1 Mapping-based all-RNA-information sequencing analysis (ARIsseq) pipeline
2 simultaneously revealed taxonomic composition, gene expression, and their correlation
3 in an acidic stream ecosystem

4

5 Arisa Tsuboi^{1, 2}, Misao Itoga¹, Yuichi Hongoh², and Shigeharu Moriya^{1, 3, *}

6

7 Short title: New pipeline for rRNA and mRNA sequence analyses for environmental
8 microbial communities

9

10

11 1) RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan

12 Arisa Tsuboi, Misao Itoga, Shigeharu Moriya

13

14 2) Department of Biological Sciences, Tokyo Institute of Technology, Tokyo, Japan

15 Arisa Tsuboi, Yuichi Hongoh

16

17 3) Graduate School of Medical Life Science, Yokohama City University, Yokohama,

18 Kanagawa, Japan

19 Shigeharu Moriya

20

21 * Corresponding author

22 E-mail: smoriya@riken.jp (SM)

23

24 Abstract

25

26 We developed a new pipeline for simultaneous analyses of both rRNA profile
 27 as a taxonomic marker and mRNA profile as a functional marker, to understand microbial
 28 ecosystems in natural environments. Our pipeline, named All-RNA-Information
 29 sequencing (ARIsSeq), comprises a high-throughput sequencing of reverse transcribed
 30 total RNA and several widely used computational tools, and generates quantitatively
 31 reliable information on both community structures and gene expression patterns, which
 32 were verified by quantitative PCR analyses in this study. Particularly, correlation network
 33 analysis in the pipeline can reveal microbial taxa and expressed genes that share patterns
 34 of dynamics among different time and/or geographical points. The pipeline is primarily
 35 mapping-based, using a public database for small subunit rRNA genes and obtained
 36 contigs as the reference database for protein-coding genes. We applied this pipeline to
 37 biofilm samples, as examples, collected from an acidic spring water stream in the
 38 Chyatsubomi-goke Park in Gunma prefecture, Japan. Our analyses revealed the
 39 predominance of iron and sulfur-oxidizing bacteria and *Pinnularia* diatoms, and also
 40 indicated that the distributions of the iron-sulfur-oxidizing bacterial consortium and the
 41 *Pinnularia* diatoms largely overlapped but showed distinct patterns. In addition, our
 42 analyses showed that the iron-oxidizing bacterial genus *Acidithiobacillus* and
 43 co-occurring *Acidiphilium* shared similar distribution pattern whereas another
 44 iron-oxidizing genus *Leptospirillum* exhibited a distinct pattern. Our pipeline enables
 45 researchers to more easily capture the outline of microbial ecosystems based on the
 46 taxonomic composition, protein-coding gene expression, and their correlations.

47

48

49 Introduction

50

51 Microbial communities play various and crucial roles in their habitat
 52 ecosystems. Because most microbial species in many environments are yet uncultivable,
 53 culture-independent approaches are generally adopted to reveal both taxonomic

54 compositions and potential functions. Sequencing analysis of small subunit (SSU) rRNA
55 gene amplicons is widely used methodology to obtain information on the taxonomic
56 composition of a microbial community. Metagenome and metatranscriptome analyses are
57 also performed to investigate the community structure based on rRNA and/or other
58 taxonomic marker genes as well as to obtain the functional information on the microbial
59 community. In general, researchers choose a combination of these methodologies to
60 comprehensively understand a microbial ecosystem.

61 Among these currently available methodologies, RNA-based analyses are
62 essential to the assessment of microbial activities. The large proportion of rRNA,
63 compared to mRNA, has been an obstacle to functional analyses based on mRNA
64 sequences; however, high-throughput sequencing has enabled researchers to analyze
65 sufficient mRNA sequences without depleting rRNA in many cases. Simultaneous
66 acquisition of both mRNA sequences as functional markers and rRNA sequences as
67 taxonomic markers from the same RNA sample is advantageous to depict the precise
68 picture of a microbial ecosystem [1-4]. Furthermore, amplicon-based analyses of rRNA
69 sequences are subjected to PCR amplification bias [5]. Thus, development of a simple
70 and quantitatively reliable pipeline for the simultaneous analysis of both mRNA and
71 rRNA data will greatly help researchers.

72 Here, we developed a pipeline comprising several steps: a standard quality
73 filtering of sequence reads, mapping reads to a public SSU rRNA sequence database, *de*
74 *novo* assembly of mRNA reads and annotation of the contigs, mapping reads to the
75 mRNA contigs, and finally statistical analyses of the expression level of respective gene
76 categories and the frequency of microbial taxa. Our pipeline, named
77 All-RNA-Information sequencing (ARIsseq), yields quantitatively reliable information
78 for both microbial community structure and gene expression, which was tested by
79 quantitative PCR (qPCR) analyses. The quantitative information is also used for a
80 correlation network analysis, which can reveal microbial taxa and expressed genes that
81 share patterns of dynamics among different time and/or geographical points.

82 We applied this newly developed tool, ARIsseq, to an analysis of biofilm
83 samples collected from an acidic spring water stream at the Chyatsubomi-goke Park in
84 Japan. This acidic stream is a well-known ecosystem for bio-mineralization by
85 iron-oxidizing microbial consortia [6]. ARIsseq successfully provided quantitative

information on both the microbial community structure and gene expression. The results showed that the similarity and dissimilarity of distribution patterns among dominant organisms and expressed genes.

Materials and Methods

Sample collection

Biofilm samples were collected at three different points along an acidic stream at the Chyatsubomi-goke Park in Gunma prefecture, Japan [6]. The stream originated from an acidic spring and then flowed down to the inlet point of the Shirakinu waterfall. The three points were: site 1 upstream point (36°38'58.14"N 138°35'10.66"E), site 2 upstream point (36°38'58.14"N 138°35'11.24"E), and site 3 downstream point (36°38'57.05"N 138°35'27.82"E). A schematic drawing of these sampling sites with water temperature, pH, and ⁵⁷Fe element concentration was shown in Fig 1. The values of pH and water temperature were directly measured by a portable pH meter, HM-30P (TOA DDK, Tokyo, Japan). The total iron concentration in the stream water was determined for ⁵⁷Fe element in the diluted solution by 0.01 mol L⁻¹ hydrochloric acid (088-02265, WAKO, Japan) by an inductive coupled plasma mass spectrometry, NexION300 (PerkinElmer Japan, Yokohama, Japan).

Fig 1. Schematic drawing of sampling site topology and physical and chemical conditions.

Three stones (indicated as a, b, c) per site were sampled. Each stone was rinsed with purified water, and biofilm was brushed off from the stone surface using a sterile toothbrush and purified water. Suspension of the collected biofilm was filtered with a 0.22-μm polyvinylidene difluoride (PVDF) Durapore® membrane (Millipore, Billerica, MA, USA). The filter membrane with trapped materials was preserved in 1 ml of RNeasy lysis solution (Ambion, Austin, TX, USA) at -80°C until being processed. In total, nine samples were collected (designated as samples 1a–c, 2a–c, and 3a–c).

117

118 **RNA extraction and sequencing**

119 The materials trapped on the filter membrane was resuspended in RNAlater®
120 by pipetting, and precipitated by centrifugation at 15,000 rpm for 5 min. The pellet was
121 subjected to RNA extraction, using the PowerBiofilm RNA Isolation Kit (MO Bio,
122 Carlsbad, CA, USA), according to the manufacturer's instructions. RNA was eluted with
123 100 µl of purified water. RNA concentration was measured using a Qubit system
124 (Invitrogen, Carlsbad, CA, USA) and adjusted to 100 ng/µl with purified water. RNA was
125 not sufficiently recovered from samples 1b, 1c, and 3c, which were removed from further
126 experiments.

127 Sequencing libraries were prepared using the SMARTer® Stranded RNA-Seq
128 Kit for Illumina (Takara, Kyoto, Japan), following the manufacturer's instructions. The
129 concentration and length of DNA fragments in the sequence library were measured using
130 the Qubit system and a Bioanalyzer 2100 (Agilent Technologies, Carlsbad, CA, USA).
131 When libraries contained DNA fragments < 80 bp, the libraries were further processed
132 with the Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA) to remove small
133 fragments, according to the manufacturer's instructions. Sequencing was performed on
134 the Illumina MiSeq platform (Illumina, CA, USA) with the Reagent Kit v3 (600 cycles,
135 paired-end mode). Sequence data have been deposited at DDBJ with the accession
136 number DRA005571.

137

138 **Processing sequence data**

139 The scheme of sequence processing is outlined in Fig. 2. Sequence reads were
140 trimmed and quality-filtered using program Trimmomatic [7] in paired-end mode with a
141 seed mismatch value of 5, a palindrome clip threshold of 30, a simple clip threshold of 7,
142 a minimum read length of 100 bp and a headcrop of 6 bp. Trimmed reads were used for
143 mapping to the SSU rRNA sequence database SILVA release 108 for QIIME [8,9].
144 Mapping was performed using Bowtie2 [10] with a local alignment mode and single- or
145 paired-end modes. The data generated by Bowtie2 were converted to the sorted binary
146 sequence alignment/map (BAM) format, using samtools [11], and the number of mapped
147 reads were counted using the eXpress program package [12]. The read count data were

integrated into taxon information and utilized for secondary analyses (see below).
 Reads unmapped to the SILVA SSU rRNA database by paired-end mode were assembled using the Trinity program package [13] with the Jaccard clip option, to construct contigs. Open reading frames (ORFs) and the encoded protein sequences were predicted using the TransDecoder program package (<https://transdecoder.github.io/>). The ORF data were used as the reference database for mapping reads. Functional annotation of the identified ORFs was conducted with the Trinotate program package (<https://trinotate.github.io/>) that uses a combination of a BlastP search against the “UniProt/Swiss-Prot database for Trinotate”, an hmmer search in the Pfam database, and RNAMMER analysis. BlastP searches against the non-redundant (nr) and standard Swiss-Prot/UniProtKB databases were conducted, and the results were added to the annotation. Furthermore, a Ghost KOALA search provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14] was performed, and a K number was assigned to each of the ORFs. These functional annotations were combined with the read count data, and rank abundance curves were constructed.

Secondary analysis

The read count data for both SSU rRNA and ORFs were subjected to secondary analyses. First, we performed normalization of the read count data, following a negative binomial distribution with generalized linear model, using the DESeq2 package [15]. For this analysis, the reads were divided into two groups: the “upstream group” (samples 1a and 2a–c) and the “downstream group” (samples 3a, b). SSU rRNA reads or ORFs with less than 10 mapped reads in total from all samples were excluded. The calculation with the DESeq2 package was performed using the TCC package in *R* [15].

Network analysis was performed based on Spearman’s rank correlation coefficient matrix, calculated using *R*, for the normalized read count datasets. The results were visualized using the Gephi program package [16] with a transformed matrix of connection between source and target with collected high positive correlation coefficient ($r > 0.7$) from correlation coefficient matrix [17]. The raw read count datasets were used also for calculating rarefaction curves using Past3.14 [18].

179 Quantitative PCR experiments

180 We performed qPCR to quantify the expression level of genes and to compare
181 the results with those obtained from our pipeline. We selected two SSU rRNA sequences
182 and two contigs each containing an ORF (ORF contig) as examples that showed
183 differential expression patterns. We chose taxon 17056 (18S rRNA of *Pinnularia* cf.
184 *gibba*) and the ORF contig TRINITY_DN8038_c26_g13_i2|m.9499 (peptide 9499), as
185 they were highly expressed in the “upstream group”, and taxon 50111 (18S rRNA of
186 *Chironomus tentans*) and the ORF contig TRINITY_DN6923_c3_g1_i3|m.6571 (peptide
187 6571), as they were highly expressed in the “downstream group”. Peptide 9499 was
188 annotated as an “uncharacterized protein” by both the Trinotate pipeline and a BlastP
189 search against the Swiss-Prot/UniProtKB database. Peptide 6571 was predicted to be
190 3-dehydroquinate dehydratase/shikimate dehydrogenase (K13832) by Ghost KOALA or
191 putative alpha-L1 nicotinic acetyl choline receptor by a BlastP search against the
192 Swiss-Prot/UniProtKB database. Primer sequences for qPCR are shown in Table. 1.

193 Sequencing libraries were used as templates for qPCR, although one of the
194 downstream samples (sample 3b) was excluded because the library DNA ran out by
195 sequencing. Four samples (1a and 2a–c) from the upstream group and one (3a) from the
196 downstream group were subjected to qPCR using the KAPA™ SYBR® Fast qPCR Kit
197 (KAPA Biosystems, MA USA), according to the manufacturer’s instructions. The qPCR
198 was performed in a Thermal Cycler Dice® Real Time System TP850 (Takara) in 25-μl
199 reaction volume, and the PCR program was as follows: initial denaturation at 95°C for 30
200 sec and 40 cycles of 95°C for 30 sec and 60°C for 1 min. The dilution rate of the samples
201 was 1/200. The calibration curve was constructed using sample 2c with dilution rates
202 1/100, 1/200, 1/400, 1/800, 1/1600, and 1/3200. The experiments were conducted in
203 triplicate. The amount of template DNA was adjusted, using the KAPA™ Library
204 Quantification Kit for Illumina (KAPA Biosystems, MA USA).

205

206

207 Results and Discussion

208

209 Mapping-based analysis of total RNA sequences

210 Total RNA sequencing of the six biofilm samples (1a, 2a–c, and 3b,c)
 211 generated 770,334 to 1,320,217 read pairs, and >80% reads passed the quality check
 212 (Table 2). These reads were analyzed using our mapping-based ARIsq pipeline (Fig 2).
 213 Of the reads, 45.7% were mapped to the SILVA SSU rRNA sequence database, using the
 214 single-end mode option of Bowtie2 (Table 2). When using the paired-end mode, mapped
 215 reads were only 5.1%; thus, we used the results obtained using the single-end mode for
 216 quantification data. We here only considered quantification of SSU rRNA genes sharing a
 217 high sequence identity with those in the reference database, Therefore, the unmapped
 218 reads should contain other SSU rRNA, large subunit rRNA, and non-coding RNA, which
 219 can be identified and removed in the following annotation step.

220

221

222 **Fig 2. Schematic flow chart of the ARIsq pipeline.** Green arrows indicate the flow of
 223 rRNA gene data, and red arrows indicate the flow of the other RNA.

224

225 Unmapped reads against the SILVA database by paired-end mode mapping
 226 were used for a *de novo* assembly process using Trinity, and then ORFs on the contigs
 227 were predicted by using TransDecoder. These ORFs were used as the reference sequence
 228 database for mapping analysis to obtain gene expression profiles. In the single-end mode
 229 of Bowtie2, 17.2% of the reads were mapped onto this self-made database, whereas only
 230 2.7% of the reads were mapped in the paired-end mode, (Table 2). Here, we again
 231 employed the results from the single-end mode for subsequent analyses.

232 Read count data for both SSU rRNA and ORFs were obtained using the
 233 eXpress program package. Fig 3 shows rarefaction curves, which indicated that the
 234 sequence efforts (*i.e.*, number of reads) for both rRNA and ORFs were sufficient or nearly
 235 sufficient to obtain most variations in the RNA samples except for the SSU rRNA of the
 236 downstream samples 3a and 3b. All read count datasets were subjected to the
 237 normalization processes with DESeq2 and annotation using Trinotate, BlastP, and Ghost
 238 KOALA (see Materials and Methods for details). The annotation and normalization
 239 results were listed in S1 and S2 Tables.

240

241 **Fig 3. Rarefaction curves.** (a) Rarefaction curves of taxa based on SSU rRNA sequences.
 242 Taxa was defined as reads mapped OTUs. (b) Rarefaction curves of protein-coding genes
 243 (ORF) based on cDNA sequences from mRNA. Sample names are shown aside the
 244 curves.

245

246 In the normalized read count data, gene expressions of several rRNA or ORFs
 247 greatly differed between the upstream and the downstream sample groups. For example,
 248 mapping data suggested that, a diatom, *Pinnularia cf. gibba* (taxon 17056), was abundant
 249 in the upstream samples but less represented in the downstream sample 3a based on the
 250 rRNA read count data (Fig 4a). In contrast, aquatic larvae of the non-biting midge
 251 *Chironomus tentans* (taxon 5111) were found abundant only in the downstream sample 3a
 252 (Fig 4a). Although insects are not microbes, we here included them because they
 253 probably have a great impact on the biofilm ecosystem. These expression patterns were
 254 well congruent with the results obtained by qPCR analyses (Fig 4b). ORF read count data
 255 showed that the ORF for peptide 9499 was most highly expressed in the upstream
 256 samples, whereas ORFs for peptide 6571 was most highly expressed in the downstream
 257 samples (Fig 5a). These results were also congruent with those from qPCR analyses (Fig
 258 5b). Thus, our ARIsseq analysis pipeline generated quantitatively reliable results for both
 259 rRNA and expressed ORFs.

260

261 **Fig 4. Quantitative analysis of rRNA data.** (a) Normalized read counts for an
 262 upstream-specific taxon, 17056, and a downstream-specific taxon, 5111. (b) Relative
 263 abundance of these taxa evaluated by qPCR.

264

265 **Fig 5. Quantitative analysis of ORF data.** (a) Normalized read counts of an
 266 upstream-specific ORF for peptide 9499 and a downstream-specific ORF, 6571. (b)
 267 Relative abundance of these ORFs evaluated by qPCR.

268

269 **Biofilm community structure in the acidic stream**

270 Fig 6 shows SSU rRNA-based taxonomic compositions at the domain level.

271 Eukaryotes or bacteria predominated in all samples; only a few archaeal sequences were
272 detected. The distribution pattern of each taxon was summarized by a correlation network
273 analysis (Fig 7a). Six clusters were recognized, and each distribution pattern is shown in
274 Fig 7b. Here, a “cluster” is defined as a group of taxa that exhibit similar distribution
275 pattern among the samples. Members of cluster 1 mainly inhabited the downstream sites
276 (samples 3a,b), whereas members of clusters 3, 4, and 5 were found mostly at upstream
277 sites (samples 1a and 2a–c). Because the reads assigned to clusters 2, 5, and 6 were
278 relatively few, we focused only on clusters 1, 3, and 4 for comparisons.

279

280 **Fig 6. Taxonomic compositions at domain level.** Normalized read counts are used for
281 the calculation.

282

283 **Fig 7. Correlation network analysis of SSU rRNA data.** (a) A network drawn for
284 relationships with positive correlation indexes > 0.7 . (b) Normalized read counts of SSU
285 rRNA for taxa assigned to each cluster.

286

287 Fig 8 shows the taxonomic composition of each cluster at the genus level.
288 *Chironomus* midge larvae, *Pinnularia* diatoms, and *Acidithiobacillus* bacteria
289 predominated in clusters 1, 3, and 4, respectively. This at once indicated that *Chironomus*
290 predominated in the downstream regions, and that *Pinnularia* and *Acidithiobacillus* did in
291 the upstream regions. *Acidithiobacillus* is a well-known acidophilic bacterial genus that
292 oxidizes sulfur and iron, and is frequently found in acid mine drainage [19–22]. In this
293 acidic stream, the presence of bacteria attached to iron precipitate with
294 phosphorous/sulfur crystals was previously reported [6]. It is highly possible that
295 *Acidithiobacillus* members with co-occurring, possibly symbiotic *Acidiphilium* [23], the
296 second-dominant genus in cluster 4, mainly cause iron/sulfur oxidation.

297

298 **Fig 8. Taxonomic compositions of each cluster generated by network analysis of**
299 **SSU rRNA data.** Normalized read counts are shown at the genus level. The taxonomy
300 was based on the SILVA database release 108. Original data are shown in S3 Table.

301

302 *Pinnularia* diatoms are also frequently found in acid mine drainage [19,24] and

comprise a biofilm community [25]. The presence of *Pinnularia*-like diatoms in this acidic stream was previously reported: the cells were found around or attached to precipitated Fe-P materials [6]. As shown in Fig 8, overlapping but distinct distribution pattern is observed between *Pinnularia* and *Acidithiobacillus* in the upstream region. This pattern suggested that the primary production at the upstream site 2a is largely attributable to the iron-sulfur oxidation by *Acidithiobacillus*, while photosynthesis by *Pinnularia* contributes to the primary production broadly in the upstream region

In the downstream samples, the larvae of the non-biting midge genera *Chironomus* and *Acricotopus* were predominant (Fig 8). These chironomid larvae are detritivores and frequently found in biofilms in freshwater [26]. The third-dominant genus in the downstream samples in cluster 1 was *Leptospirillum*. *Leptospirillum* members are also iron-oxidizers in general and produce iron/sulfur granules inside their extracellular polymeric substances that compose the biofilm [19,27,28]. This different distribution pattern between *Acidithiobacillus* and *Leptospirillum* as revealed in our correlation analysis implies their niche differentiation.

Gene expression profiles and correlation with taxonomic compositions

ORFs showing a similar expression pattern among the samples were also clustered by a correlation network analysis (Fig 9a). Two large clusters were generated: cluster 1 comprised ORFs that were highly expressed in the downstream samples; cluster 2 comprised those highly expressed in the upstream samples (Fig 9b). In cluster 2, genes related to photosynthesis, such as electron transportation and photosystem components were dominant (Table 3). This coincided with the predominance of diatoms including *Pinnularia* in the upstream region (Fig 8).

Fig 9. Correlation network analysis of expressed ORF data. ORFs were categorized and bundled according to the KEGG orthology. Normalized read counts were used for the analysis. (a) A network drawn for relationships with positive correlation indexes >0.7. (b) Normalized read counts of expressed ORFs assigned to each cluster.

Furthermore, we constructed a correlation network based on expression

patterns of both rRNA and ORF datasets (Fig 10). Here, “cluster” comprised both SSU rRNA and ORFs that exhibited similar distribution patterns among the sampling sites. The clustering pattern resembled that based on SSU rRNA (Fig 8 and 11); the ORF expression patterns and the SSU rRNA distribution patterns were well congruent. Dominant gene categories in each cluster were mostly house-keeping genes (Tables 4 and 5) or photosynthesis-related genes (Table 6). Rank abundance curves of taxa that expressed genes related to photosynthesis and assigned to cluster 6 are shown in Fig 12, which suggested that photosynthesis-related genes were mainly originated from diatoms. In Fig 12, closest species that were identified by BlastP searches against the SwissProt/UniProtKB database are shown. *Phaeodactylum tricornutum*, listed as the predominant species in Fig 12, is a marine diatom and one of the few diatoms with genome sequence being analyzed [29]; thus, this most likely represented the predominant diatom *Pinnularia* in the upstream region of this acidic stream.

Fig 10. Correlation network analysis of combined data of SSU rRNA and expressed ORFs. (a) A network was drawn for relationships with positive correlation indexes >0.7. (b) Normalized read counts of SSU rRNA-based taxa and expressed ORFs assigned to each cluster.

Fig 11. Taxonomic compositions of each cluster generated by network analysis of combined data of SSU rRNA and expressed ORFs. The taxonomic compositions were based only on the SSU rRNA data. Original data are shown in S4 Table. See also the legend to Fig 8.

Fig 12. Rank abundance curves of taxa expressing genes involved in photosynthesis. The species names are the closest taxa found by BlastP searches for photosynthesis-related genes against the SwissProt/UniprotKB database. Genes assigned to cluster 6 in Fig 10 were used for the analysis. (a) Rank abundance curves of organisms that expressed genes related to electron transportation (K00330, K00339, K00343, K00412, K02256, K02261, K02262, K03881, K03883, K03934, and K03935). (b) Rank abundance curves of organisms that expressed genes related to photosystem (K02689, K02690, K02703, K02704, L02705, K02706, and K08910). (c) Rank abundance curves

366 of organisms that expressed genes for the RuBisCO components (K01601 and K01602).

367 Asterisks indicate diatoms.

368

369 **Conclusions**

370 Our All-RNA-Information Sequencing analysis (ARIsseq) successfully revealed both
371 microbial community structures and expressed gene categories in quantitatively reliable
372 forms. Particularly, our ARIsseq pipeline was able to show the correlation among the
373 community members based on rRNA frequency and also the correlation among the gene
374 categories in the biofilm samples collected from the acidic spring water stream. Our
375 pipeline is suitable for researches to capture comprehensive information from one RNA
376 sequencing analysis and facilitates understanding of ecological functions of organismal
377 communities in natural environments.

378

379

380 **Acknowledgements**

381 Sampling was carried out in an environmental research permitted by the board of
382 education, Gunma Prefecture, Japan.

383

384

385 **References**

- 386 1. Urich T, Lanzén A, Stokke R, Pedersen RB, Bayer C, Thorseth IH, et al. Microbial
387 community structure and functioning in marine sediments associated with diffuse
388 hydrothermal venting assessed by integrated meta-omics. *Environ Microbiol.*
389 2014;16: 2699–2710. doi:10.1111/1462-2920.12283
- 390 2. Berry D, Schwab C, Milinovich G, Reichert J, Ben Mahfoudh K, Decker T, et al.
391 Phylotype-level 16S rRNA analysis reveals new bacterial indicators of health state
392 in acute murine colitis. *ISME J.* 2012;6: 2091–2106. doi:10.1038/ismej.2012.39
- 393 3. Radax R, Rattei T, Lanzén A, Bayer C, Rapp HT, Urich T, et al.
394 Metatranscriptomics of the marine sponge *Geodia barretti*: tackling phylogeny

- 395 and function of its microbial community. *Environ Microbiol.* 2012;14: 1308–1324.
396 doi:10.1111/j.1462-2920.2012.02714.x
- 397 4. Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. Simultaneous
398 Assessment of soil microbial community structure and function through analysis
399 of the meta-transcriptome. *PLoS ONE.* 2008;3: e2527.
400 doi:10.1371/journal.pone.0002527.s019
- 401 5. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced
402 sequence artifacts and bias: insights from comparison of two 16S rRNA clone
403 libraries constructed from the same sample. *Appl Environ Microbiol.* 2005;71:
404 8966–8969. doi:10.1128/AEM.71.12.8966-8969.2005
- 405 6. Akai J, Akai K, Ito M, Nakano S, Maki Y. Biologically induced iron ore at Gunma
406 iron mine, Japan. *Am Min.* 1999;84: 171-182
- 407 7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
408 sequence data. *Bioinformatics.* 2014;30: 2114–2120.
409 doi:10.1093/bioinformatics/btu170
- 410 8. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA
411 ribosomal RNA gene database project: improved data processing and web-based
412 tools. *Nucleic Acids Res.* 2013;41: D590–D596. doi:10.1093/nar/gks1219
- 413 9. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK,
414 et al. QIIME allows analysis of high-throughput community sequencing data. *Nat*
415 *Methods.* 2010;7: 335–336. doi:10.1038/nmeth.f.303
- 416 10. Ben Langmead, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat*
417 *Methods.* 2012;9: 357–359. doi:10.1038/nmeth.1923
- 418 11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
419 alignment/map format and SAMtools. *Bioinformatics.* 2009;25: 2078–2079.
420 doi:10.1093/bioinformatics/btp352

- 421 12. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of
422 sequencing experiments. *Nat Methods*. 2013;10: 71–73. doi:10.1038/nmeth.2251
- 423 13. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.
424 Full-length transcriptome assembly from RNA-Seq data without a reference
425 genome. *Nat Biotechnol*. 2011;29: 644–652. doi:10.1038/nbt.1883
- 426 14. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG
427 Tools for functional characterization of genome and metagenome sequences. *J*
428 *Mol Biol*. 2016;428: 726–731. doi:10.1016/j.jmb.2015.11.006
- 429 15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion
430 for RNA-seq data with DESeq2. *Genome Biol*. 2014;15: 550.
431 doi:10.1186/s13059-014-0550-8
- 432 16. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring
433 and manipulating networks. *Proceedings of the Third International ICWSM*
434 *Conference*. 2009: 361–362.
- 435 17. Ito K, Sakata K, Date Y, Kikuchi J. Integrated analysis of seaweed components
436 during seasonal fluctuation by data mining across heterogeneous chemical
437 measurements with network visualization. *Anal Chem*. 2014;86: 1098–1105.
438 doi:10.1021/ac402869b
- 439 18. Hammer Ø, Harper D, Ryan PD. PAST: Paleontological Statistics Software
440 Package for Education and Data Analysis. *Palaeontol Electronica*. 2001;4: 1–9.
- 441 19. MÃ ndez-GarcÃ a C, PelÃ ez AI, Mesa V, SÃ nchez J, Golyshina OV, Ferrer M.
442 Microbial diversity and metabolic networks in acid mine drainage habitats. *Front*
443 *Microbiol*. 2015;6: 687. doi:10.1002/mbo3.17
- 444 20. Hua Z-S, Han Y-J, Chen L-X, Liu J, Hu M, Li S-J, et al. Ecological roles of
445 dominant and rare prokaryotes in acid mine drainage revealed by metagenomics
446 and metatranscriptomics. *ISME J*. 2014;9: 1280–1294.
447 doi:10.1038/ismej.2014.212

- 448 21. Bonnefoy V, Holmes DS. Genomic insights into microbial iron oxidation and iron
449 uptake strategies in extremely acidic environments. *Environ Microbiol.* 2011;14:
450 1597–1611. doi:10.1111/j.1462-2920.2011.02626.x
- 451 22. Hedrich S, Schlomann M, Johnson DB. The iron-oxidizing proteobacteria.
452 *Microbiology.* 2011;157: 1551–1564. doi:10.1099/mic.0.045344-0
- 453 23. Liu H, Yin H, Dai Y, Dai Z, Liu Y, Li Q, et al. The co-culture of *Acidithiobacillus*
454 *ferrooxidans* and *Acidiphilium acidophilum* enhances the growth, iron oxidation,
455 and CO₂ fixation. *Arch Microbiol.* 2011;193: 857–866.
456 doi:10.1007/s00203-011-0723-8
- 457 24. DeNicola DM. A review of diatoms found in highly acidic environments.
458 *Hydrobiologia.* 2000;433: 111–122.
- 459 25. Aguilera A, Souza-Egipsy V, Gómez F, Amils R. Development and structure of
460 eukaryotic biofilms in an extreme acidic environment, Río Tinto (SW, Spain).
461 *Microb Ecol.* 2007;53: 294–305. doi:10.1007/s00248-006-9092-2
- 462 26. K Johnson R, Boström B, van de Bund W. Interactions between *Chironomus*
463 *plumosus* (L.) and the microbial community in surficial sediments of a shallow,
464 eutrophic lake. *Limnol Oceanogr.* 1989;34: 992–1003.
465 doi:10.4319/lo.1989.34.6.0992
- 466 27. Sand W, Gehrke T. Extracellular polymeric substances mediate
467 bioleaching/biocorrosion via interfacial processes involving iron(III) ions and
468 acidophilic bacteria. *Res Microbiol.* 2006;157: 49–56.
469 doi:10.1016/j.resmic.2005.07.012
- 470 28. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al.
471 Community structure and metabolism through reconstruction of microbial
472 genomes from the environment. *Nature.* 2004;428: 37–43.
473 doi:10.1038/nature02340

474 29. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K. The *Phaeodactylum*
475 genome reveals the evolutionary history of diatom genomes. *Nature*. 2008;456:
476 239–244. doi:10.1038/nature07410

477

478

479 **Contributions**

480 SM, AT, and YH designed the research. AT and MI performed the sampling. AT and SM
481 performed the experiments and analyses. AT, SM, and YH wrote the paper.

482

483

484 **Competing interests**

485 The authors declare no competing financial interests.

486

487

488 **Corresponding author**

489 Shigeharu Moriya

490

491

492 **Supporting information**

493 **S1_Table. Annotation of SSU rRNA and data matrix.**

494 **S2_Table. Annotation of expressed ORFs and data matrix.**

495 **S3_Table. Taxonomic composition of each cluster based on SSU rRNA.**

496 **S4_Table. Taxonomic composition of each cluster based on SSU rRNA and**
497 **expressed ORFs.**

Fig. 1

Fig. 2

Fig. 3

Fig. 4

a

b

Fig. 5

a

b

Fig. 6

Fig. 7

Fig. 8

Fig. 10

Fig. 11

Fig. 12