacteria determined at genus level with each platform are shown using heatmaps in Fig. S12 and Fig. S13. For bacteria, in some soil samples, genera such as *Nitrospira*, *Terrimonas*, *Ferruginibacter*, and *Gemmatimonas* were only detected in full-length sequencing, while *Acinetobacter* was only identified by short-read sequencing. For fungi, genera such as *Podospora*, *Kamienskia*, and *Rhizophagus* were only detected in full-length sequencing for some soil samples, while *Saitozyma*, *Chaetomium*, and *Inocybe* were only identified in short-read sequencing for some soils. Compared to short-read sequences, a larger proportion of the full-length sequences with the relative abundance below 0.01% was identified. It further illustrated that the full-length sequencing provided a higher resolution to analysis of microbial community (Fig. 2).

Moreover, the comparisons of alpha-diversity indices (Shannon and Pielou) of the soil bacteria were significantly different between the two platforms (Fig. 3). It was demonstrated that the PacBio platform exhibited a significantly higher level of bacterial α-diversity than the Illumina platform ($p < 0.01$). Both the results of full-length sequencing and short-read sequencing can reveal the beta-diversity differences between the bacterial communities of the two site samples, with the higher dissimilarity values of full-length sequencing (ANOSIM test, R = 0.9352 vs R = 0.8056). And for fungi, there is no significant difference in alpha-diversity indices (Shannon and Pielou index) of the soil obtained by the two sequencing platforms ($p >$

316 0.05). However, in the analysis of beta- diversity, full-length sequencing can better
317 distinguish the differences between the soil of different sites than short-read sequencing
318 (ANOSIM test, R = 0.8537 vs R = 0.3148).

319 **3.3 Taxonomic and functional composition of different subcommunities**

320 Microbial community compositions differed between abundant and rare taxa,
321 whiles the phylum *Proteobacteria* dominated both rich and rare bacteria
322 subcommunities (Fig. 4a), and *Ascomycota* were the key phylum in both the abundant
323 and rare fungi subcommunities (Fig. 4c). For bacteria, compared with the abundant
324 taxa (> 2.07×10^8 copies g^{-1}), the rare taxa (< 2.07×10^7 copies g^{-1}) have a higher
325 proportion of *Acidobacteriota* (19.5%) and *Actinobacteriota* (9.0%), and a lower
326 proportion of *Gemmatimonadota* (7.6%) and *Bacteroidota* (7.4%). In addition, at the
327 species level, a large number of rare bacteria such as *Parasegetibacter luojiensis*,
328 *Massilia violaceinigra*, *Povalibacter uvarum*, *Flavisolibacter ginsengiterrae* and
329 *Chryseobacterium indologenes* were identified in full-length sequencing (Table S5).
330 Moreover, we found that defining different groups of species based on absolute
331 thresholds reduced the bias caused by different microbial biomass of samples (Fig.
332 S14). For example, *Nocardioides iriomotensis* had an average relative abundance of
333 0.027% but was defined as rare taxa under absolute thresholding results (Table. S5).
334 This result was more convincing because *Nocardioides iriomotensis* only richened in
335 some samples with low microbial biomass, which means, its actual absolute abundance
336 was still rare under site habitat. The function ratio of abundant bacteria in the TCA
337 Cycle, Embden Meyerhof-Parnos, and Superpathway of thiosulfate metabolism were
338 higher than that of the rare taxa (Wilcoxon rank sum tests, p <0.01), and these
339 functions are often necessary to maintain basic life activities (Fig. 4b). While the rare
340 bacteria contained more metabolic functional diversity, and the ratio of functions such

as Hydrocarbon degradation and Transporters were higher than that of the abundant taxa (Fig. 4b).

As for fungi, the rare fungi (< 5.83×10^5 copies g^{-1}) have a higher proportion of *Basidiomycota* (11.9%), *Rozellomycota* (7.1%) and *Glomeromycota* (4.76%) than abundant fungi (> 5.83×10^6 copies g^{-1}), while the abundant taxa have a higher proportion of *Mortierellomycota* (12.28%) (Fig. 4c). Moreover, the abundant fungi such as *Mortierella alpine*, *Mrakia frigida*, *Glaciozyma martini* and *Guehomyces pullulans* have negatively correlated with the heavy metals and positively correlated with organic pollution in soil (Fig. 5). Fungal functional guilds prediction showed that the abundant fungi had a higher proportion of "Pathotroph (12.1%)" and "Pathotroph-Saprotroph (6.0%)" and a lower proportion of "Pathotroph-Symbiotroph" and "Pathotroph-Saprotroph-Symbiotroph" than the rare fungi. While, there is a large proportion of rare fungi (43.1%) whose function is unknown (Fig. 4d).

3.4. Co-occurrence patterns and assembly mechanism of different subcommunities

Based on the correlation results, we further performed network analysis to disentangle the ecological role and co-occurrence patterns of abundant and rare subcommunities in the soil of contaminated site. Results showed that the average value of topological properties of abundant taxa was higher than that of rare taxa, suggesting that the abundant taxa were located in central positions within the network more often than the rare taxa (Fig. 6a, d-g). Moreover, the bacterial-fungal co-occurrence networks were composed of 9786 edges and 371 nodes, which included 77.63% bacteria and 22.37% fungi, and their average absolute abundance were 3.56×10^7 copies/g and 3.08×10^7 copies/g, respectively (Fig. 6b). Species with the highest betweenness centrality scores were considered as the keystone species. The abundant fungi

366 (*Mortierella alpina* and *Fusarium solani*, etc.) served as the keystone of the network
367 (Fig. 6a-b, Table S6), and were more involved in maintaining the microbiomes (Fig.
368 6c). *Mortierell alpina* and *Fusarium solani* also have highest degree (> 100) and
369 closeness centrality (> 0.50) (Fig. 6c, Table S6). Moreover, the specific networks
370 without top 20 keystone ASVs (abundant fungi) showed that microbial networks were
371 less complex, and the important parameters such as edges, degree, and average
372 weighted degree decreased (Fig. 6c). Specifically, in the co-occurrence network of
373 Anshan site (Fig. S15a), abundant fungi make up half of the top 10 ASVs for the
374 betweenness centrality. In addition, the abundant fungi *Fusarium solani* (ASV1146)
375 and *Mortierella elongate* (ASV4424) also have high value of betweenness centrality
376 (Fig. S15b) in the co-occurrence network of Taizhou site.

377 Also, with the help of full-length sequencing, we can identify keystone species at
378 the species level, and clarify which the abundant fungi maintain the coexistence of soil
379 microbial community in the contaminated site. For examples, the co-occurrence
380 network of fungi in Anshan site contains 2656 edges and 307 nodes (average network
381 path was 3.562, modularity coefficient was 0.815, and average clustering coefficient
382 was 0.868), and the *Chaetomium homopilatum* (ASV926), *Gibellulopsis nigrescens*
383 (ASV1077), *Fusarium oxysporum* (ASV1440), *Mrakia frigida* (ASV1973), and
384 *Nigrospora oryzae* (ASV733) paly as the keystone species. Moreover, the keystone
385 species in fungal co-occurrence network in Taizhou site were also mainly played by
386 abundant fungi (Fig. 6g), such as *Trichoderma virens*, *Cladosporium delicatulum*,
387 *Curvularia lunata*, etc.

388 Mantel test suggested that heavy metal and soil texture exhibited the strongest
389 correlations with microbial community, and other significant environmental variables
390 were PAHs and pH (Fig. 7a, Table S7). The abundant bacteria had stronger

391 correlations with soil texture, As and Low-ring PAHs, while the rare bacteria had
392 stronger correlations with soil texture, As, and Cu (Table S7). In addition, the abundant
393 fungi also had correlations with soil texture, As and Cd, while the rare fungi had
394 stronger correlations with the content of sand and clay in the soil (Fig. 7a). The NST
395 results showed that the deterministic and stochastic processes were more important to
396 the abundant and rare taxa, respectively (Fig.7 b). Additionally, a significantly lower
397 NST value was observed in the bacterial communities with an average of 49.96% than
398 that in the fungal communities with an average of 57.97%.
399
400 **4. Discussion**
401 As shown in Fig. 1, Figs. S2 and S3, a number of species were identified only
402 through full-length sequencing, suggesting that short-read sequences of the 16S
403 rRNA/ITS gene couldn't provide the same taxonomic resolution as full-length
404 sequences. This was in line with Lam et al. who compared the anaerobic digesters
405 microbiome as revealed by the two sequencing platforms [13]. In addition to the
406 PacBio platform, Nanopore MinION Technologies (ONT) also can achieve full-length
407 sequencing [14, 17, 57], and previous studies using ONT platform have also shown
408 that full-length sequencing reported greater taxonomic resolution than Illumina MiSeq
409 for the dust samples [57]. In addition, the higher taxonomic resolution not only
410 required longer read lengths, but also needed the suitable reference databases [39, 41].
411 The "silva_nr99_v138_wSpecies_train_set.fa.gz" database is suitable for assigning
412 bacterial taxonomy down to species level. However, due to a lack of best hits in the
413 reference database, the taxonomic resolution of fungi was not as high as expected (Fig.
414 1).

Furthermore, the different sub-regions of 16S rRNA/ITS showed bias led to the inconsistent identified taxonomic outcomes between the full-length and short-read sequencings [16, 58]. For examples, compared with full-length sequencing, short-read sequencing underestimated the relative abundance of *Bacteroidota*, *Gemmatimonadota*, *Myxococcota*, *Rozellomycota* and *Blastocladiomycota* while overestimated the relative abundance of *Acidobacteriota* and *Olpidiomycota* (Fig. S18). Likewise, the α-diversity of microbial diversity, the richness components (Shannon's index) and the evenness (Pielou's index), could be influenced by the type of platform. Similar results have been observed in the previous studies [13], and only sequencing limited variable region of 16S rRNA gene caused the underestimation of community diversity [16]. The clustering of samples was performed using PCoA and NMDS in this study, and both methods suggested differentiation between both platforms and the test samples (Fig. 3), suggesting that some differentiations were only be observed in full-length sequencing platform [13, 15].

Since few microbial community studies performed with full-length 16S rRNA/ITS sequences, the genus level was commonly used for comparison of environments samples [7-8, 10-12]. However, identification to the genus level is not enough for a comprehensive understanding of microbial community [13, 15-17]. In addition, analysis at species level could provide a more detailed description of the microbial communities and a better integrated understanding of functional characterization [16-17, 39]. For example, Earl et al. found that full-length sequencing enabled the species-level analysis of the sinonasal microbiome, and *Cutibacterium acnes* was the predominant species in patients' sinonasal samples [15]. Similarly, Lam et al. reported that methanogenic species, i.e., *Methanosarcina horonobensis* and *Methanosarcina flavescens*, were dominant in different digesters, and these species can

serve as guide for inoculum selection [13]. In our research, according to the species-level analysis (Fig. S16-S17), we also found that a number of species such as *Lysobacter dokdonensis*, *Nitrospira defluvii* and *Vicinamibacter silvestris* were negatively correlated with the As contents of soil (Fig. 5). Moreover, according to the principles of transitive prediction schema and collaborative filtering predictor combined with full-length sequencing base information [59], the cultivation conditions of keystone species were predicted (Table S8). If the keystone species can be obtained by the predicted cultivation conditions, their functions will be further explored and verified. In total, compared with traditional 16S rDNA sequencing of the MiSeq platform, the PacBio platform improved its read length and annotated the nucleotide sequence of soil microbiota to the species level. Full-length sequencing may be optimal for soil microbiota sequencing due to its long reads and high performance, while platforms such as Illumina MiSeq will have the advantage of cost-efficiency [13-15].

Further analysis of the microbial co-occurrence patterns indicated that abundant fungi were the critical network nodes in soil microbial associations. And the rare taxa were not evenly distributed throughout the soil of contaminated site, most of them were only present in a few samples (Fig. 5). This could be explained by the fact that abundant fungi can effectively utilize a broader range of resources than others [23, 28]. Moreover, compared to the bacteria, fungi have a high tolerance of extreme conditions such as acidic or alkaline pH, low moisture, and higher concentration of metals, etc [60]. In this study, the abundant fungi *Mortierella alpine*, *Fusarium solani*, *Mrakia frigida*, *Nigrospora oryzae*, *Trichoderma virens*, and *Curvularia lunata* had the highest betweenness centrality scores and were regarded as keystone species in the co-occurrence network (Table S6). The previous studies have pointed out that these microorganisms have the ability to degrade PAHs, resist heavy metal, and even can act

as sensitive indicators for contaminated sites [61-64]. In the present study, we found that abundant taxa were more ubiquitous than the rare taxa, and it could be explained by the possibility that the abundant taxa occupied a diverse niche, competitively utilized a broader array of resources than the rare taxa, and could effectively adapt to the environment [21]. These conclusions were supported by the NST value of the abundant taxa (Fig. 6), which reflected that the abundant taxa were less limited and impacted by stochastic processes (i.e., dispersal limitation) compared with the rare taxa. This results also in line with Jiao et al. who found that the abundant taxa were ubiquitous in contaminated soils from different regions and strongly impacted by deterministic filtering [23]. In addition, we also found that the rare taxa contained more functional diversity (Fig. 4). It suggested that the rare species were indicated to act as a "seed bank" that could become dominant under the proper conditions [21-22].

5. Conclusion

In summary, we analyzed the community composition, functional differences, ecological status, and community assembly mechanisms of the abundant/rare taxa in the soil of the contaminated site for the first time using full-length and short-read sequencings from PacBio and Illumina platforms, respectively. Full-length 16S rRNA gene sequencing gave better resolution for bacterial identification at all levels (especially at the level of genus and species), while there was no significant difference between the two platforms for fungal identification in some samples. In the view of high-resolution, abundant fungi (*Mortierella alpine*, *Fusarium solani*, *Mrakia frigida*, *Nigrospora oryzae*, *Trichoderma virens*, and *Curvularia lunata*) were located in central positions within the network more often than the others, emphasizing the dominant role of abundant fungi in the microbial community. Our results also indicated

that heavy metal and soil texture affect microbial community structure significantly, and the abundant taxa preferred deterministic filtering, whereas rare taxa randomly formed due to weak selection (stochastic processes). Moreover, we believe that through the expansion of the database and improvement of the sequencing protocol, the application of the full-length sequencing on environmental samples can be more promising.

Declaration of Competing Interest

None of the authors has any competing interest with this work.

Acknowledgements

We gratefully acknowledge funding from the National Key Research and Development Program of China (No. 2019YFC1803704), the National Natural Science Foundation of China (41771344), and the National College Students′ Innovation and Entrepreneurship Training Program (202210335039).

References

1. Chen R, Ye C. Resolving soil pollution in China. Nature. 2014;505(7484):483. https://doi.org/10.1038/505483c.
2. Yao Y. Spend more on soil clean-up in China. Nature. 2016;533(7604):469. https://doi.org/10.1038/533469a.
3. Niu S, Tao W, Chen R, Hageman KJ, Zhu C, Zheng R, et al. Using polychlorinated naphthalene concentrations in the soil from a southeast China e-waste recycling area in a novel screening-level multipathway human cancer risk

514 assessment. Environ Sci Technol. 2021;55(10):6773-82.
515 https://doi.org/10.1021/acs.est.1c00128.
516 4. Li P, Du B, Maurice L, Laffont L, Lagane C, Point D, et al. Mercury isotope
517 signatures of methylmercury in rice samples from the wanshan mercury mining
518 area, china: environmental implications. Environ Sci Technol. 2017;51(21):12321-
519 8. https://doi.org/10.1021/acs.est.7b03510.
520 5. Yanqun Z, Yuan L, Schvartz C, Langlade L, Fan L. Accumulation of Pb, Cd, Cu
521 and Zn in plants and hyperaccumulator choice in Lanping lead–zinc mine area,
522 China. Environ Int. 2004;30(4):567-76. https://doi.org/10.1016/j.envint.2003.10.012.
523 6. Li X, Jiao W, Xiao R, Chen W, Liu W. Contaminated sites in China:
524 countermeasures of provincial governments. J Clean Prod. 2017; 147:485-96.
525 https://doi.org/10.1016/j.jclepro.2017.01.107.
526 7. Mukherjee S, Juottonen H, Siivonen P, Lloret Quesada C, Tuomi P, Pulkkinen P,
527 et al. Spatial patterns of microbial diversity and activity in an aged creosote-
528 contaminated site. ISME J. 2014;8(10):2131-42. https://doi.org/10.1038/ismej.2014.151.
529 8. Hou D, Zhang P, Zhang J, Zhou Y, Yang Y, Mao Q, et al. Spatial variation of
530 sediment bacterial community in an acid mine drainage contaminated area and
531 surrounding river basin. J Environ Manage. 2019; 251:109542.
532 https://doi.org/10.1016/j.jenvman.2019.109542.
533 9. van Dillewijn P, Caballero A, Paz JA, González-Pérez MM, Oliva JM, Ramos JL.
534 Bioremediation of 2,4,6-trinitrotoluene under field conditions. Environ Sci
535 Technol. 2007;41(4):1378-83. https://doi.org/10.1021/es062165z.
536 10. Sun W, Xiao E, Xiao T, Krumins V, Wang Q, Häggblom M, et al. Response of
537 soil microbial communities to elevated antimony and arsenic contamination

indicates the relationship between the innate microbiota and contaminant fractions. Environ Sci Technol. 2017;51(16):9165-75. https://doi.org/10.1021/acs.est.7b00294.

11. Jiao S, Liu Z, Lin Y, Yang J, Chen W, Wei G. Bacterial communities in oil contaminated soils: biogeography and co-occurrence patterns. Soil Boil Biochem. 2016; 98:64-73. https://doi.org/10.1016/j.soilbio.2016.04.005.

12. Li D, Li G, Zhang D. Field-scale studies on the change of soil microbial community structure and functions after stabilization at a chromium-contaminated site. J Hazard Mater. 2021; 415:125727. https://doi.org/10.1016/j.jhazmat.2021.125727.

13. Lam TYC, Mei R, Wu Z, Lee PKH, Liu W, Lee P. Superior resolution characterisation of microbial diversity in anaerobic digesters using full-length 16s rRNA gene amplicon sequencing. Water Res. 2020; 178:115815. https://doi.org/10.1016/j.watres.2020.115815.

14. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345-53. https://doi.org/10.1038/nature24286.

15. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, et al. Species-level bacterial community profiling of the healthy sinonasal microbiome using pacific biosciences sequencing of full-length 16S rRNA genes. Microbiome. 2018;6(1). https://doi.org/10.1186/s40168-018-0569-2.

16. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, et al. High-resolution phylogenetic microbial community profiling. ISME J. 2016;10(8):2020-32. https://doi.org/10.1038/ismej.2015.249.

17. Johnson JS, Spakowicz DJ, Hong B, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10(1). https://doi.org/10.1038/s41467-019-13036-1.

18. Eisenstein M. An ace in the hole for DNA sequencing. Nature. 2017;550(7675):285-8. https://doi.org/10.1038/550285a.

19. Fichot EB, Norman RS. Microbial phylogenetic profiling with the pacific biosciences sequencing platform. Microbiome. 2013;1(1):10. https://doi.org/10.1186/2049-2618-1-10.

20. Tedersoo L, Tooming Klunderud A, Anslan S. Pacbio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol. 2018;217(3):1370-85. https://doi.org/10.1111/nph.14776.

21. Jia X, Dini-Andreote F, Falcão Salles J. Community assembly processes of the microbial rare biosphere. Trends Microbiol. 2018;26(9):738-47. https://doi.org/10.1016/j.tim.2018.02.011.

22. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol. 2015;13(4):217-29. https://doi.org/10.1038/nrmicro3400.

23. Jiao S, Chen W, Wei G. Biogeography and ecological diversity patterns of rare and abundant bacteria in oil-contaminated soils. Mol Ecol. 2017;26(19):5305-17. https://doi.org/10.1111/mec.14218.

24. Mo Y, Peng F, Gao X, Xiao P, Logares R, Jeppesen E, et al. Low shifts in salinity determined assembly processes and network stability of microeukaryotic plankton communities in a subtropical urban reservoir. Microbiome. 2021;9(1). https://doi.org/10.1186/s40168-021-01079-w.

585 25. Liu L, Yang J, Yu Z, Wilkinson DM. The biogeography of abundant and rare
586 bacterioplankton in the lakes and reservoirs of china. ISME J. 2015;9(9):2068-77.
587 https://doi.org/10.1038/ismej.2015.29.
588 26. Liang Y, Xiao X, Nuccio EE, Yuan M, Zhang N, Xue K, et al. Differentiation
589 strategies of soil rare and abundant microbial taxa in response to changing climatic
590 regimes. Environ Microbiol. 2020;22(4):1327-40. https://doi.org/10.1111/1462-
591 2920.14945.
592 27. Xiong C, He JZ, Singh BK, Zhu YG, Wang JT, Li PP, et al. Rare taxa maintain the
593 stability of crop mycobiomes and ecosystem functions. Environ Microbiol.
594 2021;23(4):1907-24. https://doi.org/10.1111/1462-2920.15262.
595 28. Jiao S, Wang J, Wei G, Chen W, Lu Y. Dominant role of abundant rather than rare
596 bacterial taxa in maintaining agro-soil microbiomes under environmental
597 disturbances. Chemosphere. 2019; 235:248-59.
598 https://doi.org/10.1016/j.chemosphere.2019.06.174.
599 29. Lou J, Yang L, Wang H, Wu L, Xu J. Assessing soil bacterial community and
600 dynamics by integrated high-throughput absolute abundance quantification. PeerJ.
601 2018;6: e4514. https://doi.org/10.7717/peerj.4514.
602 30. Props R, Kerckhof F, Rubbens P, De Vrieze J, Hernandez Sanabria E, Waegeman
603 W, et al. Absolute quantification of microbial taxon abundances. ISME J.
604 2017;11(2):584-7. https://doi.org/10.1038/ismej.2016.117.
605 31. Shi W, Li M, Wei G, Tian R, Li C, Wang B, et al. The occurrence of potato
606 common scab correlates with the community composition and function of the
607 geocaulosphere soil microbiome. Microbiome. 2019;7(1):1-18.
608 https://doi.org/10.1186/s40168-019-0629-2.

609 32. Zhang Z, Qu Y, Li S, Feng K, Wang S, Cai W, et al. Soil bacterial quantification
610 approaches coupling with relative abundances reflecting the changes of taxa. Sci.
611 Rep. 2017;7(1):4837. https://doi.org/10.1038/s41598-017-05260-w.
612 33. Yang L, Lou J, Wang H, Wu L, Xu J. Use of an improved high-throughput
613 absolute abundance quantification method to characterize soil bacterial community
614 and dynamics. Sci Total Environ. 2018; 633:360-71.
615 https://doi.org/10.1016/j.scitotenv.2018.03.201.
616 34. Ma B, Wang H, Dsouza M, Lou J, He Y, Dai Z, et al. Geographic patterns of co-
617 occurrence network topological features for soil microbiota at continental scale in
618 eastern China. ISME J. 2016;10(8):1891-1901.
619 https://doi.org/10.1038/ismej.2015.261.
620 35. Jones D, Willett V. Experimental evaluation of methods to quantify dissolved
621 organic nitrogen (DON) and dissolved organic carbon (DOC) in soil. Soil Biol.
622 Biochem. 2006;38(5):991-9. https://doi.org/10.1016/j.soilbio.2005.08.012.
623 36. Zhai W, Qin T, Li L, Guo T, Yin X, Khan MI, et al. Abundance and diversity of
624 microbial arsenic biotransformation genes in the sludge of full-scale anaerobic
625 digesters from a municipal wastewater treatment plant. Environ Int. 2020;
626 138:105535. https://doi.org/10.1016/j.envint.2020.105535.
627 37. Gu H, Yan K, You Q, Chen Y, Pan Y, Wang H, et al. Soil indigenous
628 microorganisms weaken the synergy of *Massilia* sp. WF1 and *Phanerochaete*
629 *chrysosporium* in phenanthrene biodegradation. Sci Total Environ. 2021;
630 781:146655. https://doi.org/10.1016/j.scitotenv.2021.146655.
631 38. Fierer N, Jackson JA, Vilgalys R, Jackson RB. Assessment of soil microbial
632 community structure by use of taxon-specific quantitative PCR assays. Appl

633 Environ Microb. 2005;71(7):4117-20. https://doi.org/doi:10.1128/AEM.71.7.4117-
634 4120.2005.
635 39. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, et al. High-
636 throughput amplicon sequencing of the full-length 16S rRNA gene with single-
637 nucleotide resolution. Nucleic Acids Res. 2019;47(18): e103.
638 https://doi.org/10.1093/nar/gkz569.
639 40. Callahan BJ, Mcmurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP.
640 DADA2: high-resolution sample inference from Illumina amplicon data. Nat
641 Methods. 2016;13(7):581-3. https://doi.org/10.1038/nmeth.3869.
642 41. Callahan BJ, Grinevich D, Thakur S, Balamotis MA, Yehezkel TB. Ultra-accurate
643 microbial amplicon sequencing with synthetic long reads. Microbiome.
644 2021;9(1):1-13. https://doi.org/10.1186/s40168-021-01072-3.
645 42. Schiro G, Colangeli P, Müller MEH. A metabarcoding analysis of the mycobiome
646 of wheat ears across a topographically heterogeneous field. Front Microbiol.
647 2019;10. https://doi.org/10.3389/fmicb.2019.02095.
648 43. Brede M, Orton T, Pinior B, Roch F, Dzieciol M, Zwirzitz B, et al. PacBio and
649 Illumina Miseq amplicon sequencing confirm full recovery of the bacterial
650 community after subacute ruminal acidosis challenge in the RUSITEC system.
651 Front Microbiol. 2020;11. https://doi.org/10.3389/fmicb.2020.01813.
652 44. Community U. Unite general fasta release: UNITE Community Shadwell; 2017.
653 45. Team RC. R: a language and environment for statistical computing. 2013.
654 46. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O Hara R, et al.
655 Package ‘vegan’. Community ecology package, version 2. 2013.

656 47. Wickham H. Ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics.
657 2011;3(2):180-5.
658 48. Langer JAF, Sharma R, Schmidt SI, Bahrdt S, Horn HG, Algueró-Muñiz M, et al.
659 Community barcoding reveals little effect of ocean acidification on the
660 composition of coastal plankton communities: Evidence from a long-term
661 mesocosm study in the Gullmar Fjord, Skagerrak. PLoS One. 2017;12(4):
662 e175808. https://doi.org/10.1371/journal.pone.0175808.
663 49. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al.
664 Picrust2 for prediction of metagenome functions. Nat Biotechnol. 2020;38(6):685-
665 8. https://doi.org/10.1038/s41587-020-0548-6.
666 50. Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, et al. Funguild:
667 an open annotation tool for parsing fungal community datasets by ecological guild.
668 Fungal Ecol. 2016; 20:241-8. https://doi.org/10.1016/j.funeco.2015.06.006.
669 51. Xue Y, Chen H, Yang JR, Liu M, Huang B, Yang J. Distinct patterns and
670 processes of abundant and rare eukaryotic plankton communities following a
671 reservoir cyanobacterial bloom. ISME J. 2018;12(9):2263-77.
672 https://doi.org/10.1038/s41396-018-0159-0.
673 52. Revelle WR. Psych: procedures for personality and psychological research. 2017.
674 53. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring
675 and manipulating networks. Poster presented at: Third international AAAI
676 conference on weblogs and social media; 2009.
677 54. Banerjee S, Kirkby CA, Schmutter D, Bissett A, Kirkegaard JA, Richardson AE.
678 Network analysis reveals functional redundancy and keystone taxa amongst
679 bacterial and fungal communities during organic matter decomposition in an

680 arable soil. Soil Boil Biochem. 2016; 97:188-98.
681 https://doi.org/10.1016/j.soilbio.2016.03.017.
682 55. Vick-Majors TJ, Priscu JC, Amaral-Zettler LA. Modular community structure
683 suggests metabolic plasticity during the transition to polar night in ice-covered
684 antarctic lakes. ISME J. 2014;8(4):778-89. https://doi.org/10.1038/ismej.2013.190.
685 56. Ning D, Deng Y, Tiedje JM, Zhou J. A general framework for quantitatively
686 assessing ecological stochasticity. Proc Nat Acad Sci. 2019;116(34):16892-8.
687 https://doi.org/10.1073/pnas.1904623116.
688 57. Nygaard AB, Tunsjø HS, Meisal R, Charnock C. A preliminary study on the
689 potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to
690 characterize building-dust microbiomes. Sci Rep. 2020;10(1).
691 https://doi.org/10.1038/s41598-020-59771-0.
692 58. Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene
693 sequence regions for phylogenetic analysis of microbiomes. J Microbiol Meth.
694 2011;84(1):81-7. https://doi.org/10.1016/j.mimet.2010.10.020.
695 59. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk H, Gophna U, et al.
696 Harnessing the landscape of microbial culture media to predict new organism–
697 media pairings. Nat Commun. 2015;6(1). https://doi.org/10.1038/ncomms9493.
698 60. Gostincar C, Grube M, de Hoog S, Zalar P, Gunde-Cimerman N.
699 Extremotolerance in fungi: evolution on the edge. FEMS Microbiol Ecol.
700 2010;71(1):2-11. https://doi.org/10.1111/j.1574-6941.2009.00794.x.
701 61. Ani E, Adekunle AA, Kadiri AB, Njoku KL. Rhizoremediation of hydrocarbon
702 contaminated soil using *Luffa aegyptiaca* (Mill) and associated fungi. Int J
703 Phytoremediation. 2021:1-13. https://doi.org/10.1080/15226514.2021.1901852.

62. Martorell MM, Ruberto LAM, de Castellanos LIF, Mac Cormack WP. Bioremediation abilities of antarctic fungi. Fungi in Extreme Environments: Ecological Role and Biotechnological Significance. Springer, Cham. 2019:517-534. https://doi.org/10.1007/978-3-030-19030-9_26.

63. Kannangara S, Ambadeniya P, Undugoda L, Abeywickrama K. Polyaromatic hydrocarbon degradation of moss endophytic fungi isolated from *Macromitrium* sp. In Sri Lanka. Journal of Agricultural Science and Technology A. 2016;6(03):171-182. https://doi.org/10.17265/2161-6256/2016.03.004.

64. Babu AG, Shim J, Bang K, Shea PJ, Oh B. *Trichoderma virens* PDR-28: a heavy metal-tolerant and plant growth-promoting fungus for remediation and bioenergy crop production on mine tailing soil. J Environ Manage. 2014,132:129-34. https://doi.org/10.1016/j.jenvman.2013.10.009.

Figures legends

Fig. 1 Statistical comparison the number of taxa identified between full-length 16S rRNA/ITS gene amplicon sequencing using PacBio Sequel and short-read amplicon sequencing using the Illumina platform. Venn diagram showing the shared and specific taxonomic units assigned at the Phylum, Class, Order, Family, Genus and Species levels between PacBio and Illumina sequencing data. The box plots show the number of taxa per sample, and the Venn diagrams the total number of taxa across all 12 samples. The red asterisks (**) indicate full-length 16S rRNA/ITS gene amplicon sequenceing is significantly identified more taxa than short-read sequenceing (*P* value < 0.01, Wilcoxon test).

Fig. 2 Correlation of identified taxa at (a) the genus of bacteria, (b) species of bacteria, (c) genus of fungi, and (d) species of fungi between full-length sequencing using PacBio Sequel and short read amplicon sequencing using the Illumina platform.for all 12 samples. The dashed lines mark a 0.01% relative abundance threshold for each taxon for Pacbio and Illumina sequence data. Spearman rank correlation was used to compare the samples microbial community compositions as revealed by the sequencing platforms at the level of Genus and Species.

Fig. 3 Comparison of α-diversity (a, d) and β-diversity (b, c, e, f) between Pacbio (P) and Illumina (I) platforms. (a) Shannon index and Pielou index of soil bacteria in the two sequencing platforms, the red asterisks (**) indicate the α-diversity index obtained by full-length sequencing was significantly higher than short-read sequencing (*P* value <0.01, Wilcoxon test). (d) Shannon index and Pielou index of soil fungi, and there was no significant difference between two sequencing platforms. (b) and (e) are principal coordinate analysis (PCoA) of bacterial (b) / fungi (e) community structure based on Bray-Curtis distance matrices. (c) and (f)

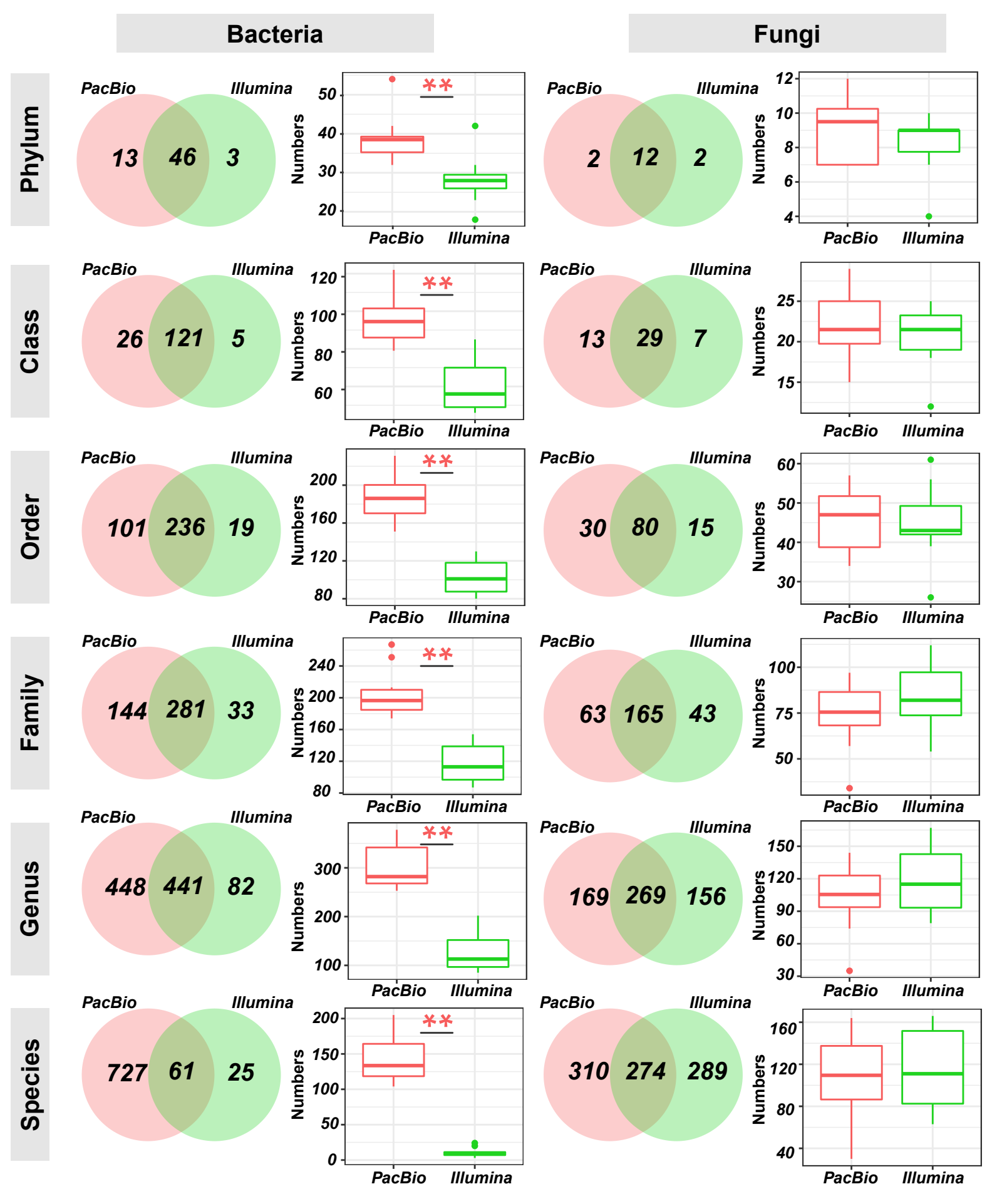
31

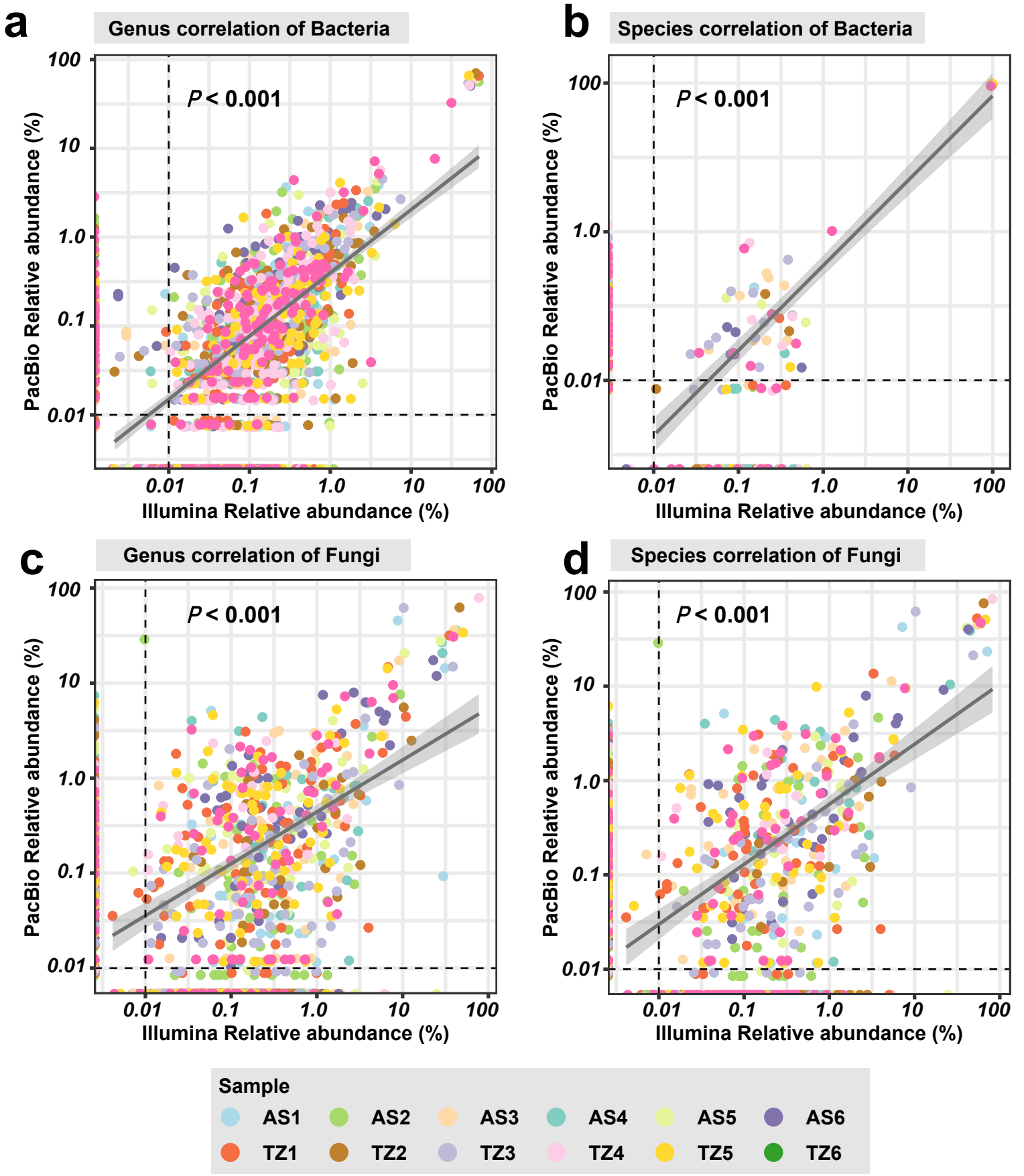
741 are non-metric multidimensional scaling (NMDS) analysis of bacterial (c) /
742 fungus (f) community structure base on Bray-Curtis dissimilarity.
743 **Fig. 4** Taxonomic and functional composition of abundant and rare sub-communities
744 in soil of contaminated sites. (a) and (c) are chord diagram to show the
745 community compositions of abundant and rare sub-communities at phylum level
746 in bacterial (a) / fungal (c) community. (b) The relative abundance of community
747 groups of functional traits in abundant and rare subcommunities based on the
748 FOAM Database level 1. The error bars represent standard deviation of sample
749 replicates and red asterisks (**) indicate metabolic categories that are
750 significantly more predominant in abundant subcommunity (p value < 0.01,
751 Wilcoxon test) and green asterisks indicate categories that are significantly more
752 predominant in rare subcommunity (double asterisks, $p < 0.01$; single asterisk, $p <$
753 0.05, Wilcoxon test). (d) Pie plot showing the functional guilds of fungal sub-
754 communities.
755 **Fig. 5** Heatmap of Spearman's rank correlation between main species(bacteria(a) and
756 fungi(b)) and chemical characteristics (main pollutants) in soil of contaminated
757 site. Spearman's rank correlation coefficient ranges from 1.0 to–1.0, from a
758 strongly positive to a strongly negative correlation. $*p < 0.05$, $**p < 0.01$. LR
759 PAHs means Low-ring PAHs (two and three rings, including Nap, AcPy, Ace,
760 Flu, Phe, and Ant), and HR PAHs means High-ring PAHs (five and six rings,
761 including BaP, IND, DBA, and BghiP).
762 **Fig. 6** Co-occurrence networks of abundant, intermediate, and rare taxa in the soil of
763 contaminated site. (a) and (b) are the bacterial-fungal co-occurrence network. (a)
764 was established by calculating correlations among abundant, rare and intermediate
765 ASVs, and the nodes of (b) were colored according nodes belong to bacteria or

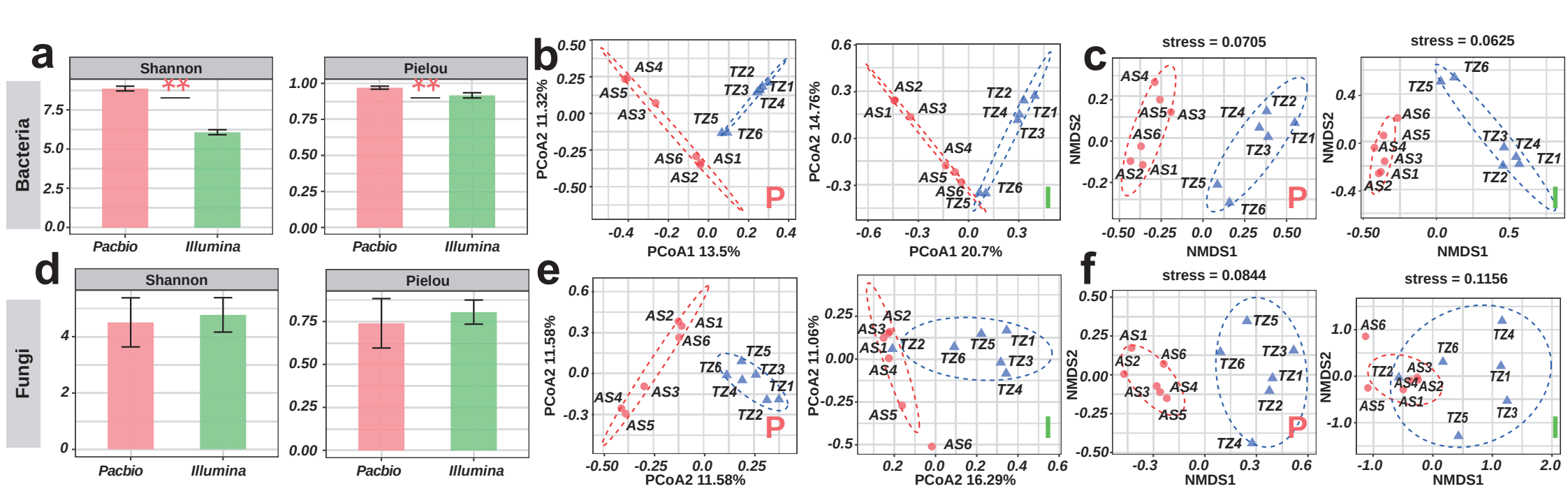
fungi. A connection indicates a strong (Spearman's $\rho > 0.6$) and significant (FDR-corrected $p < 0.01$) correlation. The size of each node is proportional to the absolute abundance of the ASVs. Red lines show positive correlations, while green lines show negative correlations. (c) Node-level topological feature of different sub-communities and the comparison of node-level topological features (degree and closeness centrality) among different sub-communities. The table show the network topological features after exclude keystone fungi. (d) and (e) are the bacterial co-occurrence network in Anshan (d) and Taizhou(e), while (f) and (g) are the fungal co-occurrence network in Anshan (f) and Taizhou (g). The nodes that the blue arrow points to are the keystone species (top five based on the betweenness centrality score) in the network.

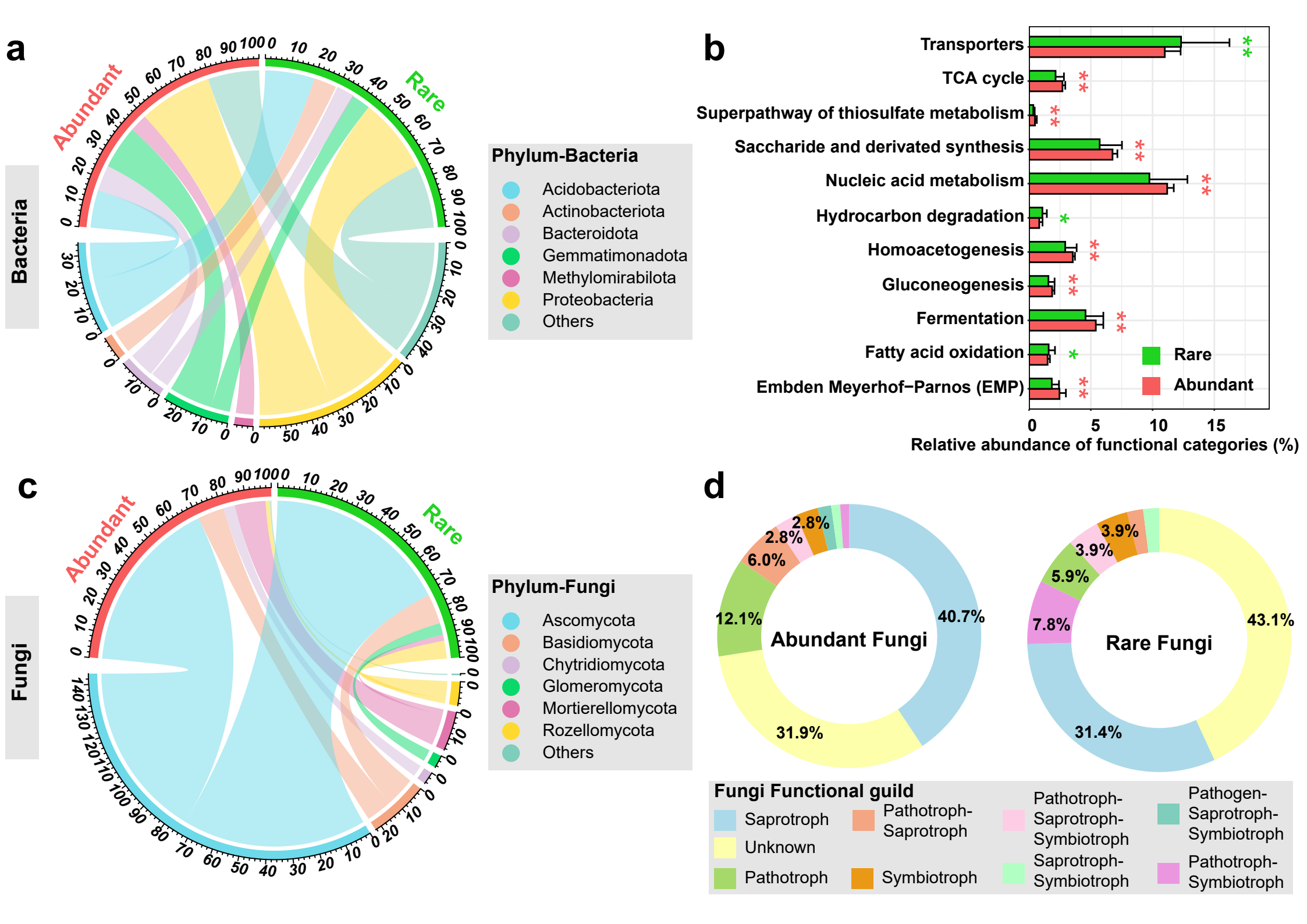
Fig. 7 Relationships between environmental variables and the microbial community and community assembly mechanism. (a) The correlations between microbial communities, soil physicochemical characteristics. Different microbial communities were related to each environmental factor by Mantel tests. Edge width corresponds to the Mantel's r statistic, and edge color denotes the statistical significance based on 999 permutations. The proportion of the pie indicates correlation strength, with a higher proportion representing higher correlation strength. (B) Box plot showing the differences in the bacterial and fungal NST values under different sub-communities, and different letter indicated significant difference ($p < 0.001$, Kruskal-Wallis test).

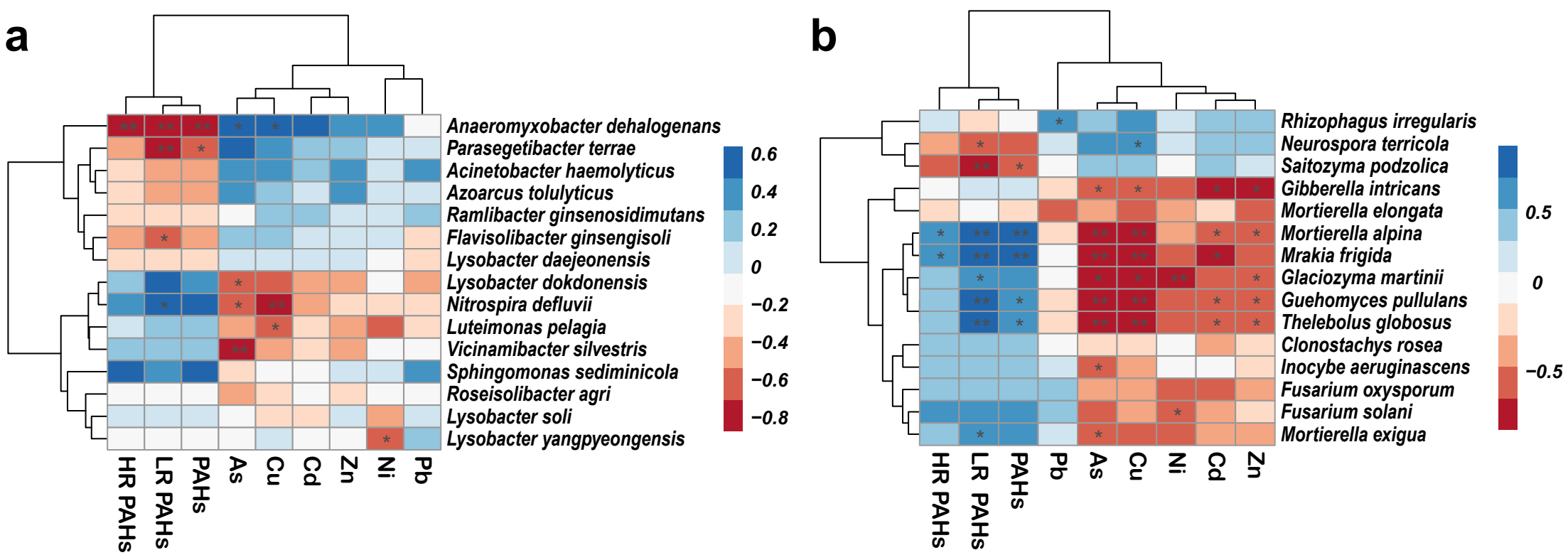
33

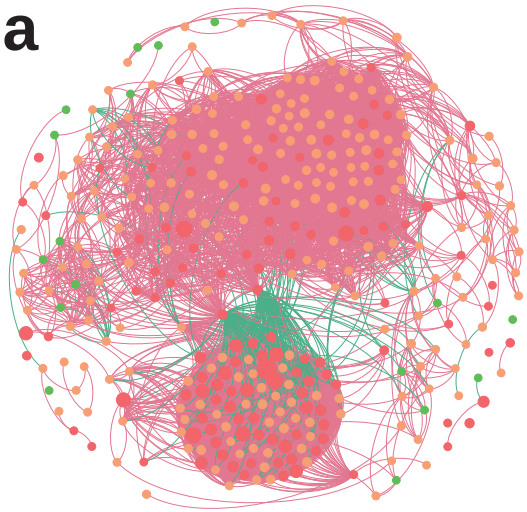

Bacteria
Fungi
Phylum
Class
Order
Family
Genus
Species
PacBio
Illumina
13 46 3
2 12 2
26 121 5
13 29 7
101 236 19
30 80 15
144 281 33
63 165 43
448 441 82
169 269 156
727 61 25
310 274 289
Numbers
**

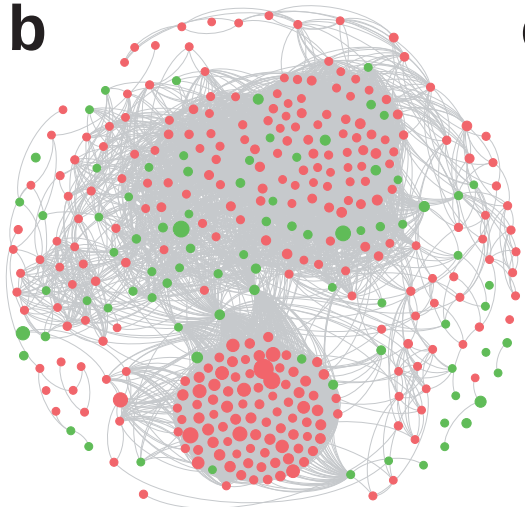

a
Genus correlation of Bacteria
P < 0.001
PacBio Relative abundance (%)
Illumina Relative abundance (%)
b
Species correlation of Bacteria
P < 0.001
PacBio Relative abundance (%)
Illumina Relative abundance (%)
c
Genus correlation of Fungi
P < 0.001
PacBio Relative abundance (%)
Illumina Relative abundance (%)
d
Species correlation of Fungi
P < 0.001
PacBio Relative abundance (%)
Illumina Relative abundance (%)
100
10
1.0
0.1
0.01
Sample
AS1
AS2
AS3
AS4
AS5
AS6
TZ1
TZ2
TZ3
TZ4
TZ5
TZ6

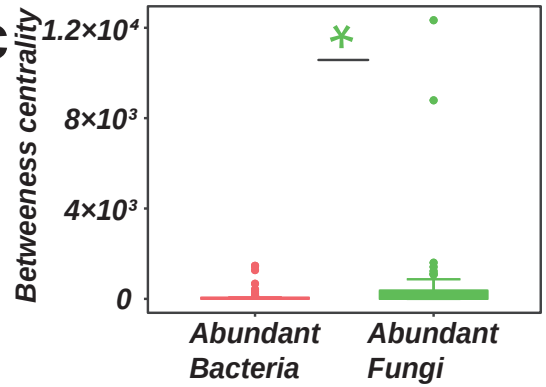

a
Bacteria
Shannon
Pielou
Pacbio
Illumina
b
PCoA2 11.32%
PCoA1 13.5%
PCoA2 14.76%
PCoA1 20.7%
c
stress = 0.0705
stress = 0.0625
NMDS1
NMDS2
d
Fungi
e
PCoA2 11.58%
PCoA1 11.58%
PCoA2 11.06%
PCoA2 16.29%
f
stress = 0.0844
stress = 0.1156
P


a
Bacteria
Abundant
Rare
Phylum-Bacteria
Acidobacteriota
Actinobacteriota
Bacteroidota
Gemmatimonadota
Methylomirabilota
Proteobacteria
Others
b
Transporters
TCA cycle
Superpathway of thiosulfate metabolism
Saccharide and derivated synthesis
Nucleic acid metabolism
Hydrocarbon degradation
Homoacetogenesis
Gluconeogenesis
Fermentation
Fatty acid oxidation
Embden Meyerhof−Parnos (EMP)
Rare
Abundant
0
5
10
15
Relative abundance of functional categories (%)
c
Fungi
Abundant
Rare
Phylum-Fungi
Ascomycota
Basidiomycota
Chytridiomycota
Glomeromycota
Mortierellomycota
Rozellomycota
Others
d
Abundant Fungi
40.7%
31.9%
12.1%
6.0%
2.8%
2.8%
Rare Fungi
43.1%
31.4%
7.8%
5.9%
3.9%
3.9%
Fungi Functional guild
Saprotroph
Unknown
Pathotroph
Pathotroph-Saprotroph
Symbiotroph
Pathotroph-Saprotroph-Symbiotroph
Saprotroph-Symbiotroph
Pathogen-Saprotroph-Symbiotroph
Pathotroph-Symbiotroph

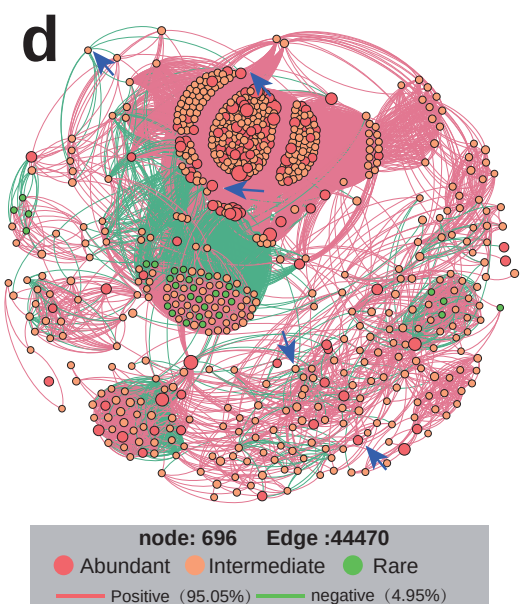

a
Anaeromyxobacter dehalogenans
Parasegetibacter terrae
Acinetobacter haemolyticus
Azoarcus tolulyticus
Ramlibacter ginsenosidimutans
Flavisolibacter ginsengisoli
Lysobacter daejeonensis
Lysobacter dokdonensis
Nitrospira defluvii
Luteimonas pelagia
Vicinamibacter silvestris
Sphingomonas sediminicola
Roseisolibacter agri
Lysobacter soli
Lysobacter yangpyeongensis
HR PAHs
LR PAHs
PAHs
As
Cu
Cd
Zn
Ni
Pb
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
b
Rhizophagus irregularis
Neurospora terricola
Saitozyma podzolica
Gibberella intricans
Mortierella elongata
Mortierella alpina
Mrakia frigida
Glaciozyma martinii
Guehomyces pullulans
Thelebolus globosus
Clonostachys rosea
Inocybe aeruginascens
Fusarium oxysporum
Fusarium solani
Mortierella exigua
HR PAHs
LR PAHs
PAHs
Pb
As
Cu
Ni
Cd
Zn
0.5
0
−0.5

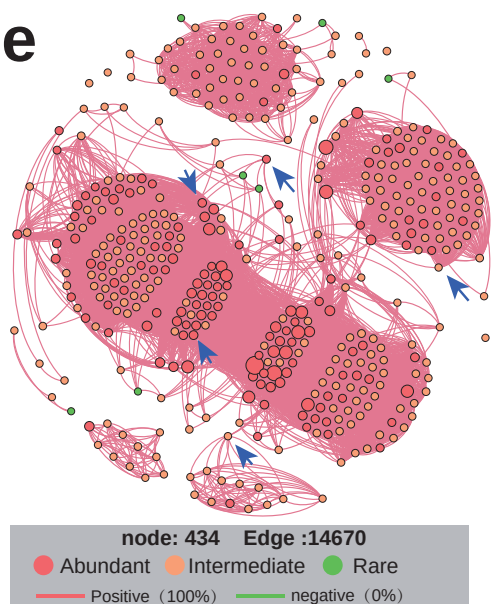


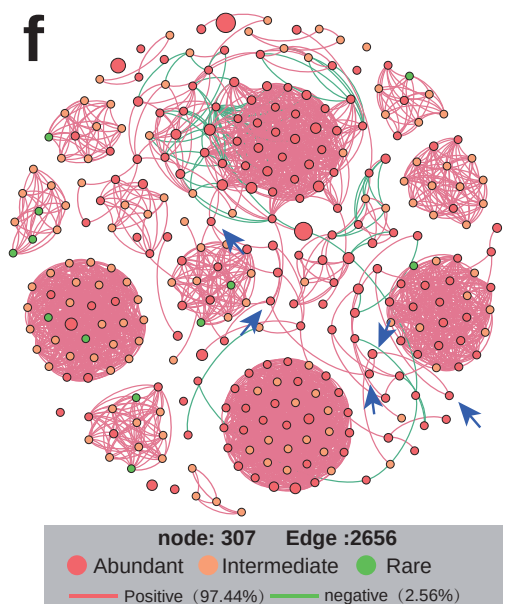


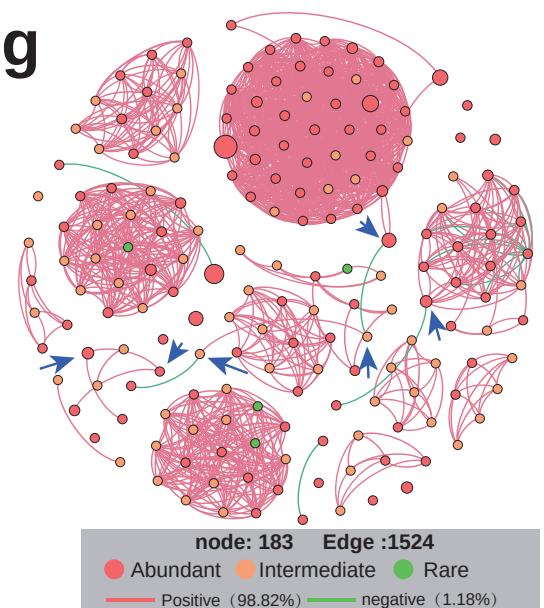


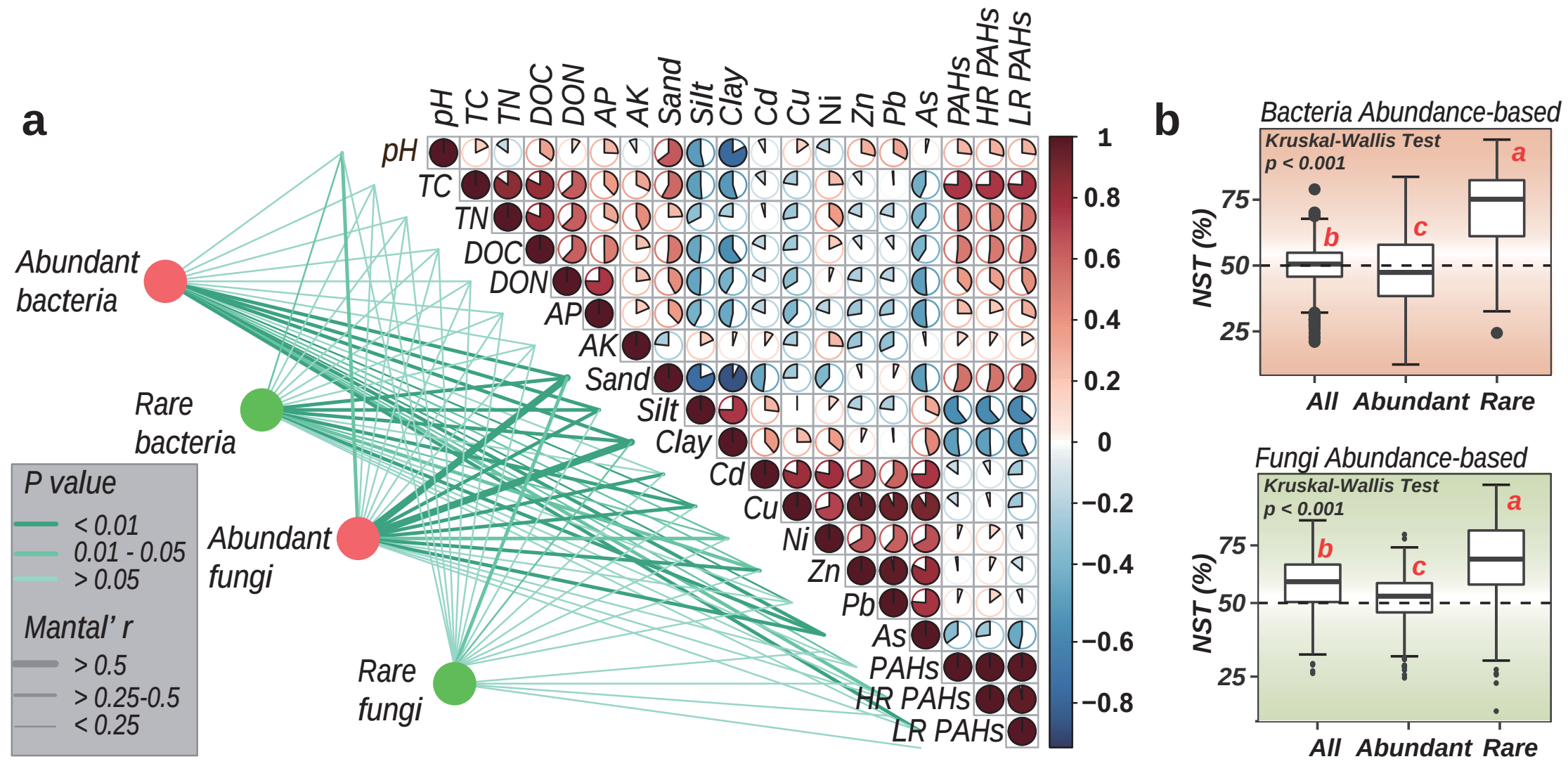
Network	Edges	Avg. Degree	Avg. Weighted Degree	Avg. Clustering Coefficient	Avg. Eccentrivity
Origianl	9786	52.788	46.014	0.735	6.42
Exclude keystone fungi	8462	48.217	42.712	0.749	6.61










a
pH
TC
TN
DOC
DON
AP
AK
Sand
Silt
Clay
Cd
Cu
Ni
Zn
Pb
As
PAHs
HR PAHs
LR PAHs
Abundant bacteria
Rare bacteria
Abundant fungi
Rare fungi
P value
< 0.01
0.01 - 0.05
> 0.05
Mantal' r
> 0.5
> 0.25-0.5
< 0.25
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
b
Bacteria Abundance-based
Kruskal-Wallis Test
p < 0.001
NST (%)
75
50
25
a
b
c
All
Abundant
Rare
Fungi Abundance-based
Kruskal-Wallis Test
p < 0.001
NST (%)
75
50
25
a
b
c
All
Abundant
Rare