Supplementary information

Stochastic process determines the spatial variations in microbial community inhabiting terrestrial mud volcanoes across the Eurasian continent

Tzu-Hsuan Tu[1,2,3], Li-Ling Chen[2], Yi-Ping Chiu[3], Li-Hung Lin[3,4], Li-Wei Wu[5], Francesco Italiano[6], J. Bruce H. Shyu[3], Seyed Naser Raisossadat[7,8], and Pei-Ling Wang[2,4*]

1. Department of Oceanography, National Sun Yat-sen University, Kaohsiung, Taiwan
2. Institute of Oceanography, National Taiwan University, Taipei, Taiwan
3. Department of Geosciences, National Taiwan University, Taipei, Taiwan
4. Research Center for Future Earth, National Taiwan University, Taipei, Taiwan
5. Department of Life Science, Tunghai University, Taichung, Taiwan
6. National Institute of Geophysics and Volcanology, Palermo, Italy
7. Department of Geology, University of Birjand, Birjand, Iran
8. Earth Science Research Group, University of Birjand, Birjand, Iran

Material and methods

Sampling sites and procedures

Muddy fluids from bubbling pools and a total of 16 sediment cores from the adjacent mud platform were retrieved from 15 MVs across the Eurasian continent during 2011 to 2017 (Fig. 1; plotted using the ggmap package (Kahle & Wickham, 2013) in R; Table S1) for geochemical and molecular analyses. In brief, bubbling fluids and cores were collected using sterilized cups and liners, respectively. The lengths of the cores ranged between 20 and 160 cm. Samples were transported to the nearby laboratory or accommodation within 5 hours after retrieval. The cores were immediately sectioned at an interval of 1.5 to 5 cm (Table S1) with the average depth of individual sectioned intervals as the representative depth. For gas geochemistry, we preserved 6 mL of sediments in a 36-mL serum bottle with 10 mL of 1 M NaOH, and sealed with a butyl rubber and an aluminum ring. Following the gas sampling, 3 mL of sediments were collected in a 15-mL centrifuge tube for the determination of water content. Samples for pore water content were subject to freeze drying. The weight difference was used to calculate the water weight content or porosity assuming the density of dry sediment was 2.5g cm^{-3} and the pore space was completely saturated with pore water. For aqueous geochemistry, the remaining sediments were placed in a 50-mL centrifuge tube and centrifuged at 8,2000 x g for 15 minutes to collect pore water. The obtained pore water was decanted from the centrifuge tube, 0.22-μm-filtered using syringe filters and split it into two fractions with one for ion chromatographic analyses of anion abundances and the other for dissolved inorganic carbon (DIC). For molecular analyses, sediments were placed in a 50-mL centrifuge tube. Upon arriving at the laboratory, anion and DNA samples were stored in a 4°C refrigerator and a −80°C freezer, respectively, until further analysis.

Geochemical analyses

Concentrations of gaseous hydrocarbon compounds in head space and dissolved inorganic carbon (DIC) in pore water were analyzed using a 6890N gas chromatograph (GC; Agilent Technologies, Santa Clara, CA, USA) equipped with a Porapak Q packed column (3 m) in line with a flame ionization detector and a thermal conductivity detector (6890N, Agilent Taiwan, Taipei, Taiwan). The measured partial pressure of a specific gas compound was used to calculate the equilibrium dissolved concentration with the Henry's law constant (Wiesenburg and Guinasso, 1979). The total moles in headspace and dissolved phase were summed up and normalized to the volume of pore water in order to obtain the dissolved concentration.

The $\delta^{13}C$ values of DIC was determined by acidifying pore water with 85% phosphoric acid for the production of CO_2 from DIC. Carbon isotope compositions of both methane and DIC were measured using a MAT253 isotope ratio mass spectrometer (IRMS) connected with a GC Isolink (Thermo Fisher Scientific, Waltham, MA, USA). The isotopic compositions were reported as the δ notation (δ value = $(R_{sample}/R_{standard} - 1) \times 1000‰$), where R is the ratio of ^{13}C to ^{12}C, and the standard is Pee Dee Belemnite (PDB) for carbon isotopes, was used in reporting the data.

Two anions in pore water, chloride, and sulfate were analyzed using an ICS-3000 ion chromatograph (Thermo Fisher Scientific, Waltham, MA, USA). Concentrations of particulate total organic carbon (TOC), total inorganic carbon (TIC), total nitrogen (TN), and total sulfur (TS) were determined by an elemental analyzer (MICROcube, Elementar, Germany). The uncertainties for aqueous and gas geochemistry, elemental abundance, and $\delta^{13}C$ value are ±2%, ±5%, ±2%, and ±0.3‰, respectively. The detectable limits for anions with the consideration of dilution were 10 ppm.

Microbial community compositions

DNA extraction and amplification of 16S rRNA gene

Crude DNA for 16S rRNA gene analyses was extracted from 2 to 5 g of fluids/sediments using the PowerSoil DNA Isolation Kit (Qiagen, Hilden, Germany). Bubbling fluids (if available) and sediments distributed across geochemical transition were selected for DNA extraction. These samples are representative of communities inhabiting the subsurface source region (represented by bubbling fluids) or subjected to the redox gradient developed after the sediment deposition (represented by cored sediments in adjacent mud platform). A total of 136 DNA extracts were obtained and stored at −80 °C for subsequent analyses. Polymerase chain reaction (PCR) amplification was applied to amplify the V4 hypervariable region of 16S rRNA genes using the primers F515 (5′–GTG CCA GCM GCC GCG GTA A–3′) and R806 (5′–CCC GTC AAT TCM TTT RAG T–3′) that target both bacterial and archaeal communities (Kozich et al., 2013). Sample specific barcodes and Illumina-specific adapters were appended with both forward and reverse primers. The ingredients of each PCR mixture contained 1.1–1.5 ng of purified genomic DNA, 1 U of ExTaq polymerase (TaKaRa Bio, Japan), 0.2 mM of dNTPs, 0.2μM of each primer, and 2.5μl of 10 × PCR buffer in a total volume of 25 μL. The program of thermal cycling involved a denaturation step at 94°C for 3 minutes followed by 30 cycles of denaturation at 94°C for 45 seconds, annealing at 55°C for 45 seconds, extension at 72°C for 90 seconds, and a final extension step at 72°C for 10 minutes. The products of three PCR runs for individual samples were pooled, analyzed by gel electrophoresis for size verification (~350 bp), and purified using the DNA Clean

and Concentrator Kit (Zymo Research, United States). Amplicons from different samples were pooled in equal quantities sufficient for sequencing on an Illumina MiSeq platform (Illumina, United States).

Sequence processing

Sequences of 16S rRNA gene amplicons were analyzed using the Mothur and QIIME2 (Schloss et al., 2009; Bolyen et al., 2018). Specifically, sequences for individual samples were binned in accordance with the barcode sequences. To minimize the effects of random-sequencing errors, reads that had two or more mismatches to the barcode sequences were discarded. The split raw FASTQ data were processed with the DADA2 (Callahan et al., 2017) implemented in the QIIME2 (version 2018.8; http://qiime2.org/) (Bolyen et al., 2018; Caporaso et al., 2010) to calculate the amplicon sequence variants (ASVs) in each sample. After removing the sequencing adapters, the first 31 nucleotides of primer sequences were trimmed off. Due to the decrease of quality at the end of each read, forward and reverse sequences were truncated to a length of 220 and 200 base pairs, respectively, to obtain individual sequences with a quality score greater than 20. Subsequently, denoised reads were assembled to full sequences, aligned, and taxonomically assigned against the Silva v.132 reference set using Mothur. Sequences identified as chloroplasts and mitochondria were removed. The obtained sequences were deposited in GenBank with accession number PRJNA560274.

Statistics

Microbial community analyses

Samples were rarefied to 9,413 sequences per sample through 100 sequence random re-sampling (without replacement) of the original amplicon sequence variants (ASV) table to account for the difference in sequencing depth for the calculation of alpha diversity indices such as observed ASV richness, Chao1 and Shannon indices (Hill, 1973; Chao et. al., 1984). Based on the rarefied dataset, alpha diversity indices, such as observed ASV richness, Chao1 and Shannon indices were computed. For the beta diversity calculation, the entire ASV table was used and normalized using the function cumNorm from the R package metagenomeSeq (Paulson et al., 2013). A cumulative-sum scaling method was used to calculate the scaling factors equal to the sum of counts up to a particular quantile in order to normalize the read counts with uneven sequencing depth (Paulson et al., 2013). The dissimilarity matrix between samples was computed using the Bray-Curtis method (Bray et al., 1957) and visualized through the ordination of non-metric multidimensional scaling (NMDS) and constrained correspondence analysis (CCA). The significance of environmental variables relative to the CCA ordinations was computed using "envfit" and 999 permutations. All statistical analyses were performed in R using the packages *vegan, ggplot2,* and *phyloseq.*

Estimation of habitat similarities

Approaches described in Ranjard et al. (2013) and Powell et al. (2015) were adopted with some modification. An estimation of habitat similarities was calculated from the Euclidean distances between paired 126 samples with the available concentrations of chloride, sulfate,

methane, TN, TS, TIC, and TOC. To reduce the effects of using large concentration scales, such as for chloride, environmental factors were normalized to their minimum and maximum values to scale the data to a fixed range between 0 to 1. The transformed dataset was used to evaluate habitat similarity using the following formula (Ranjard et al., 2013; Powell et al.,2015):

$$E_d = \left(1 - {\mathrm{Euc}_d}/{\mathrm{Euc}_{max}}\right) \quad \text{Eq.1}$$

where Euc_d is the Euclidean distance, and Euc_{max} is the maximum distance between sites in the matrix. To test whether community similarity was significantly correlated with a variety of spatial components, non-parametric Mantel tests based on the Pearson correlation coefficient were applied with significance assessed based on 1000 Monte Carlo permutations. All statistical analyses were performed in R using the package *vegan.*

Distance decay relationships (DDR)

To assess the DDR, pairwise community similarities between samples were calculated using the Sørensen-Dice index (Dice, 1945). The pairwise similarity was transformed in a logarithmic space to enhance the linear fitting using the following H formula:

$$\log_{10}(S_{com}) = \log_{10}(a) + \beta \log_{10}(D) \quad \text{Eq.2}$$

where S_{com} is the pairwise similarity in community composition, *D is* the geographic and/or vertical distance between two samples, and β is the slope. Null values in the similarity/distance matrices were assumed to be 0.001 prior to the log-transformation. The distance between samples was aggregated from two categories for samples in separate cores or within the individual cores. For samples in separate cores, the distance represents the geographic distance between MVs and was calculated using the function *geodist* in the R package 'gmt'. For samples within the individual cores, the distance represents the depth difference between samples. Samples collected from the bubbling pools were regarded as the surface material (0 cm) of each sediment core. The DDR relationships were assessed for data encompassing all samples or either categories. The significance of β was tested by 1000 Monte Carlo permutations of the residuals under the full regression [25], and β was found to be significant for each sample surveyed ($P < 0.001$).

Results

Physical and geochemical characteristics

The pairwise distance between samples ranged from 2.5 to 160 cm within cores and 0.005 to 9,924 km between cores (Fig. 1). Geochemical profiles of pore water showed various characteristics related to abiotic and microbial processes. Chloride concentrations varied highly among sites (ranging between 82 mM at SI02 in Myanmar and 4890 mM at GG01 in Iran) and generally decreased with increasing depth in individual cores (Figure S1). Exceptions occurred at

PA02, SH01, SI02, and LGH03, with substantial fluctuations in the middle or bottom part of the cores. Sulfate concentrations ranged from below the detectable level at SM22, AK03, GJ01, TA, PA01, PA02, and LGH03 to 288 mM at GG01, with most data clustering between 0.5 and 2 mM. Variations in sulfate concentration for cores with detectable sulfate were further categorized into three patterns, including depth-dependent decrease (DSZ01 and SYNH02C4) and increase (GG01, COM01, and SH01), and substantial fluctuation along the depth (AR01 and SI02).

Methane concentrations ranged between 0.006 mM (PA02) and 3.98 mM (SYMH02C4), with most data clustering between 0.2 and 1 mM (Figure S2). Methane concentrations either increased (DSZ01, SM22, GJ01, TA, SYNH02C4, and LGH03) or decreased (GG01, AR01, COM01, SH01, and SI02) with increasing depth. The $\delta^{13}C$ values of methane clustered between -58‰ and -35‰ and exhibited a trend opposite to that of methane concentration. The molar ratios of methane over ethane and propane (C1 (methane) / (C2 (ethane) + C3 (propane))) were variable and ranged from 22 (SI02) to approximately 1200 (AR01 and COM01; Figure S3).

Community structures and compositions

Analyses of 16S rRNA genes yielded a total of 4,562,760 sequences. The number of observed ASVs for individual samples ranged between 58 and 1,462 with an average value of 449 ± 250 when singletons (presence of one sequence for an ASV at only one depth) were included. The trends of diversity indices exhibited a similar pattern (Figure S4). The lowest values of alpha diversity indices occurred at SI02 and SH01 in Myanmar, whereas the highest values were found for AR01 in Italy. The number of observed ASVs for individual MVs ranged between 204 (SI02) and 4,203 (AR01). Accumulation curves at coarse taxonomic resolution (i.e., phylum to family) revealed the sufficiency of our sequencing effort. In contrast, at the level of ASV, the accumulation curve showed a continuously increasing trend, indicating that the diversity of the entire MV community was not fully captured (Figure S5).

Supplementary Tables

Table S1. Core length and coordinate of sampled terrestrial mud volcanoes across the Eurasian continent.

Sample name	Country	Core Length (cm)	Section interval (cm)	Longitude	Latitude
AR01C1	Italy	67	3	13.60000	37.37667
COM01C1	Italy	45	2.5	13.65194	37.44306
PA01C1	Italy	55	3	14.91972	37.54472
PA02C1	Italy	41	2.75	14.89028	37.57278
AK03C1	Georgia	49	2.5	45.91322	41.41953
GJ01C1	Georgia	44	2.5	45.79261	41.74531
QK01C1	Georgia	22	1.5	45.80564	41.28905
GG01C1	Iran	46	2.5	54.39608	37.11856
TA01C1	Iran	47	3	59.93306	25.46697
TA03C1	Iran	47	3	59.93306	25.46697
SM22C1	China	33	2	84.38722	44.18269
DSZ01C1	China	21	2	84.84636	44.30517
SH01C1	Myanmar	38	2	93.57119	19.36975
SI02C1	Myanmar	48	2	93.59169	19.39778
LGH03C4	Taiwan	160	5	121.20940	22.98306
SYNH02C4	Taiwan	52	2.5	120.409479	22.80313
SYNH02C11	Taiwan	20	2	120.409479	22.80313

Table S2. Multiple linear regression model for Shannon index versus significant geochemical parameters, after collinear variables were removed and Akaike information criterion (AIC) was applied. Variables were added to the model to generate the highest to lowest best fit from simple linear regression.

	Estimate	Std. Error	t-value	*P*-value	Signif. code
(Intercept)	4.02048	0.32306	12.445	< 0.0001	***
Methane	149.79125	76.02220	1.970	0.05106	.
TN	8.66657	3.13095	2.768	0.00652	**
TIC	0.43655	0.08123	5.374	< 0.0001	***

Multiple R^2: 0.2067, adjusted R^2: 0.1872
F-statistic: 10.6 on 3 and 122 degrees of freedom, p-value: 3.047E-06
Significance codes: '***': 0-0.001; '**': 0.001-0.01; '.': 0.05-0.1

Table S3. Simple linear regression of Shannon index against individual geochemical parameters.

	Slope	Std. Error	t-value	*P*-vale	R^2
Sulfate	0.0004	0.0016	0.277	0.782	0.001
Chloride	0.0001	0.0001	1.808	0.0731	0.018
Methane	94.391	83.070	1.172	0.243	0.002
TN	2.733	3.267	0.837	0.404	0.002
TS	0.0985	0.1908	0.517	0.606	0.006
TIC	0.3577	0.0793	4.511	< 0.0001	0.134
TOC	0.3584	0.1862	1.925	0.0566	0.021

Table S4. Mantel test using Spearman's correlation (permutations = 999) of Bray-Curtis dissimilarities between all communities, and each geochemical parameter or geographic distance (km).

	Rho (ρ)	*p*	Signif. code
km	0.322	< 0.001	***
env	0.178	< 0.001	***
Chloride	0.454	< 0.001	***
Sulfate	0.258	< 0.001	***
Methane	0.068	0.026	**
TIC	0.255	< 0.001	***
TOC	-0.081	0.986	-
TN	-0.001	0.481	-
TS	0.143	< 0.001	***

Signifance codes: '***': 0-0.001; '-': 0.1-1

Table S5. Permutational multivariate analysis of variance (using continuous variables only) of beta diversity using Bray-Curtis dissimilarity.

	Df	SumsOfSqs	MeanSqs	F.Model	R^2	*P*-value	Signif. codes
Chloride	1	1.504	1.50431	6.1690	0.02586	< 0.001	***
Sulfate	1	1.719	1.71883	7.0488	0.02954	< 0.001	***
Methane	1	0.509	0.50856	2.0856	0.00874	< 0.001	***
TIC	1	2.117	2.11708	8.6819	0.03639	< 0.001	***
TOC	1	1.377	1.37699	5.6469	0.02367	< 0.001	***
TN	1	1.828	1.82757	7.4947	0.03141	< 0.001	***
TS	1	1.177	1.17735	4	0.02024	< 0.001	***

Significance codes: '***': 0-0.001

Table S6. Coefficient of variation and the Pearson's correlation coefficient (r) for chloride and sulfate concentrations.

Sample name	Chloride	Sulfate	r	Signif. codes
AR01C1	8.97%	117.09%	0.15	-
COM01C1	8.37%	122.09%	0.05	-
PA01C1	3.61%	NA	NA	NA
PA02C1	13.24%	NA	NA	NA
AK03C1	7.55%	NA	NA	NA
GJ01C1	5.61%	NA	NA	NA
QK01C1	63.13%	NA	NA	NA
GG01C1	14.84%	17.49%	-0.35	-
TA01C1	2.15%	NA	NA	NA
SM22C1	7.08%	NA	NA	NA
DSZ01C1	25.927%	186.04%	0.46	-
SH01C1	15.58%	102.65%	0.91	***
SI02C1	3.45%	29.34%	0.24	.
LGH03C4	6.56%	NA	NA	NA
SYNH02C4	15.02%	125.62%	0.88	***
SYNH02C11	28.11%	138.05%	0.96	***

Significance codes: '***': 0-0.001; '.': 0.05-0.1; '-': 0.1-1
NA means concentration of sulfate below the level of detection

Supplementary Figures

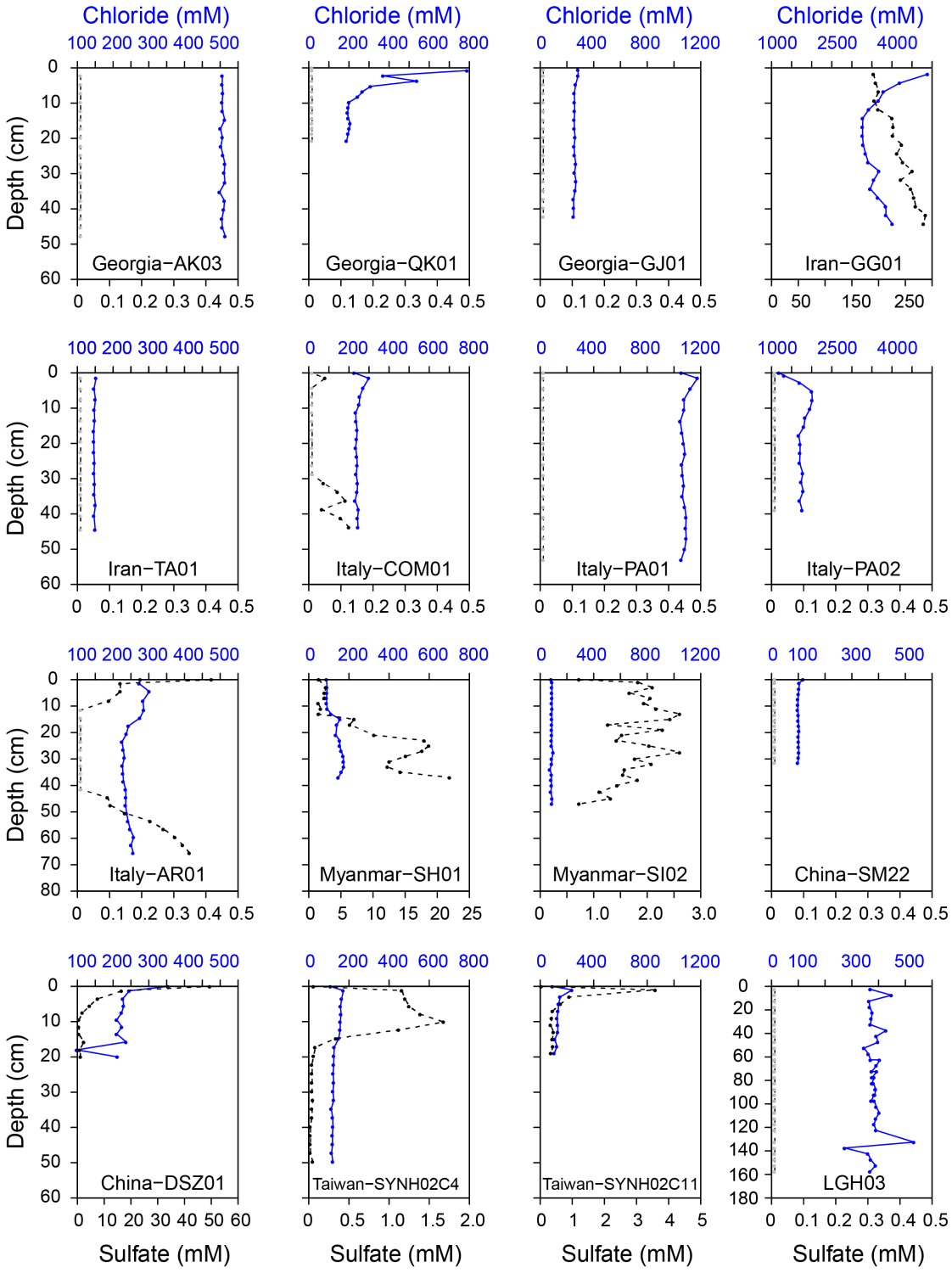

Figure S1: Chloride (in blue) and sulfate (in black) concentration profiles. Sulfate concentrations lower than the limit of detection (0.01 mM) are shown in gray.

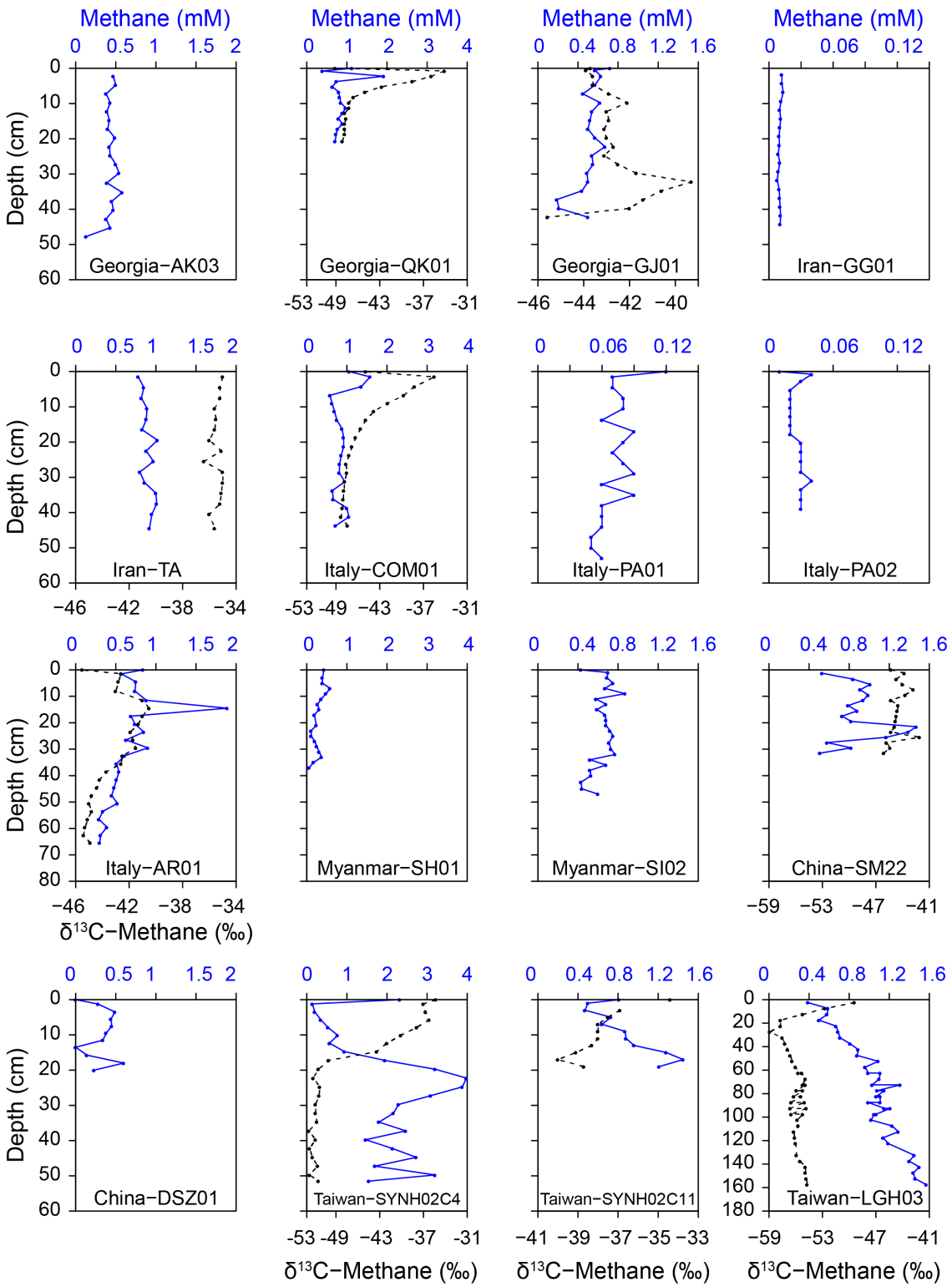

Figure S2: Profiles of methane concentrations (in blue) and δ13C values of methane (in black). Samples collected from Georgia are not sufficient for porosity measurement. Therefore, their concentrations are normalized to the weight of wet sediments. Assuming that the weight proportion of pore water is 0.5, the concentration of 1 mmole g-1 could be converted to 200 mM.

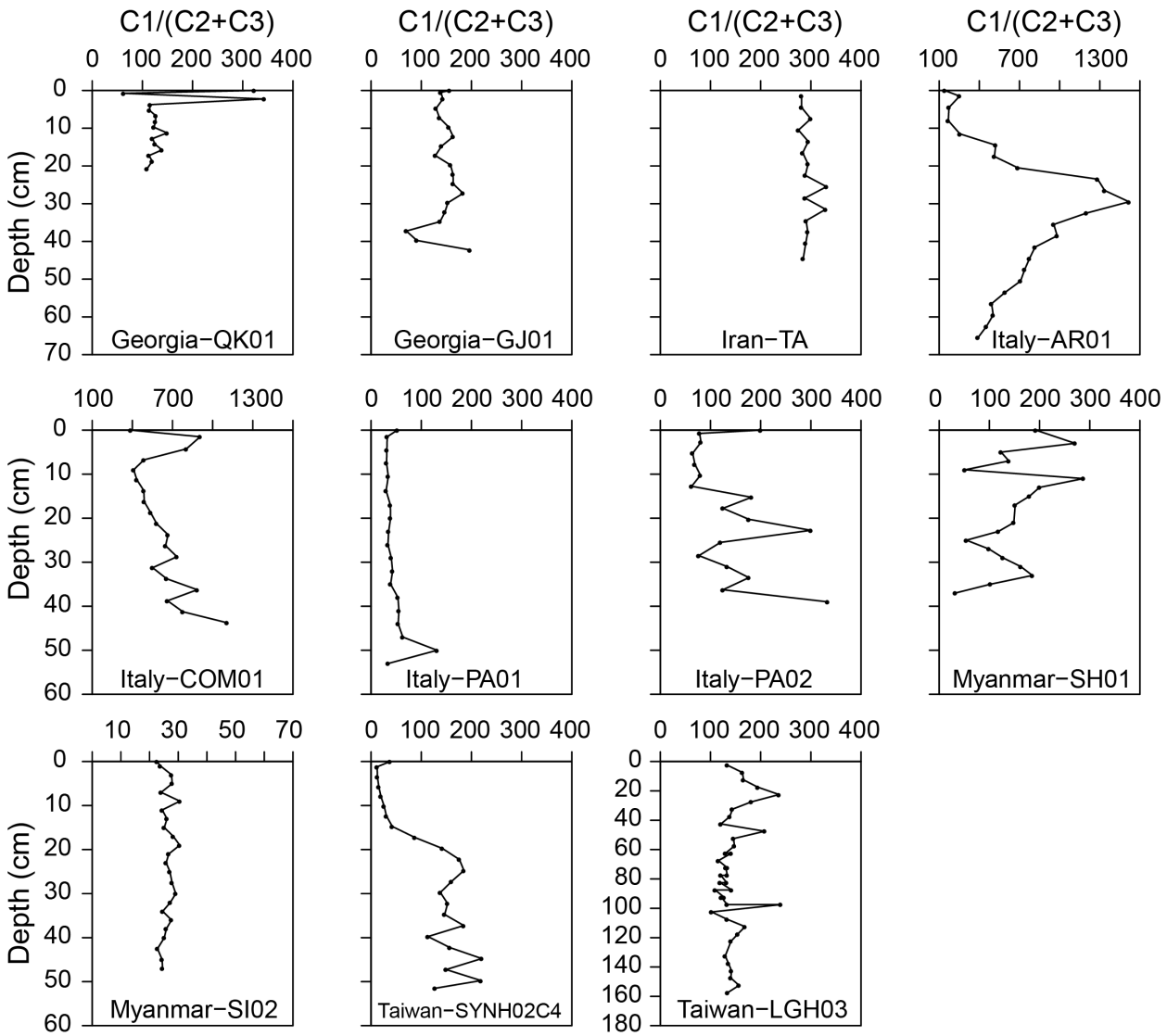

Figure S3: Molar ratio of methane over the sum of ethane and propane (C1 / C2 + C3).

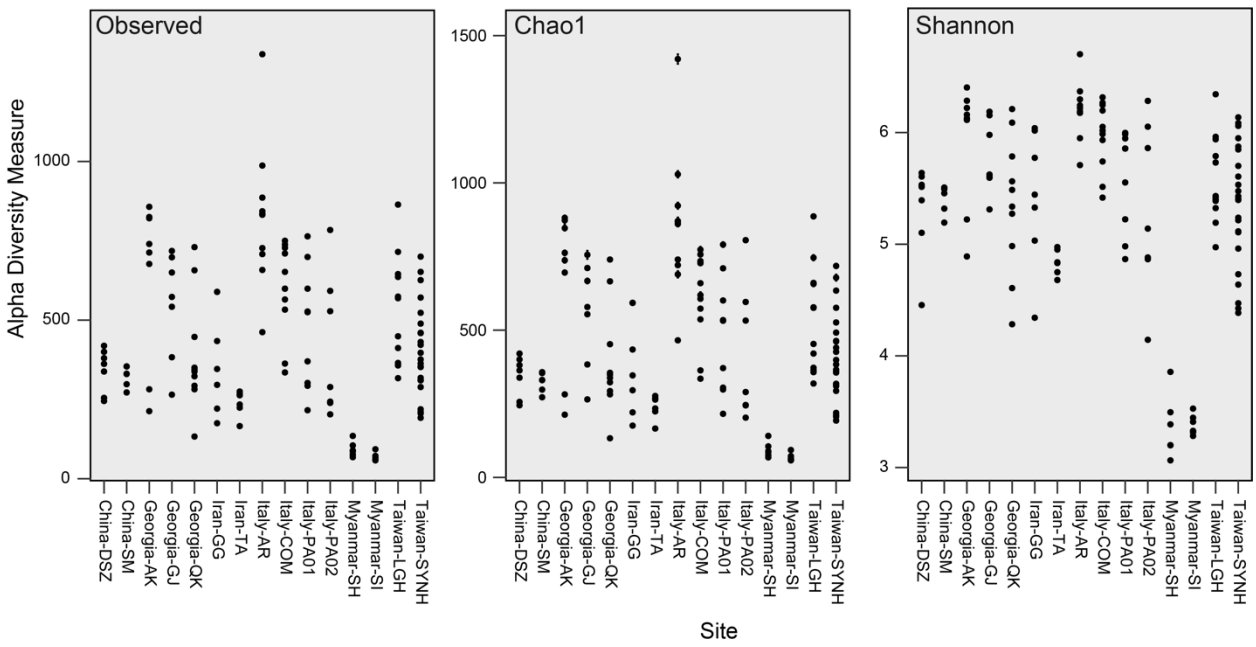

Figure S4: Alpha diversity indices calculated based on the rarefied dataset (n=9,413). No significant relationship was found between richness and sampling depth (Spearman's $\rho = 0.17$, $P > 0.01$ for the three indices).

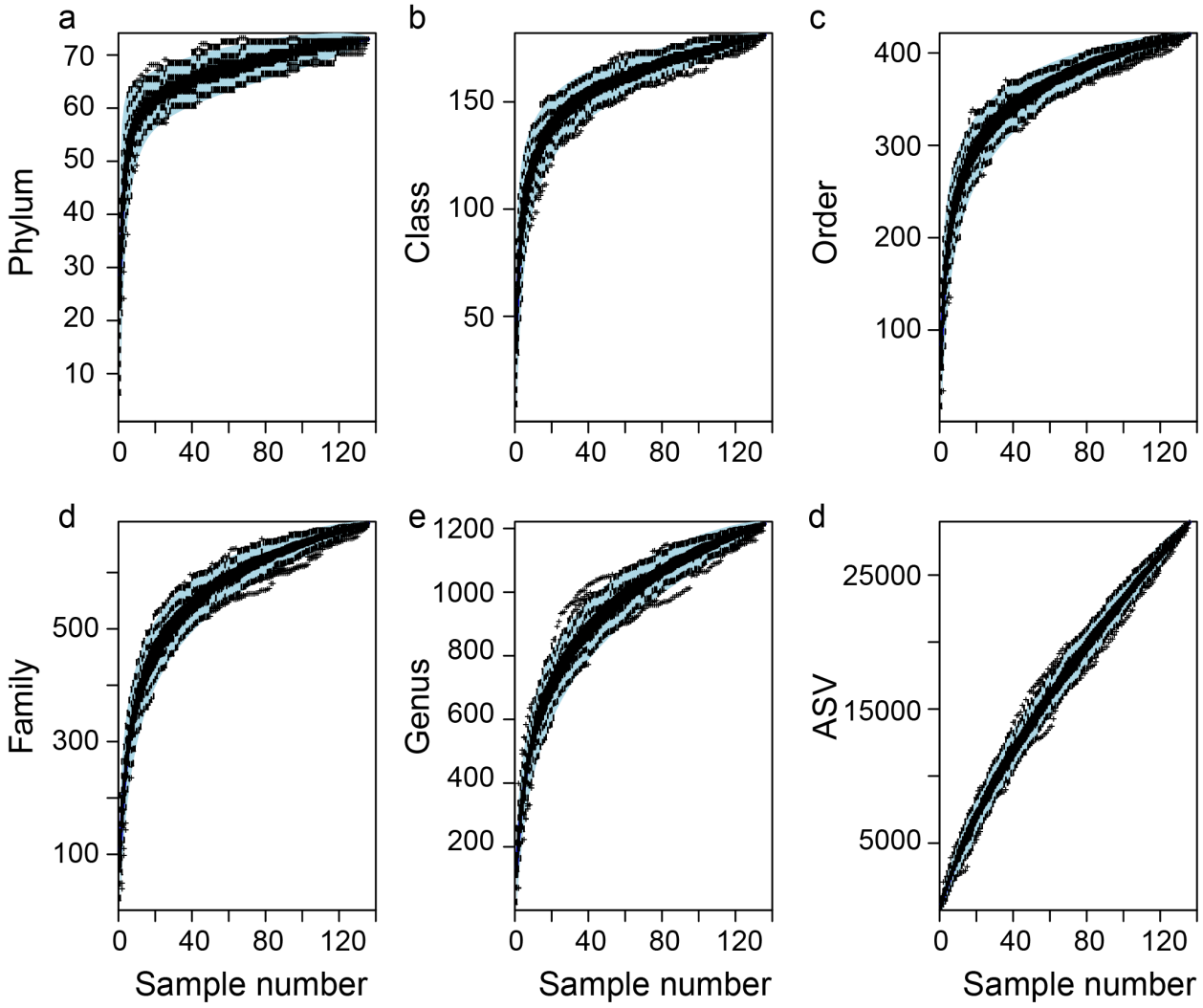

Figure S5: Accumulation curves based on different taxonomic units: (a) phylum (b) class, (c) order, (d) family, (e) genus, and (f) ASV. Boxplots show a summary of 100 permutations calculated with random subsampling. Absolute singletons were incorporated for comparison. Blue area depicts the 95% confidence interval.

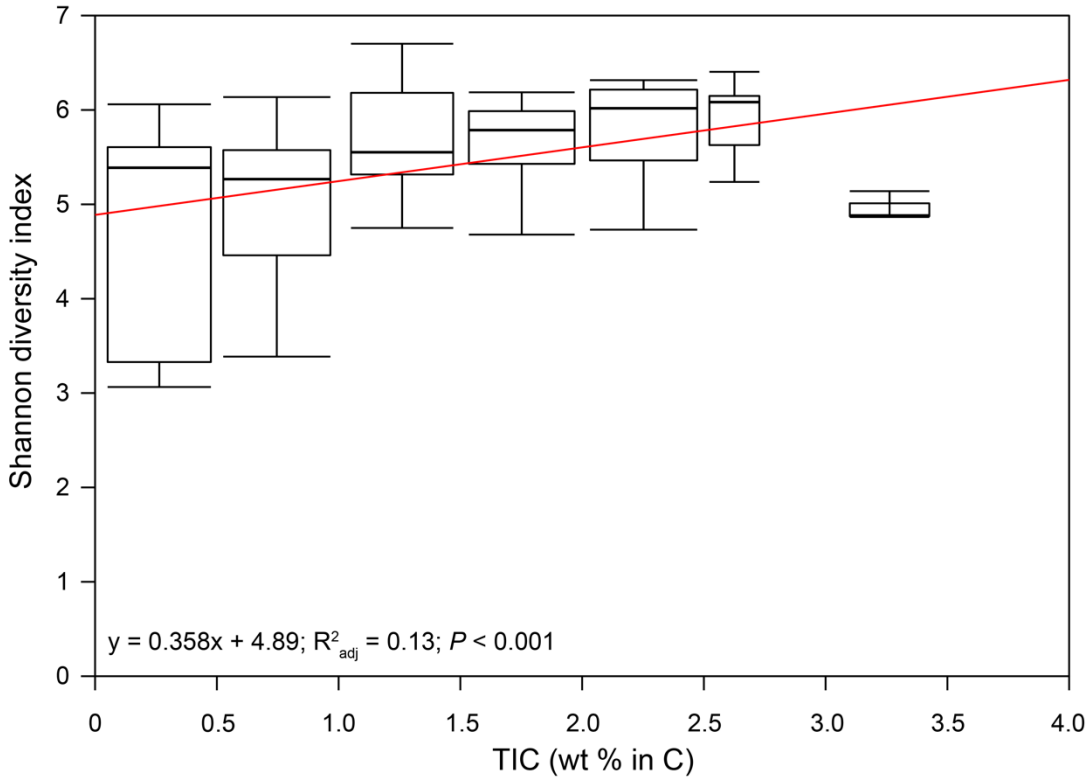

Figure S6: Plot of Shannon diversity versus TIC. Linear regression is shown in red (n=126). Box demonstrates the interquartile range that includes the first (25%), median (50%), and third quartiles. Lower and upper whiskers are the first and third quartiles minus and plus 1.5 times interquartile range, respectively.

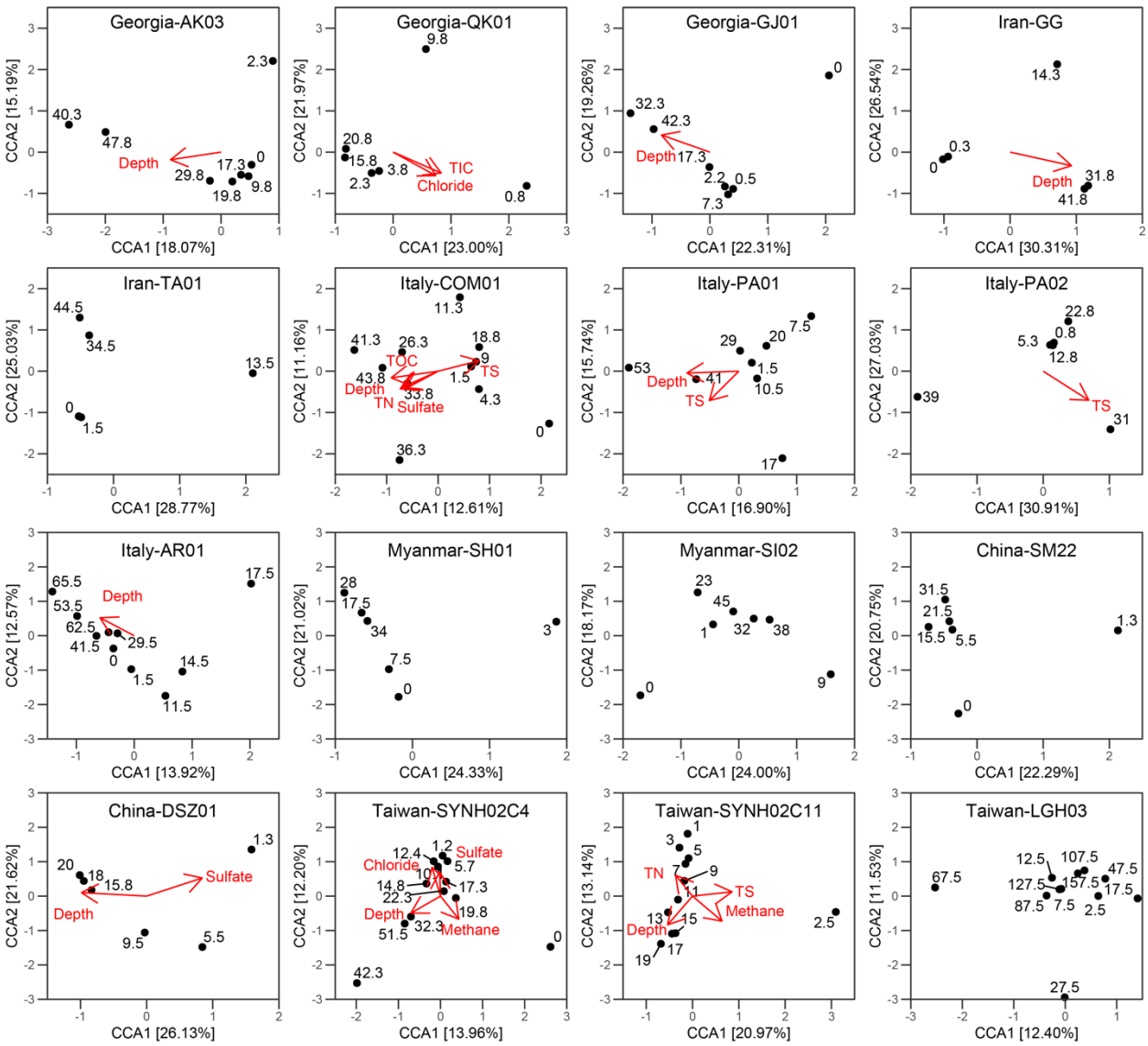

Figure S7: Constrained correspondence analysis of community relatedness quantified by the Bray–Curtis distance with the overlay of ordination for significant environmental parameters. Numbers next to each data point indicate sampling depth (in centimeter).

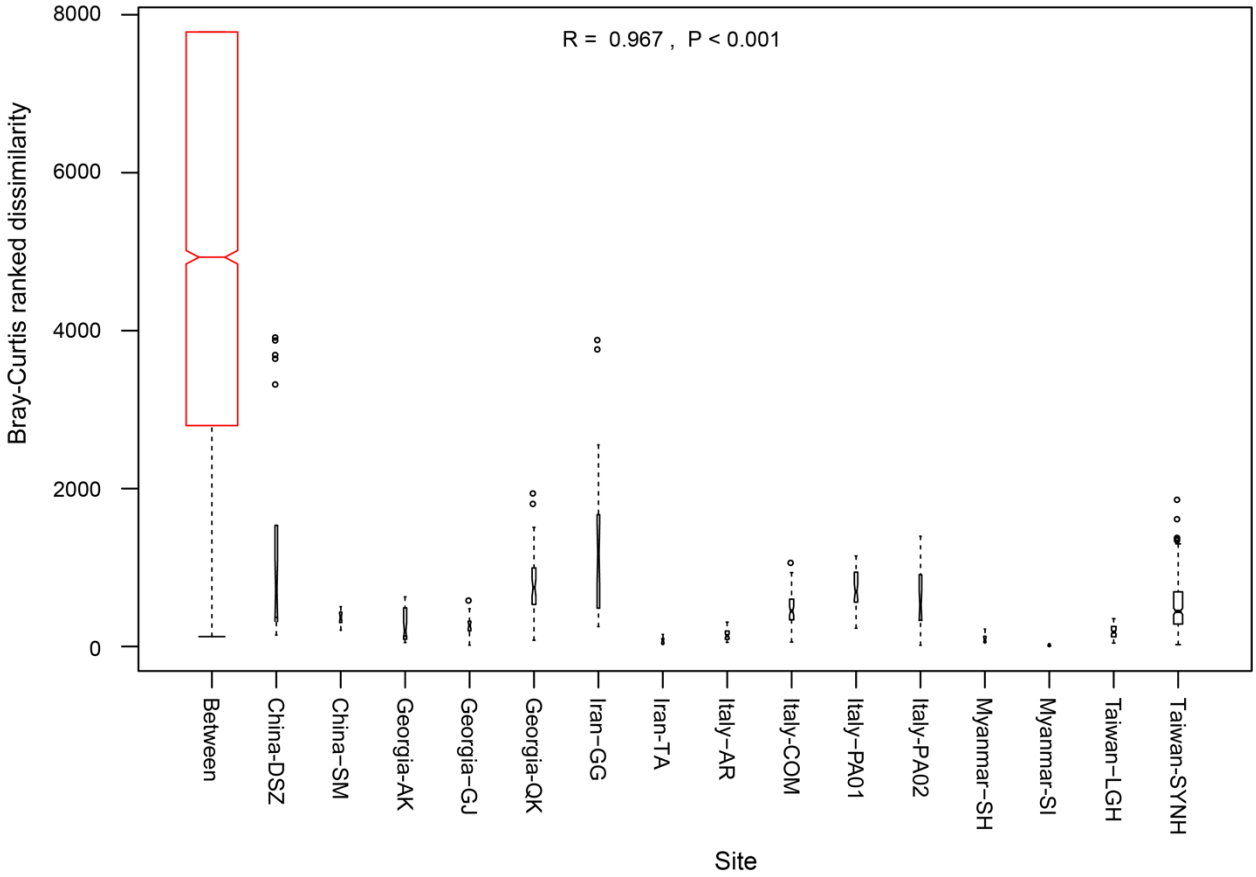

Figure S8: Analysis of similarities (ANOSIM: |R|) for community dissimilarity between all sites (in red) and within individual sites (in black). Lower and upper whiskers are first and third quartiles minus and plus 1.5 times interquartile range, respectively.

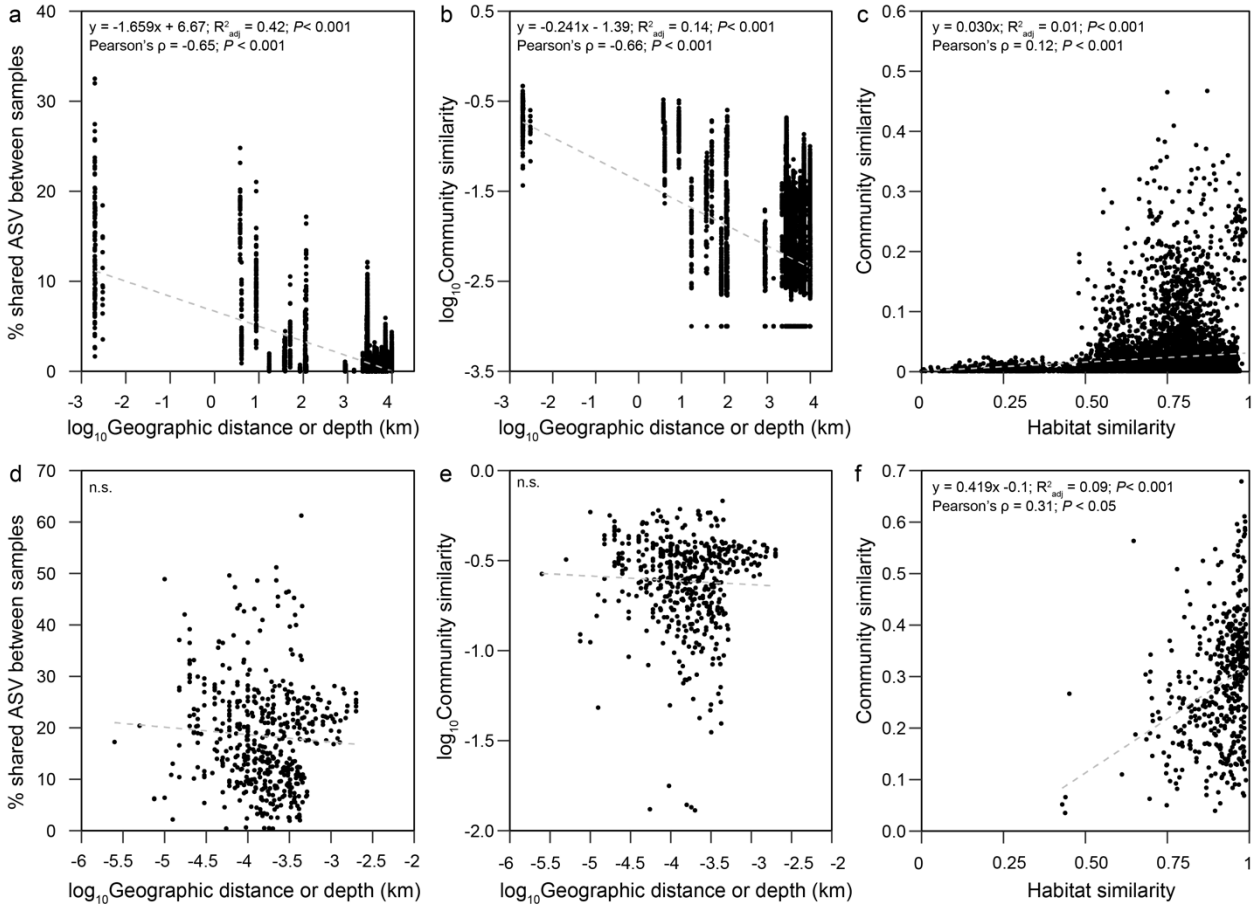

Figure S9: Distance–decay and geographic patterns of microbial communities across cores (a)–(c), and within cores (d)–(f).

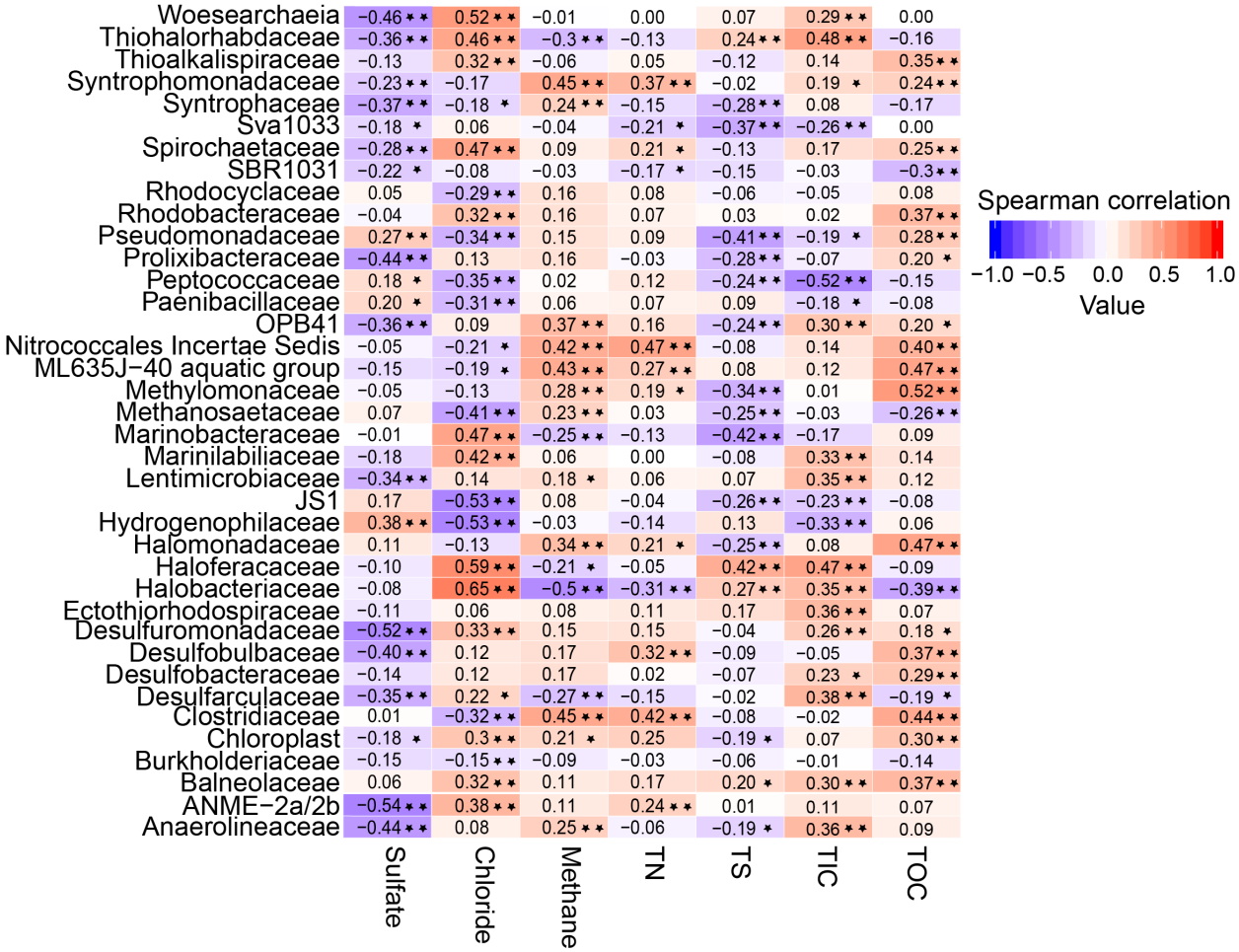

Figure S10: Correlation between concentrations of geochemical parameters and abundances of 38 major families. Color code represents the relative Spearman correlation coefficient. Major families are selected based on the top 50 most abundant families. * and ** denote *P* values less than 0.01 and 0.05, respectively.

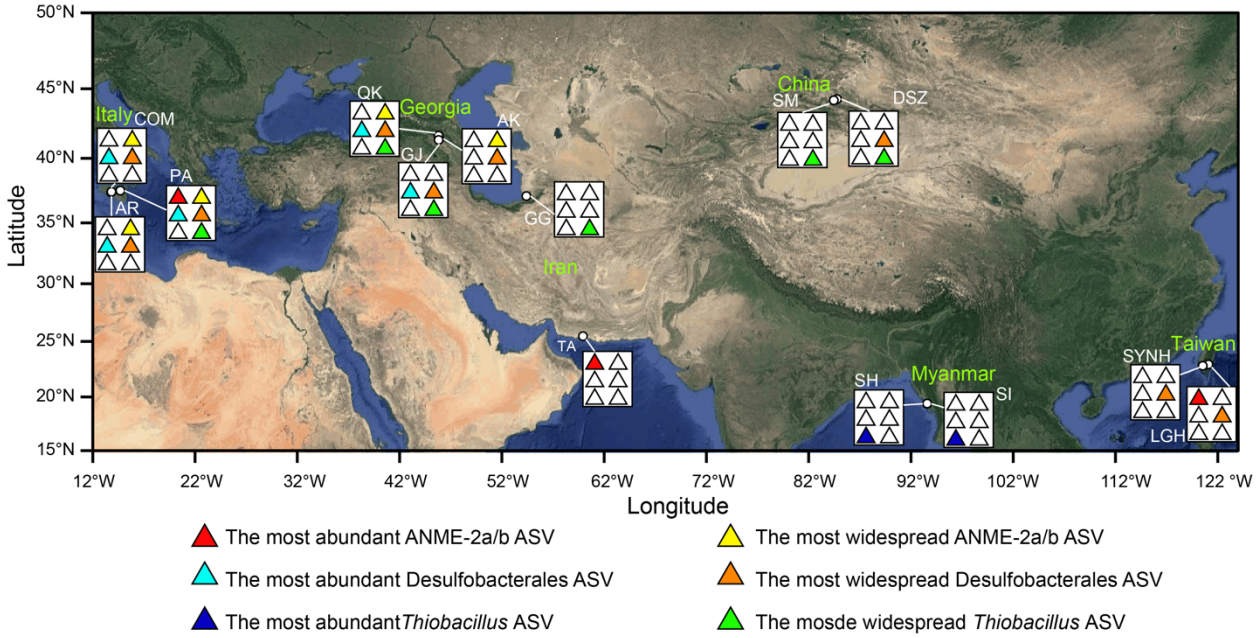

Figure S11: Occurrence of ASVs affiliated key taxa likely involved in methane and sulfur cycling. Each sub-panel consists of six color codes indicating the presence or absence of six key ASVs likely involved in methane and sulfur cycling. These target ASVs include (1) the most abundant ANME-2a (in red), Desulfobacterales (in blue-green), and *Thiobacullus* (in blue) ASVs, and the most widespread ANME-2a (in yellow), Desulfobacterales (in orange), and Thiobacillus (in light green) ASVs. The basal map is from Google Maps © Google Maps 2021.