



HAL
open science

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.

► To cite this version:

Kang Yan, Jiahang Zhou, Cong Feng, Suyuan Wang, Bart Haegeman, et al.. Abundant fungi dominate the coexistence of microbiota in soil of contaminated site: High-precision community analysis by full-length sequencing. *Science of the Total Environment*, 2023, 861, pp.160563. 10.1016/j.scitotenv.2022.160563 . hal-04219537

HAL Id: hal-04219537

<https://hal.science/hal-04219537v1>

Submitted on 27 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3

4 Kang Yan¹, Jiahang Zhou¹, Cong Feng², Suyuan Wang¹, Bart Haegeman³, Weirong
5 Zhang¹, Jian Chen⁴, Shouqing Zhao⁴, Jiangmin Zhou⁵, Jianming Xu¹, Haizhen Wang^{1,*}

6

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

10 ² Department of Bioinformatics, College of Life Sciences, Zhejiang University,
11 Hangzhou 310058, China

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

14 ⁴ Plant Protection, Fertilizer and Rural Energy Agency of Wenling, Zhejiang Province,
15 Wenling 317500, China

16 ⁵ College of Life and Environmental Sciences, Wenzhou University, Wenzhou 325035,
17 Zhejiang, China

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

19 **Abstract**

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

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

44 **1. Introduction**

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

61 Unfortunately, the use of the 16S rRNA or internal transcribed spacer (ITS) gene
62 as a taxonomic marker has been constrained by the short-read sequencing from
63 Illumina platform, which is the most commonly used second-generation sequencing for
64 microbial community profiling [13-15]. The majority of contemporary 16S rRNA or
65 ITS gene sequence information from short-read sequencing platforms limited the
66 classification of microbiota below the genus level [15]. Species-level identification and
67 functional assignment of microbiota are core objectives of microbial and soil ecology
68 [16-17]. Third-generation sequencing platforms, such as Pacific Biosciences (PacBio)

69 and Oxford Nanopore Technologies, have developed new technologies with full-length
70 sequencing [14,18]. Full-length sequencing can dramatically increase the accuracy of
71 taxonomic assignments on previously unobtainable scales, and has the potential to
72 facilitate microbial community studies [13, 15]. However, a major shortcoming of full-
73 length sequencing is its higher error rate (~ 10%) compared to short-read sequencing
74 (~ 0.5%) [14, 19]. For PacBio platform, this issue can be solved through construction
75 of a circular consensus sequence (CCS), in which individual amplicon molecules are
76 sequenced multiple times using circularized library templates that provide consensus-
77 sequence error correction [19-20]. In this way, the error rate of a CCS-generated
78 amplicon can be reduced to a level comparable to that of short-read platform [19-20].

79 In the natural ecosystem, soil microbial community typically shows a skewed taxa
80 abundance distribution, with relatively few abundant taxa and a high number of rare
81 taxa [21-23]. Clarifying the factors that influence the assembly of different
82 subcommunities (abundant taxa and rare taxa) in the soil of the contaminated site is of
83 great significance for ecosystem restoration [23]. In many studies, second-generation
84 sequencing has been used for direct identification of rare and abundant microbes with
85 the relative abundance and the microbial diversity in these subcommunities [23-28].
86 For example, the relative abundance of < 0.01% has been often used to define rare taxa
87 [22-23]. However, the heterogeneity of soil might result in large differences of
88 microbial biomass among the various samples [7-9]. Namely, the absolute abundances
89 of taxa related to the relative abundance of 0.01% were different among the various
90 samples. Moreover, the studies reviewed that the trends of the relative and absolute
91 abundances of some taxa in community were inconsistent by using the integrated high-
92 throughput absolute abundance quantification (iHAAQ) method [29]. A potential bias
93 existed in the use of relative abundances to facilitate the comparison across

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

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

109

110 **2. Material and Methods**

111 **2.1 Soil sampling and physicochemical analysis**

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

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

121 Soil physicochemical properties were determined according to previous methods
122 [34]. Specifically, soil pH was measured by pH meter (pHSJ-3F, Leici, China) using a
123 1:2.5 soil/water mixture. The soil total carbon (TC) and total nitrogen (TN) were
124 determined using an elemental analyser (Elementar Analysensysteme Gmb H,
125 Germany). The concentrations of dissolved organic carbon (DOC) and dissolved
126 organic nitrogen (DON) were extracted by 0.5 mol/L K₂SO₄ for 30 min and filtered,
127 and the extracts were analyzed using high-temperature combustion (Multi N/C 3100,
128 Analytik Jena AG, Jena, Germany) [35]. Available phosphorus (AP) was extracted by
129 NH₄F-HCl or NaHCO₃ (based on soil pH), and was determined colorimetrically at 880
130 nm. Available potassium (AK) was extracted using 1 mol/L NH₄OAc and analyzed by
131 flame atomic-absorption spectrophotometry. Soil texture composition was measured
132 using the pipette method. For the total heavy metals (Cd, Cu, Ni, Zn, Pb, and As), a
133 0.2-g soil sample was digested with an acid mixture of HNO₃, HF, and H₂O₂ (volume
134 ratio = 4:2:2) in a microwave digester (MARS6, CEM Microwave Technology Ltd.,
135 USA), then determined through ICP-MS (ICP-MS NEXION300XX, PerkinElmer, Inc.,
136 USA) [36]. The extraction of polycyclic aromatic hydrocarbons (PAHs) in soil samples
137 using the ultrasonic agitation for 30 min with 10 mL of 1:1 acetone/n-hexane (v/v) [37],
138 and were analyzed by GC-MS (N6890/5975B, Agilent, USA). The results of
139 physicochemical properties and each pollutant were shown in Table S1-S2

140 **2.2 Soil microbial DNA extraction and sequencing**

141 Soil DNA was extracted from soil samples using the TGuide S96 Magnetic
142 Soil/Stool DNA Kit (DP812, TIANGEN BIOTECH, China), following the
143 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 μ L, 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 30 μ L 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

169 μL) contained 5 μL of targeted PCR product, 2.5 μL of MPPI-a (2 μM), 2.5 μL of
170 MPPI-b (2 μM), and 10 μL of 2 \times Q5 HF MM. The PCR protocol was as follows: an
171 initial denaturation at 98 $^{\circ}\text{C}$ for 30 s; 10 cycles of 10 s at 98 $^{\circ}\text{C}$, 30 s at 65 $^{\circ}\text{C}$, and 30 s
172 at 72 $^{\circ}\text{C}$; and a final extension at 72 $^{\circ}\text{C}$ for 5 min. The amplified DNA for each
173 samples were sequenced by Illumina NovaSeq6000 platform (Illumina, USA).

174 **2.3 qPCR determination**

175 Each DNA sample with three analytical replicates (Roche LightCycler 480 Real-
176 Time PCR Machine) was performed to determine the copy numbers of 16S rDNA and
177 ITS (internal transcribed spacer) region of rDNA [38]. Primers and the qPCR
178 conditions were listed in Table S3. Amplifications were performed in a final volume of
179 20 μL containing 10 μL of SYBR Green qPCR SuperMix UDG (Invitrogen, USA), 1
180 μL DNA, 0.4 μL of each primer, and 8.2 μL nucleic acid-free water. Plasmids carrying
181 the corresponding gene were used to establish qPCR standard curves (additional details
182 in Table S1). All runs had standard efficiency curves of $R^2 > 0.999$ and the efficiencies
183 reached 90-110% (result were in Table S4).

184 **2.4 Analysis of sequencing outputs**

185 For the full-length sequencing data, circular consensus sequence was clustered
186 into amplicon sequence variants (ASVs) with the software packages DADA2 (version
187 1.18.0) in R v4.0.3 [39-41]. Briefly, primers were detected and removed from reads
188 using the removePrimers function. The remaining sequences were filtered using the
189 filterAndTrim function [39]. Then, the ASVs were inferred after learning error rates.
190 The learnErrors method learns this error model from the data, by alternating estimation
191 of the error rates and inference of sample composition until they converge on a jointly
192 consistent solution [40]. The learnErrors function with parameter:
193 errorEstimationFunction=PacBioErrfun, BAND_SIZE=32 [39-40]. And the dada

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

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

216 **2.5 Data analysis and statistics**

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

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

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

$$231 \quad AA_{i,s} = RA_{i,s} \times Q_s \quad (1)$$

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

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

$$240 \quad \overline{AA}_i = \sum_{s=1}^{12} AA_{i,s} \quad (2)$$

$$241 \quad \overline{Q} = \sum_{s=1}^{12} Q_s \quad (3)$$

242 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
243 taxon. The taxon i was subjected to the intermediate taxa when its copy number
244 between the thresholds of rare and abundant taxa.

245 Each ASV across the dataset was analyzed by using the PICRUSt2 pipeline to get
246 a deeper understanding of the metabolic profiles of abundant and rare bacterial
247 subcommunities [49], and the program FUNGuild was used to predict fungal
248 functional guilds [50]. A network analysis was performed to explore the co-occurrence
249 patterns between different subcommunities. The ASVs with low frequency were
250 removed from network analysis to avoid possible biases, as a large number of zeros
251 values may introduce spurious correlations [27,51]. The correlation matrix between
252 two ASVs was calculated in “psych (version 2.0.12)” [52]. Spearman's correlation
253 coefficient was >0.6 and the p -value was <0.01 were integrated into network analysis,
254 and the several network properties (degree, betweenness centrality, closeness centrality
255 and eigencentrality) were calculated with the interactive platform Gephi [53]. The
256 ASVs with the highest betweenness centrality scores were considered keystone [54-
257 55]. The normalized stochasticity ratio (NST) was utilized to estimate the determinacy
258 and stochasticity of the different subcommunities assembly processes with high
259 accuracy and precision [56]. The NST values were used to evaluate the deterministic
260 ($<50\%$) or more stochastic ($>50\%$) assembly processes of soil microbiota. We
261 calculated NST variations based on Ruzicka (abundance-based) dissimilarity metrics in
262 the “NST” package (version 3.0.6) [56].

263

264 **3. Results**

265 **3.1 Taxonomic analysis of different platforms**

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

274 Full-length 16S rRNA gene sequencing from PacBio platform gave better
275 resolution than the short-read from Illumina platform for bacterial identification at all
276 levels, especially at the level of genus and species ($p < 0.01$) (Fig. 1, Fig. S2-S3). In the
277 12 test soil samples, the PacBio platform detected 59 phyla, 147 classes, 337 orders,
278 425 families, 889 genera and 788 species of bacteria, while 49 phyla, 126 classes, 258
279 orders, 314 families, 523 genera, and 86 species were detected by the Illumina platform
280 (Fig. 1). For fungi, the PacBio platform detected 14 phyla, 42 classes, 110 orders, 228
281 families, 438 genera and 584 species, while 14 phyla, 36 classes, 95 orders, 208
282 families, 425 genera, and 563 species were detected by the Illumina platform (Fig. 1).
283 Full-length ITS sequencing had identified more taxa than Illumina ITS amplicon
284 sequencing at all taxonomic levels, but there had no significant difference ($p > 0.05$).

285 **3.2 Microbial community composition revealed by two platform**

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

291 of phylum (rho-value 0.665-0.909, $p < 0.001$), class (rho-value 0.511-0.755, $p <$
292 0.001), order (rho-value 0.365-0.725, $p < 0.001$) and family (rho-value 0.345-0.744, p
293 < 0.001) (Fig. S4-S7). However, the fungal compositions of some soil samples at the
294 levels of phylum to family obtained from the two sequencing platforms were different
295 ($p > 0.05$) (Fig. S8-S11).

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

307 Moreover, the comparisons of alpha-diversity indices (Shannon and Pielou) of the
308 soil bacteria were significantly different between the two platforms (Fig. 3). It was
309 demonstrated that the PacBio platform exhibited a significantly higher level of
310 bacterial α -diversity than the Illumina platform ($p < 0.01$). Both the results of full-
311 length sequencing and short-read sequencing can reveal the beta-diversity differences
312 between the bacterial communities of the two site samples, with the higher
313 dissimilarity values of full-length sequencing (ANOSIM test, $R = 0.9352$ vs $R =$
314 0.8056). And for fungi, there is no significant difference in alpha-diversity indices
315 (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 while 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 ($> 2.07 \times 10^8$ copies g^{-1}), the rare taxa ($< 2.07 \times 10^7$ 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

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

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

354 **3.4. Co-occurrence patterns and assembly mechanism of different** 355 **subcommunities**

356 Based on the correlation results, we further performed network analysis to
357 disentangle the ecological role and co-occurrence patterns of abundant and rare
358 subcommunities in the soil of contaminated site. Results showed that the average value
359 of topological properties of abundant taxa was higher than that of rare taxa, suggesting
360 that the abundant taxa were located in central positions within the network more often
361 than the rare taxa (Fig. 6a, d-g). Moreover, the bacterial-fungal co-occurrence networks
362 were composed of 9786 edges and 371 nodes, which included 77.63% bacteria and
363 22.37% fungi, and their average absolute abundance were 3.56×10^7 copies/g and
364 3.08×10^7 copies/g, respectively (Fig. 6b). Species with the highest betweenness
365 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). *Mortierella 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) play 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).

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

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

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

453 Further analysis of the microbial co-occurrence patterns indicated that abundant
454 fungi were the critical network nodes in soil microbial associations. And the rare taxa
455 were not evenly distributed throughout the soil of contaminated site, most of them
456 were only present in a few samples (Fig. 5). This could be explained by the fact that
457 abundant fungi can effectively utilize a broader range of resources than others [23, 28].
458 Moreover, compared to the bacteria, fungi have a high tolerance of extreme conditions
459 such as acidic or alkaline pH, low moisture, and higher concentration of metals, etc
460 [60]. In this study, the abundant fungi *Mortierella alpine*, *Fusarium solani*, *Mrakia*
461 *frigida*, *Nigrospora oryzae*, *Trichoderma virens*, and *Curvularia lunata* had the highest
462 betweenness centrality scores and were regarded as keystone species in the co-
463 occurrence network (Table S6). The previous studies have pointed out that these
464 microorganisms have the ability to degrade PAHs, resist heavy metal, and even can act

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

477

478 **5. Conclusion**

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

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

496

497 **Declaration of Competing Interest**

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

499

500 **Acknowledgements**

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

505

506 **References**

- 507 1. Chen R, Ye C. Resolving soil pollution in China. *Nature*. 2014;505(7484):483.
508 <https://doi.org/10.1038/505483c>.
- 509 2. Yao Y. Spend more on soil clean-up in China. *Nature*. 2016;533(7604):469.
510 <https://doi.org/10.1038/533469a>.
- 511 3. Niu S, Tao W, Chen R, Hageman KJ, Zhu C, Zheng R, et al. Using
512 polychlorinated naphthalene concentrations in the soil from a southeast China e-
513 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

- 538 indicates the relationship between the innate microbiota and contaminant fractions.
539 Environ Sci Technol. 2017;51(16):9165-75.
540 <https://doi.org/10.1021/acs.est.7b00294>.
- 541 11. Jiao S, Liu Z, Lin Y, Yang J, Chen W, Wei G. Bacterial communities in oil
542 contaminated soils: biogeography and co-occurrence patterns. *Soil Boil Biochem.*
543 2016; 98:64-73. <https://doi.org/10.1016/j.soilbio.2016.04.005>.
- 544 12. Li D, Li G, Zhang D. Field-scale studies on the change of soil microbial
545 community structure and functions after stabilization at a chromium-contaminated
546 site. *J Hazard Mater.* 2021; 415:125727.
547 <https://doi.org/10.1016/j.jhazmat.2021.125727>.
- 548 13. Lam TYC, Mei R, Wu Z, Lee PKH, Liu W, Lee P. Superior resolution
549 characterisation of microbial diversity in anaerobic digesters using full-length 16s
550 rRNA gene amplicon sequencing. *Water Res.* 2020; 178:115815.
551 <https://doi.org/10.1016/j.watres.2020.115815>.
- 552 14. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et
553 al. DNA sequencing at 40: past, present and future. *Nature.* 2017;550(7676):345-
554 53. <https://doi.org/10.1038/nature24286>.
- 555 15. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, et al. Species-level
556 bacterial community profiling of the healthy sinonasal microbiome using pacific
557 biosciences sequencing of full-length 16S rRNA genes. *Microbiome.* 2018;6(1).
558 <https://doi.org/10.1186/s40168-018-0569-2>.
- 559 16. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, et al.
560 High-resolution phylogenetic microbial community profiling. *ISME J.*
561 2016;10(8):2020-32. <https://doi.org/10.1038/ismej.2015.249>.

- 562 17. Johnson JS, Spakowicz DJ, Hong B, Petersen LM, Demkowicz P, Chen L, et al.
563 Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome
564 analysis. *Nat Commun.* 2019;10(1). <https://doi.org/10.1038/s41467-019-13036-1>.
- 565 18. Eisenstein M. An ace in the hole for DNA sequencing. *Nature.*
566 2017;550(7675):285-8. <https://doi.org/10.1038/550285a>.
- 567 19. Fichot EB, Norman RS. Microbial phylogenetic profiling with the pacific
568 biosciences sequencing platform. *Microbiome.* 2013;1(1):10.
569 <https://doi.org/10.1186/2049-2618-1-10>.
- 570 20. Tedersoo L, Tooming Klunderud A, Anslan S. Pacbio metabarcoding of Fungi and
571 other eukaryotes: errors, biases and perspectives. *New Phytol.* 2018;217(3):1370-
572 85. <https://doi.org/10.1111/nph.14776>.
- 573 21. Jia X, Dini-Andreote F, Falcão Salles J. Community assembly processes of the
574 microbial rare biosphere. *Trends Microbiol.* 2018;26(9):738-47.
575 <https://doi.org/10.1016/j.tim.2018.02.011>.
- 576 22. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. *Nat Rev*
577 *Microbiol.* 2015;13(4):217-29. <https://doi.org/10.1038/nrmicro3400>.
- 578 23. Jiao S, Chen W, Wei G. Biogeography and ecological diversity patterns of rare
579 and abundant bacteria in oil-contaminated soils. *Mol Ecol.* 2017;26(19):5305-17.
580 <https://doi.org/10.1111/mec.14218>.
- 581 24. Mo Y, Peng F, Gao X, Xiao P, Logares R, Jeppesen E, et al. Low shifts in salinity
582 determined assembly processes and network stability of microeukaryotic plankton
583 communities in a subtropical urban reservoir. *Microbiome.* 2021;9(1).
584 <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-](https://doi.org/10.1111/1462-2920.14945)
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>.

- 704 62. Martorell MM, Ruberto LAM, de Castellanos LIF, Mac Cormack WP.
705 Bioremediation abilities of antarctic fungi. *Fungi in Extreme Environments:*
706 *Ecological Role and Biotechnological Significance.* Springer, Cham. 2019:517-
707 534. https://doi.org/10.1007/978-3-030-19030-9_26.
- 708 63. Kannangara S, Ambadeniya P, Undugoda L, Abeywickrama K. Polyaromatic
709 hydrocarbon degradation of moss endophytic fungi isolated from *Macromitrium*
710 sp. In Sri Lanka. *Journal of Agricultural Science and Technology A.*
711 2016;6(03):171-182. <https://doi.org/10.17265/2161-6256/2016.03.004>.
- 712 64. Babu AG, Shim J, Bang K, Shea PJ, Oh B. *Trichoderma virens* PDR-28: a heavy
713 metal-tolerant and plant growth-promoting fungus for remediation and bioenergy
714 crop production on mine tailing soil. *J Environ Manage.* 2014,132:129-34.
715 <https://doi.org/10.1016/j.jenvman.2013.10.009>.

716 **Figures legends**

717 **Fig. 1** Statistical comparison the number of taxa identified between full-length 16S
718 rRNA/ITS gene amplicon sequencing using PacBio Sequel and short-read
719 amplicon sequencing using the Illumina platform. Venn diagram showing the
720 shared and specific taxonomic units assigned at the Phylum, Class, Order, Family,
721 Genus and Species levels between PacBio and Illumina sequencing data. The box
722 plots show the number of taxa per sample, and the Venn diagrams the total
723 number of taxa across all 12 samples. The red asterisks (**) indicate full-length
724 16S rRNA/ITS gene amplicon sequencing is significantly identified more taxa
725 than short-read sequencing (P value < 0.01 , Wilcoxon test).

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

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

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

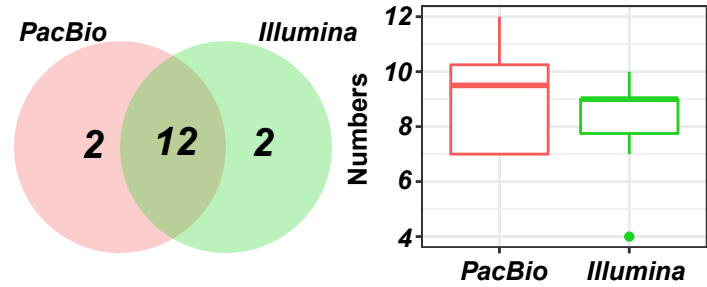
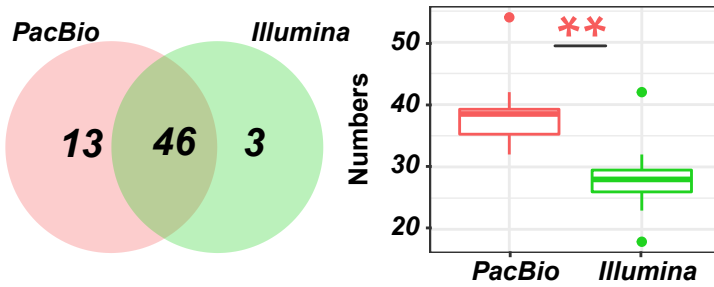
766 fungi. A connection indicates a strong (Spearman's $\rho > 0.6$) and significant (FDR-
767 corrected $p < 0.01$) correlation. The size of each node is proportional to the
768 absolute abundance of the ASVs. Red lines show positive correlations, while
769 green lines show negative correlations. (c)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

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

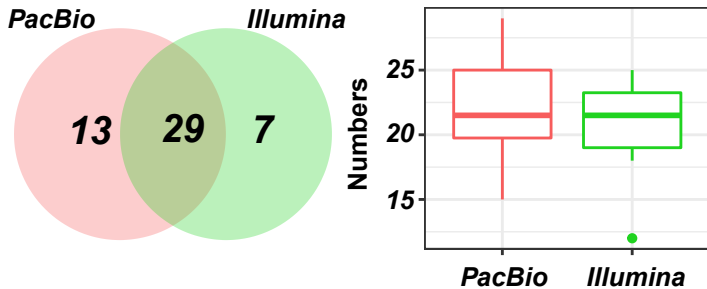
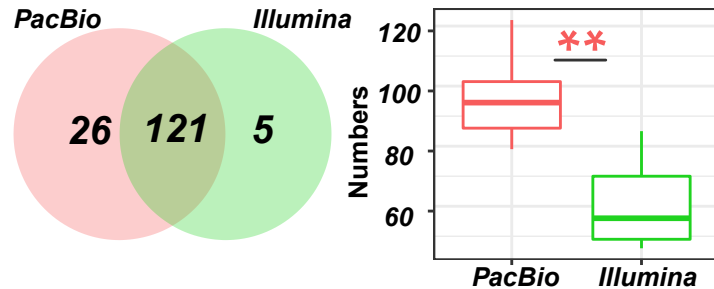
Bacteria

Fungi

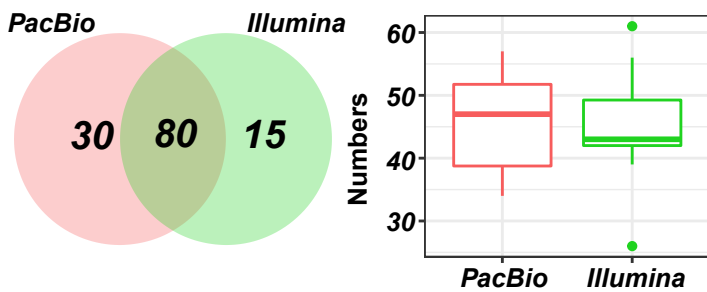
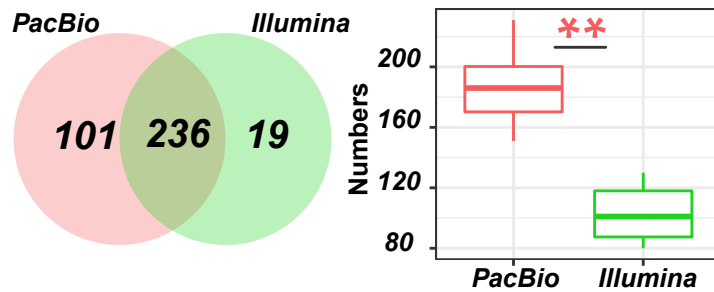
Phylum



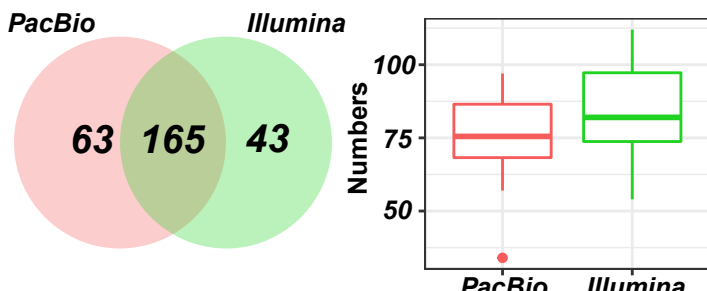
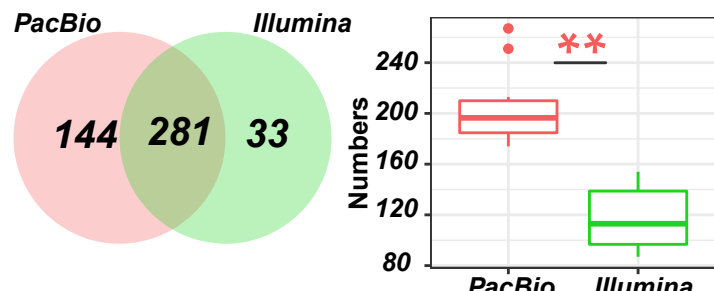
Class



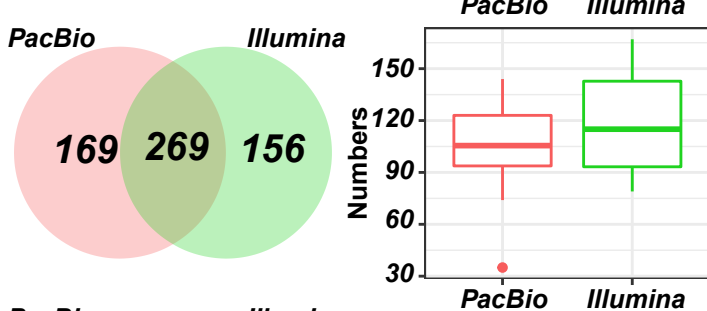
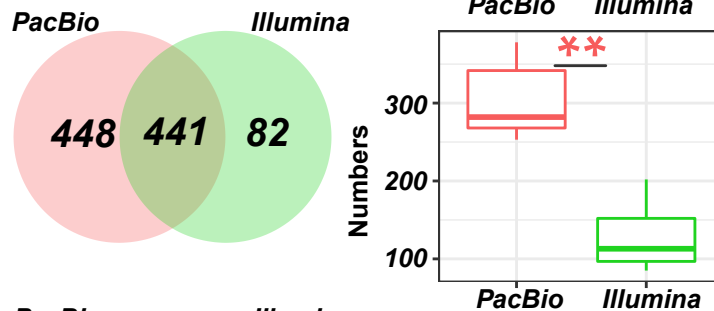
Order



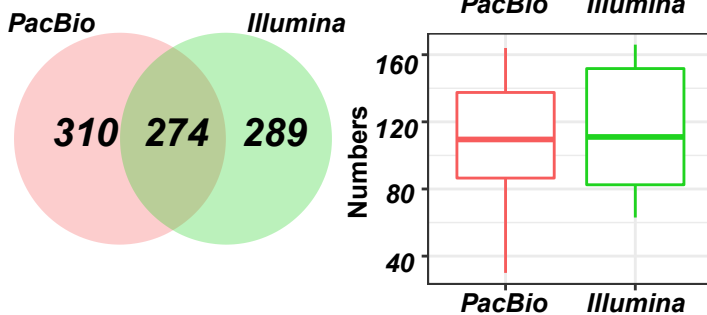
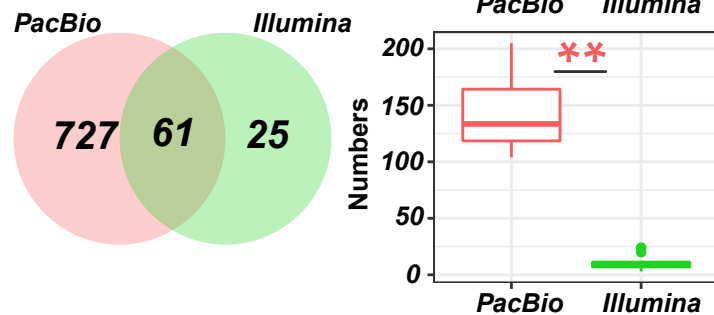
Family

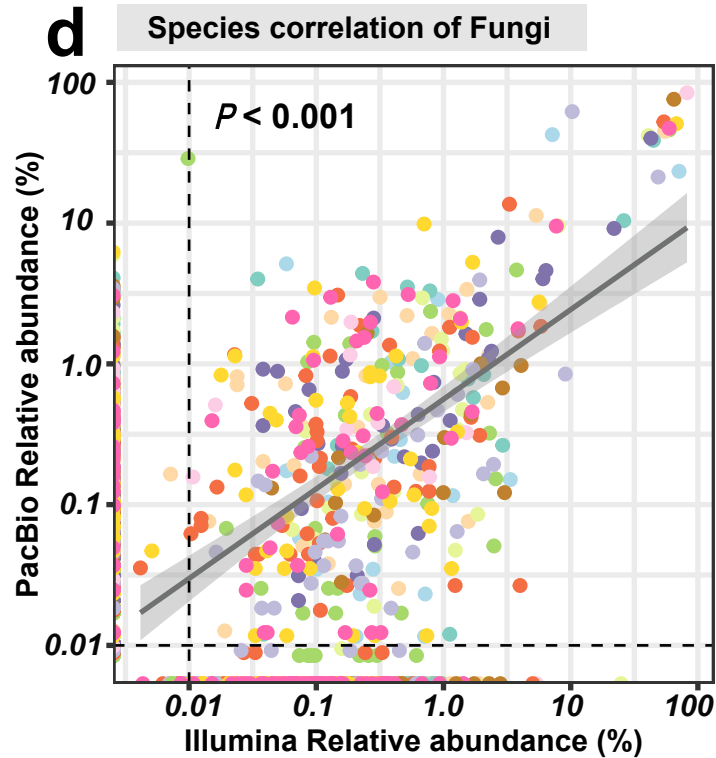
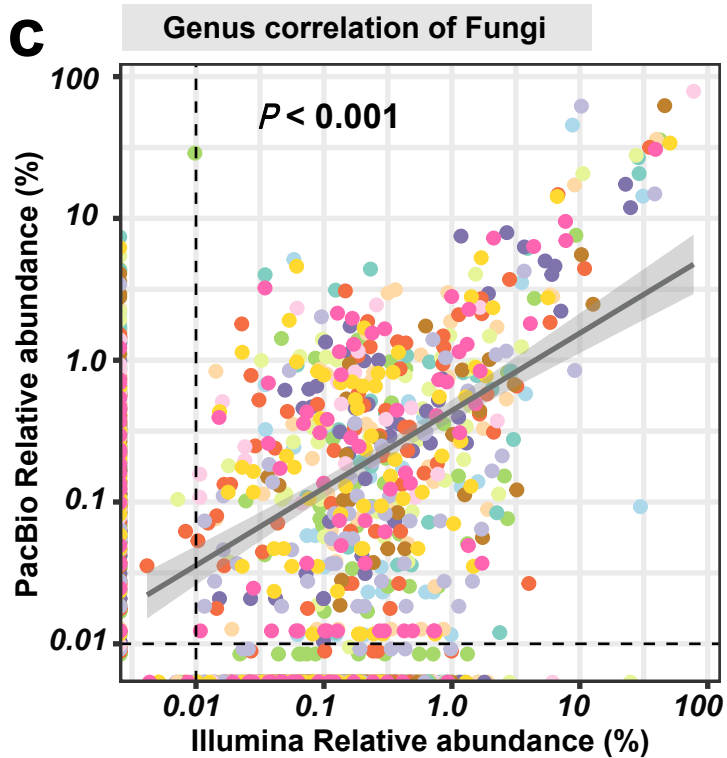
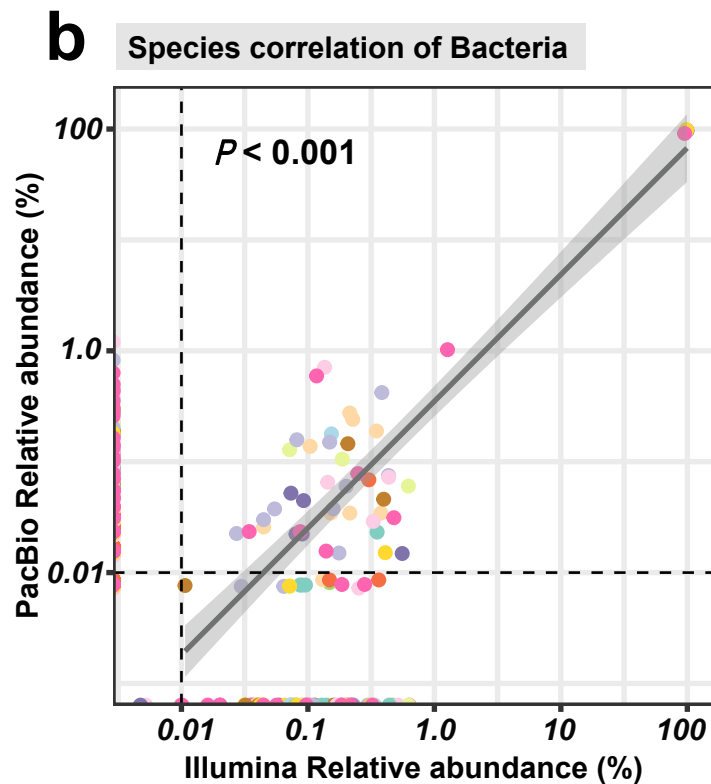
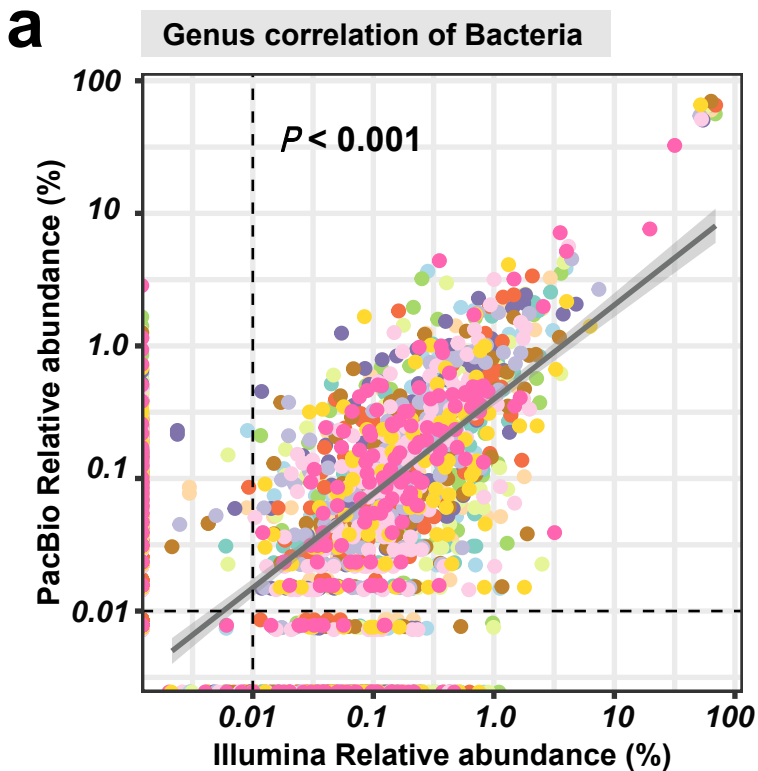


Genus

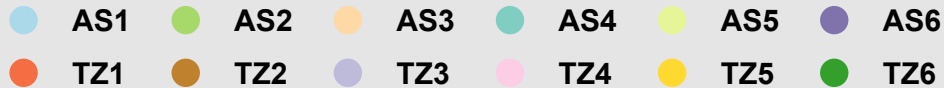


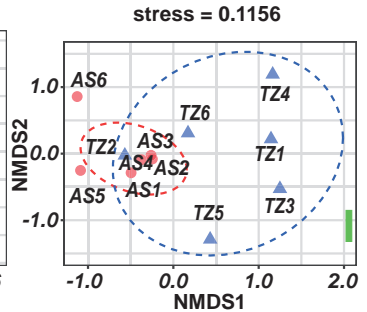
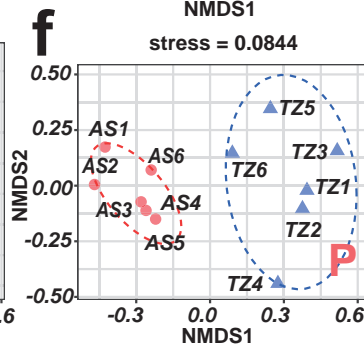
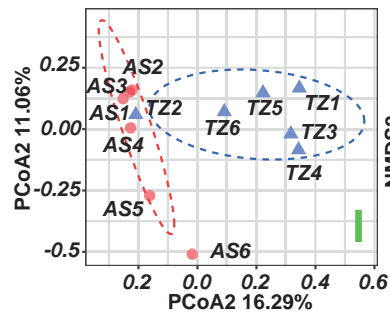
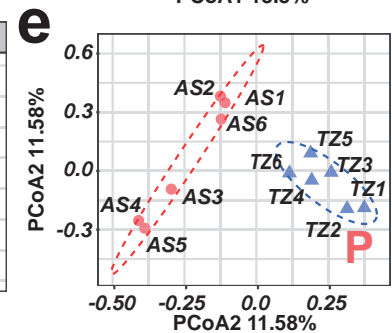
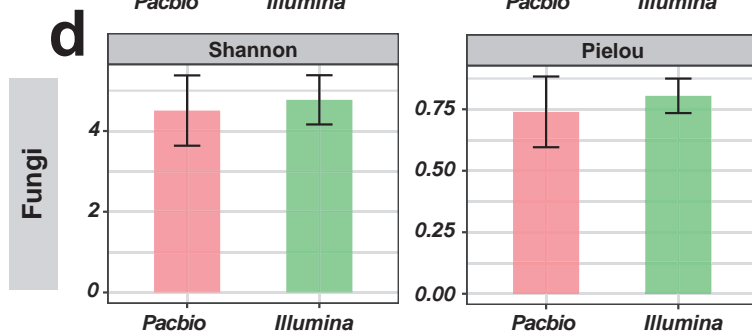
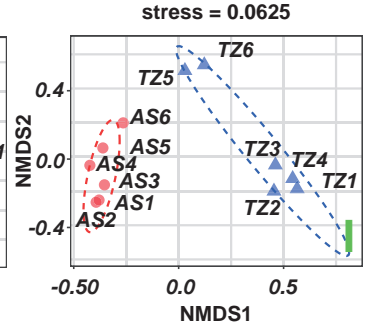
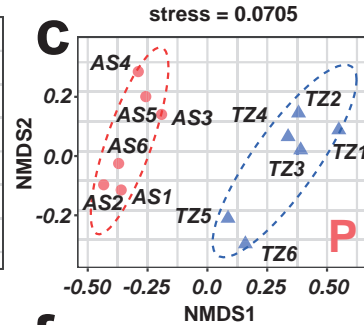
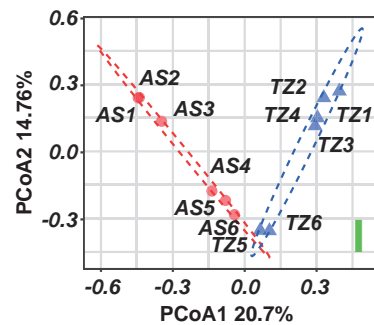
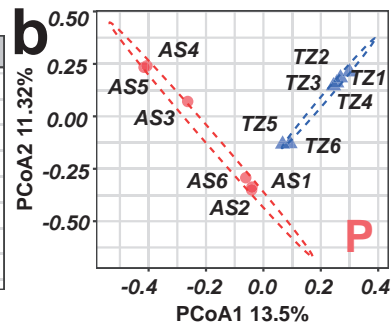
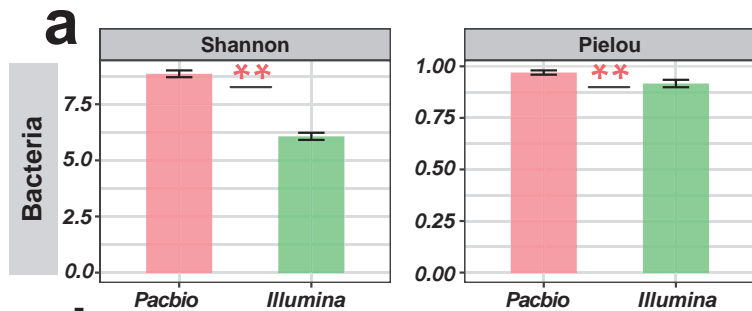
Species

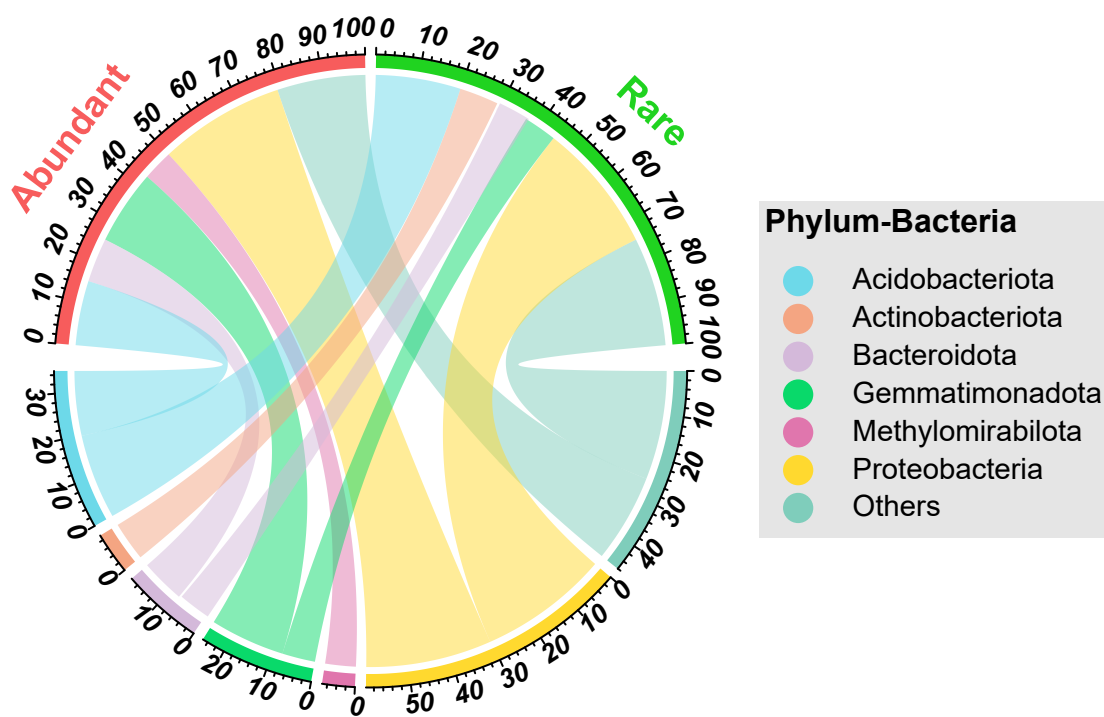




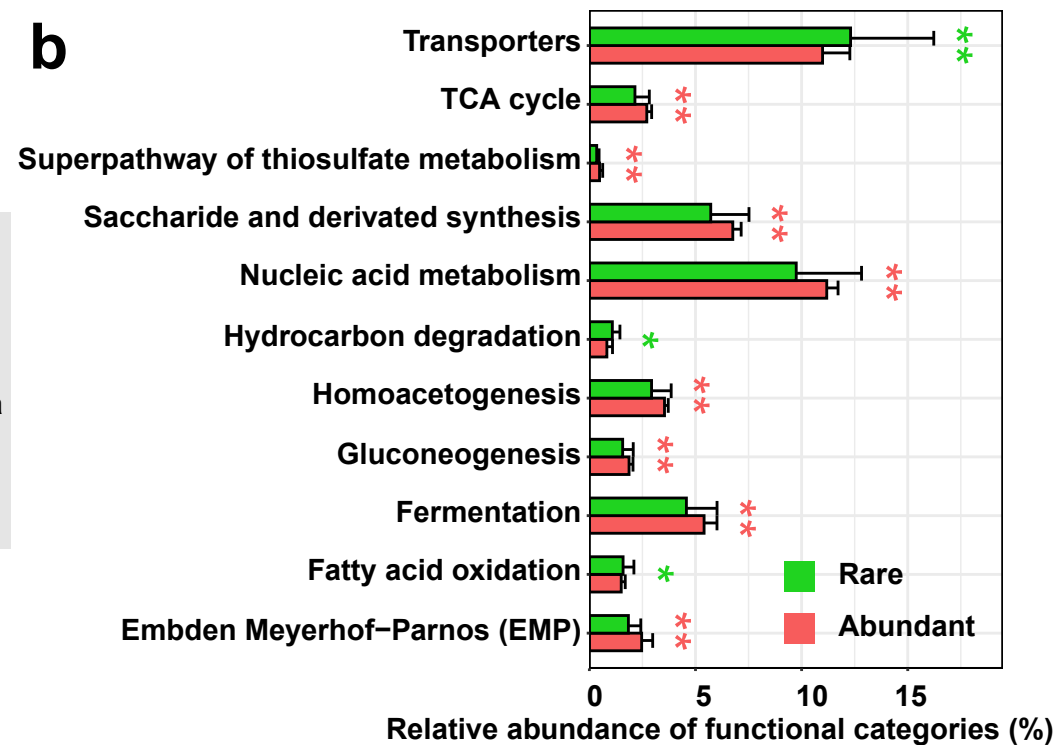
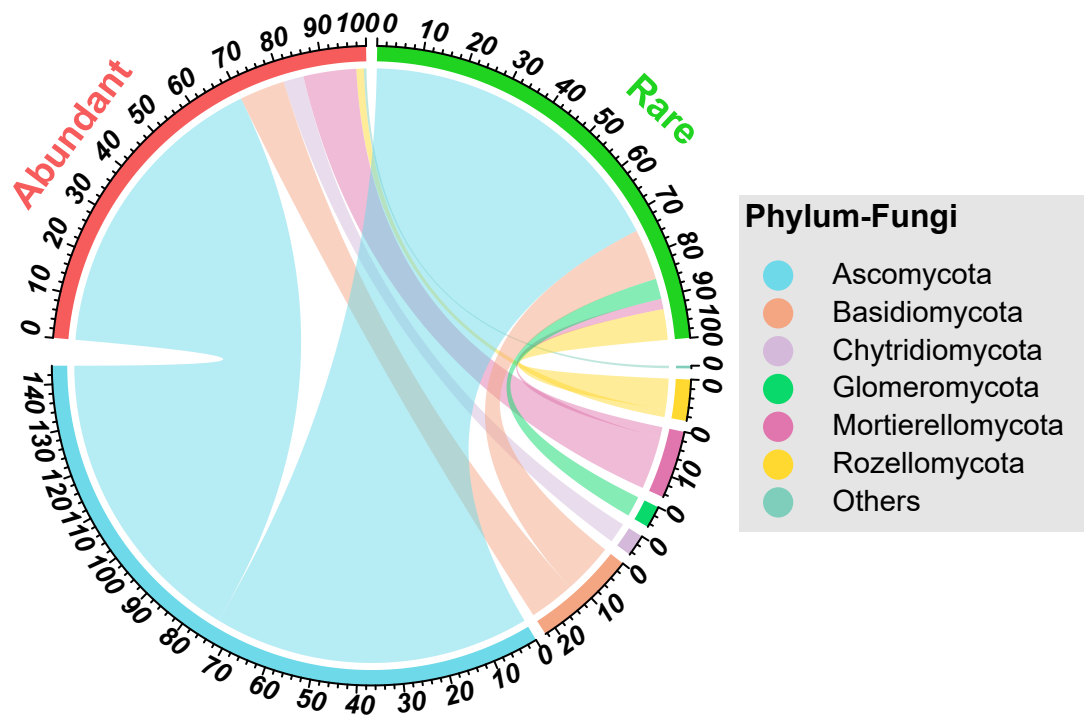
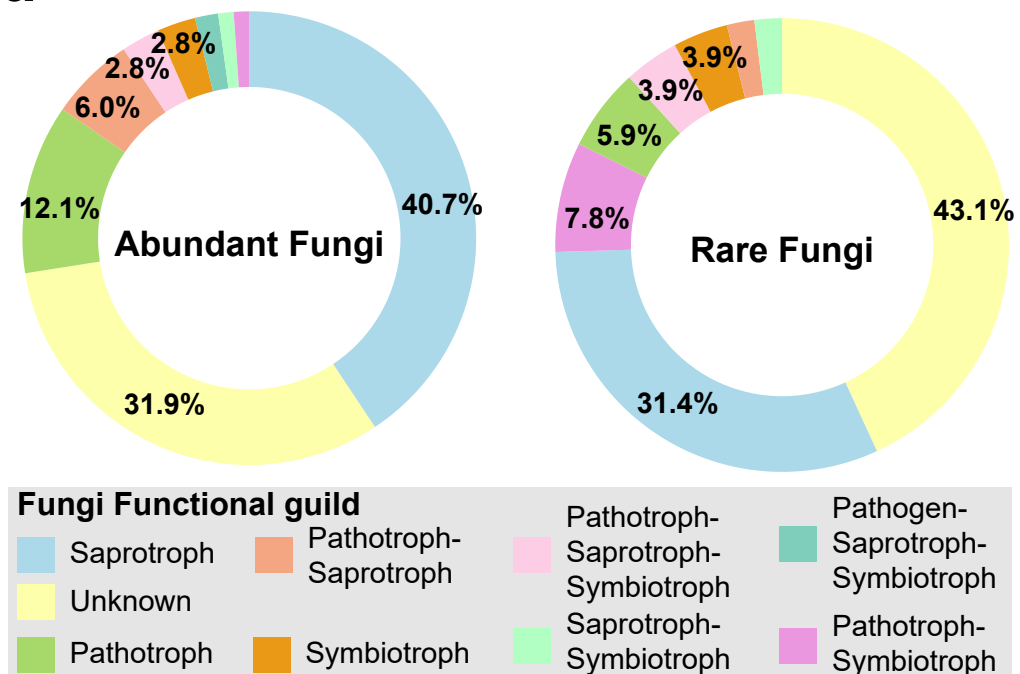
Sample

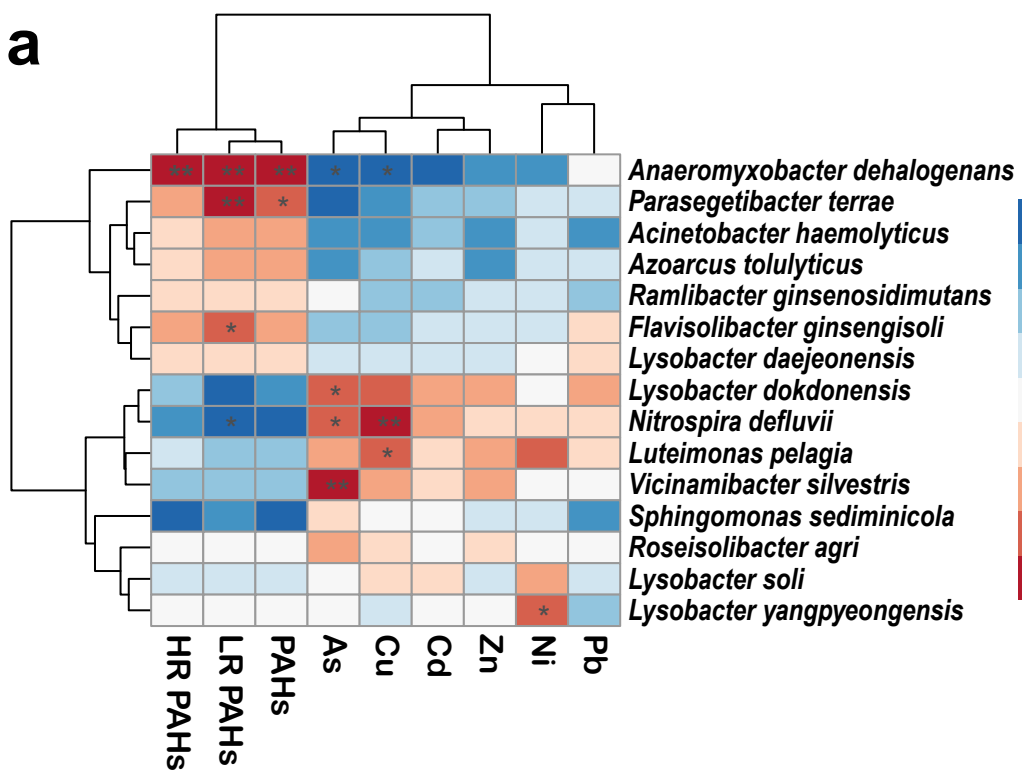
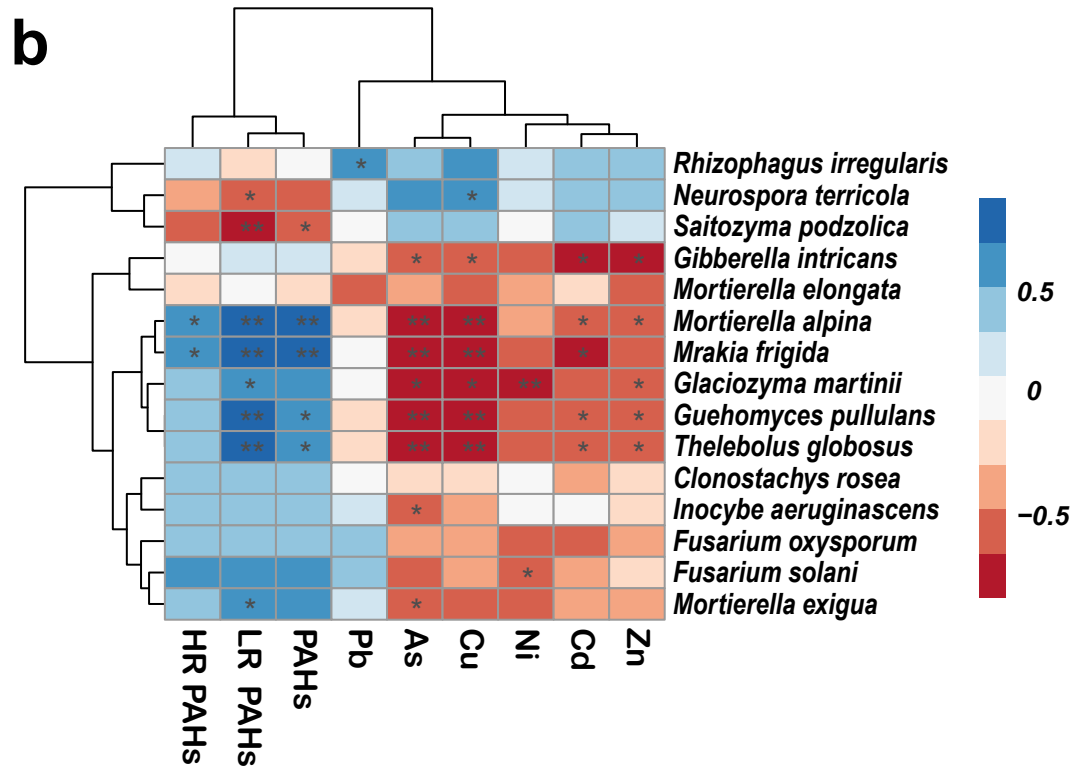


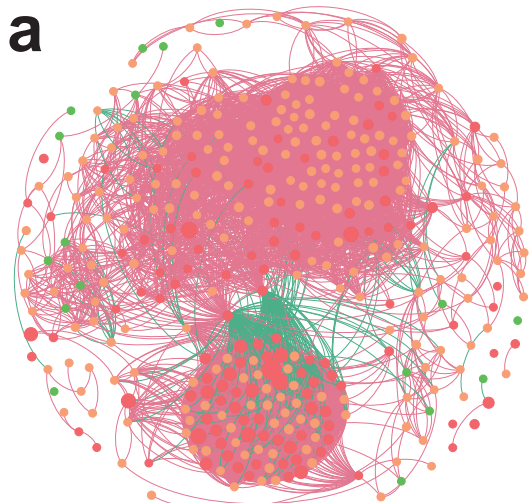


a**Bacteria****b****Superpathway of thiosulfate metabolism****Saccharide and derivated synthesis****Nucleic acid metabolism****Hydrocarbon degradation****Homoacetogenesis****Gluconeogenesis****Fermentation****Fatty acid oxidation****Embden Meyerhof-Parnos (EMP)**

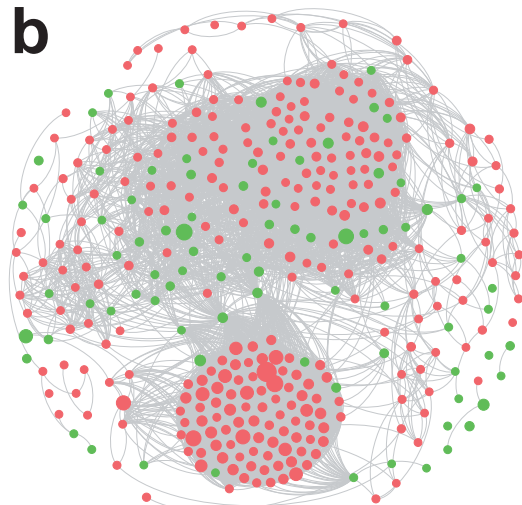
Relative abundance of functional categories (%)

**c****Fungi****d**

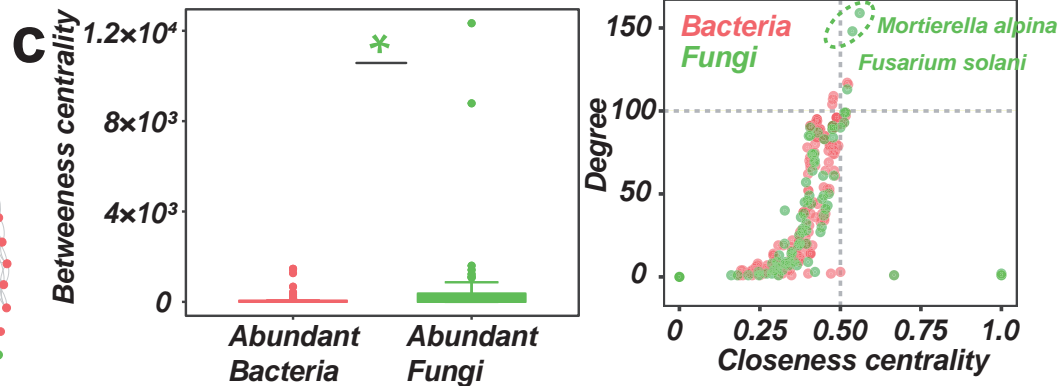
a**b**



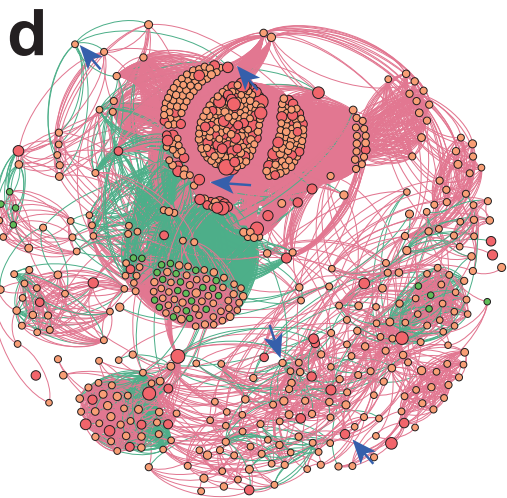
node: 371 Edge :9786
 ● Abundant ● Intermediate ● Rare
 — Positive (96.44%) — negative (3.56%)



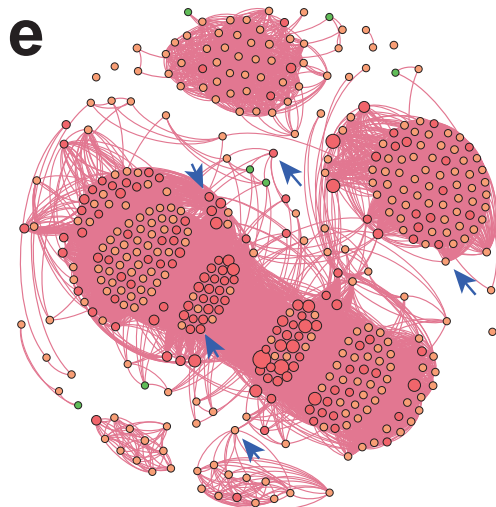
● Bacteria (77.63%) ● Fungi (22.37%)



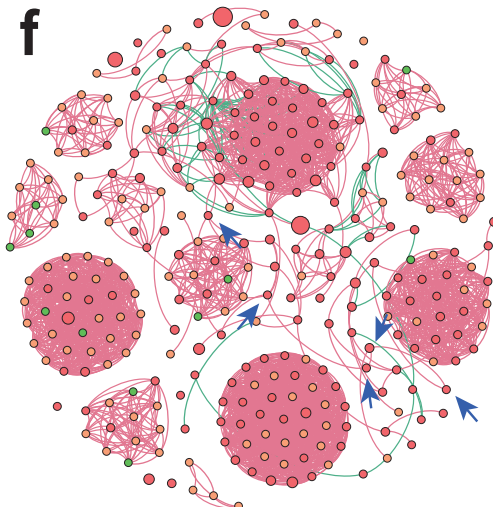
Network	Edges	Avg. Degree	Avg. Weighted Degree	Avg. Clustering Coefficient	Avg. Eccentricity
Original	9786	52.788	46.014	0.735	6.42
Exclude keystone fungi	8462	48.217	42.712	0.749	6.61



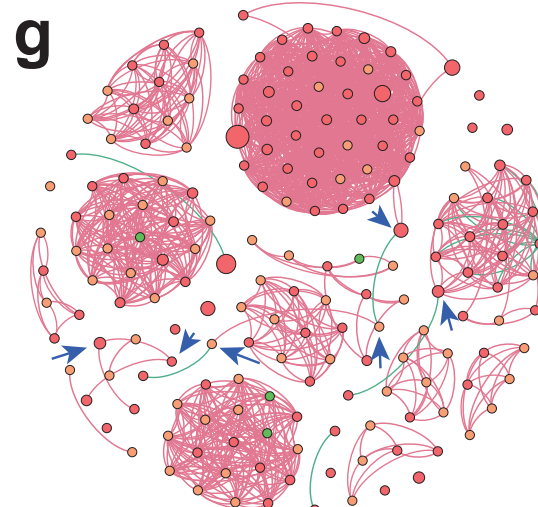
node: 696 Edge :44470
 ● Abundant ● Intermediate ● Rare
 — Positive (95.05%) — negative (4.95%)



node: 434 Edge :14670
 ● Abundant ● Intermediate ● Rare
 — Positive (100%) — negative (0%)



node: 307 Edge :2656
 ● Abundant ● Intermediate ● Rare
 — Positive (97.44%) — negative (2.56%)



node: 183 Edge :1524
 ● Abundant ● Intermediate ● Rare
 — Positive (98.82%) — negative (1.18%)

