Supplementary materials for

Novel microbial assemblages inhabiting crustal fluids within mid-ocean ridge flank subsurface basalt

Sean P. Jungbluth[1,2], Robert Bowers[3,a], Huei-Ting Lin[2], James P. Cowen[2,‡], and Michael S. Rappé[1]

[1]Hawaii Institute of Marine Biology, SOEST, University of Hawaii, P.O. Box 1346, Kaneohe, HI 96744

[2]Department of Oceanography, SOEST, University of Hawaii, 1000 Pope Rd., Honolulu, HI 96822

[3]NASA Astrobiology Institute, IfA, University of Hawaii, 2680 Woodlawn Drive Honolulu, HI 96822

[a]Current address: DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

[‡]Deceased [Hawaii, 7/27/2013]

SUPPLEMENTARY METHODS

Sample collection and preparation

Fluids were sampled using a deep sea pumping system or the GeoMICROBE instrumented sampling platform (Cowen *et al*., 2012; Lin *et al*., 2012; Jungbluth *et al*., 2013a). Borehole fluids were typically allowed to flush for an amount of time required for expulsion of at least three times the fluid delivery line volume. Because the first generation CORK observatory at borehole 1025C does not have a fluid delivery line, it was flushed for ~70 hours using a custom-designed borehole flushing system to clear the large void within the borehole casing (Jungbluth *et al*., 2014). Flushing was intended to clear 3x the volume of the borehole casing, but fluid expulsion rates were lower than expected and indicate that only 1x volume of the borehole casing was flushed. Following flushing, fluids were collected in custom acid-washed 15 liter-volume Tedlar bags (Medium Volume Bag Samples; MVBS) or 60 liter-volume Tedlar bags protected by rigid boxes (Large Volume Water Samples; LVWS) (Table S1), with the exception that foil bags were used for MVBS samples SSF18-20. In addition, particulates were filtered *in situ* using either the GeoMICROBE sled or a fluid sampling system attached to the submersible (Table S1).

Seawater samples were collected in the vicinity of the boreholes in order to provide background controls (Table S1). Seawater was collected via a Niskin rosette from within the nepheloid layer (5-10 m above seafloor) and just above the nepheloid layer (~100 m above seafloor). In addition, in 2010 and 2011 a 5 L Niskin bottle was fitted to the ROV *Jason II* and used to collect seawater from a depth of approximately 2650 m in the vicinity of CORK U1301A. Also, seawater in the vicinity of borehole

1025C was sampled 35 meters above the seafloor using the LVWS bag sampler and deep sea pumping system. Sediment samples were collected from site U1363 located nearby during IODP expedition 327 (Expedition 327 Scientists, 2011a) and are described in detail elsewhere (Expedition 327 Scientists, 2011b; Jungbluth *et al*., 2013b).

In 2008, whole fluids were filtered shipboard onto 25 mm-diameter 0.1 µm-pore sized polyethersulfone membrane filters (PES) (Pall Corporation, Port Washington, NY, USA) and stored in 0.5 ml of DNA lysis buffer [20 mM Tris-HCl, 2 mM EDTA, 1.2% Triton X-100, 2% lysozyme (w/v), pH 8]. Samples filtered *in situ* via the GeoMICROBE were passed through 47 mm-diameter 0.2 µm-pore sized PES membrane filters that were immediately processed and stored in 3 ml of DNA lysis buffer when retrieved shipboard. Whole fluid samples retrieved in 2009-2011 were filtered shipboard through 0.22 µm-pore sized Sterivex-GP filter cartridges (Millipore Corporation, Billerica, MA, USA) and stored in 2 ml of DNA lysis buffer. In 2011, Steripak cartridge filters (Millipore Corp.) were used to filter fluids *in situ*, and subsequently stored in 25 ml of DNA lysis buffer when retrieved shipboard. In 2009, a 1.9 L sample of R/V *Atlantis* shipboard distilled water was filtered through a Sterivex-GP filter cartridge and stored in DNA lysis buffer as described above. All samples were stored shipboard at -80°C until transportation back to the laboratory.

Aliquots (1 ml) of whole water from all raw borehole fluid samples collected between 2009-2011 were cryopreserved within hours of sample shipboard sample retrieval in a final solution of 10% glycerol (w/v) and stored at -80°C until further processing.

Analytical methods for geochemistry

Basement fluid and sediment porewater geochemical methods are described in additional detail elsewhere (Expedition 327 Scientists, 2011c; Lin *et al*., 2015).

Major ions (Ca^{2+}, Mg^{2+}, K^{+}, Na+, Cl^{-}, SO_4^{2-} and Br^{-}) were analyzed by ion chromatography on a Dionex ICS-1100s (Sunnyvale, CA, USA). In addition, magnesium and calcium concentrations were also analyzed by EDTA (colorimetric) and EGTA (electrometric) titration (Grasshoff *et al*., 1999), or inductively coupled plasma optical emission spectroscopy (ICP-OES) (Lin *et al*., 2012). International Association for the Physical Sciences of the Oceans (IAPSO) standard seawater was used to standardize the methods. The reproducibility of Mg and Ca measurements were ~0.5 mM for all three methods.

Silicate, nitrate, nitrite, phosphate, dissolved sulfide and dissolved manganese concentrations were measured by colorimetry (Brewer and Spencer, 1971; Phillips *et al*., 1997; Grasshoff *et al*., 1999). Samples for silicate analysis were first diluted 100 times, reacted with acidic molybdenum and reduced with ascorbic acid. The detection limit was ~0.5 μM. Nitrate analysis was performed with a flow injection analyzer. Sample nitrate was first reduced to nitrite with a Cu-Cd column and then analyzed as nitrite. The reduction efficiency of the Cu-Cd column was 99-100%. The detection limit for nitrate analysis was 0.005 μM and analytical uncertainty was 0.002 μM using a 10 cm light-path cuvette. The detection limit for both nitrite and phosphate was 0.05 μM and the analytical uncertainty was 0.02 μM using a 1 cm light-path cuvette. The detection limits

were 0.01 μM and 0.5 μM for dissolved sulfide and dissolved manganese, respectively, using a 1 cm light-path cuvette.

Ammonium concentrations were measured by a flow injection-fluorometric method (Jones, 1991). The detection limit was ~2 μM for ammonium in basement fluids and the analytical uncertainty is 0.5 μM. Ferrous iron was measured directly by a Ferrozine colorimetry method (Stookey, 1970; Gibbs, 1976). For total iron analysis, samples were first reduced with ascorbic acid and analyzed as ferrous iron. The detection limit for both ferrous iron and total iron was 0.1 μM.

Dissolved organic carbon (DOC) was measured by high-temperature combustion using a TOC-VCSH analyzer (Sharp *et al*., 2002a; Dickson *et al*., 2007) (Shimadzu Corp., Kyoto, Japan). Combustion temperature was set at 720ºC. Samples were acidified to pH <2 within the syringe of the autosampler by adding 45 μL of 2N HCl to 3 mL samples. Acid contamination was monitored throughout the analysis by analysis of low carbon deionized water. Samples were purged within the syringe for two minutes to remove inorganic carbon. Using a 50 μL volume, five to six injections were performed for each sample. The reproducibility between replicate injections was <1 μM, or <2 % of a 40 μM concentration level. Analytical reference materials (ARM) supplied by Drs. Wen-Hao Chen and Dennis Hansel (University of Miami, FL, USA) were used for control purposes. At least one ARM was measured every five samples. The average measured concentration of the ARM was 42 ± 2 μM (n=44), which is within the reported value of 41~43 μM. The detection limit for DOC was ~2 μM.

Total dissolved nitrogen (TDN) was measured with a chemiluminescence detector in-line with a Shimadzu TOC-VCSH analyzer (Sharp *et al*., 2002b). Analytical

reference materials used in DOC analysis were again used to monitor instrumental performance. The measured average value for the ARM nitrogen was 32.5 ± 0.6 μM (n=42), which was within the range of the reported value of 32.3~33.7 μM. The detection limit of TDN was ~0.5 μM. Analytical uncertainty was less than 0.6 μM when sample concentration was lower than 40 μM (seawater) and about 2 μM when total nitrogen concentration was higher than 100 μM (basement fluids).

Alkalinity was determined by acid titration. Acid (0.1N HCl) was standardized with CO_2 certified reference materials (CRMs) purchased from the office of Andrew Dickson at Scripps Institution of Oceanography. Three to five aliquots of CRM were analyzed each day for acid concentration recalibration. An Orion 911600 Semi-micro pH electrode (ThermoFisher Scientific, Waltham, MA, USA) was used to measure the pH and electrode potential during the titration process. The Gran function plot method was used to evaluate titration end-points and calculate sample alkalinity (Dickson *et al*., 2007). The analytical reproducibility for alkalinity measurements was <0.02 mM.

Estimation of basement fluid end-member content

An analysis based on nitrate concentrations was one of two mechanisms used to estimate basement fluid end-member content in samples originating from boreholes. Fluid samples with undetectable nitrate concentration (<0.005 μM) were collected from boreholes U1362A and U1362B, indicating that end-member basement fluids likely contain undetectable nitrate. Measurable nitrate in borehole U1301A fluid samples was likely from intruded seawater. Nitrate concentration was used to estimate the basement fluid end-member content in each sample using the following equation:

$$P = (C_{i\_sw} - C_{i\_smp}) / (C_{i\_sw} - C_{i\_bf}) \times 100\% \quad \text{(SE1)}$$

where P is the percentage of the end-member basement fluid in the sample collected. C_{i_sw}, C_{i_smp}, and C_{i_bf} represent the concentration of computed parameter: i (e.g. nitrate or calcium, described below) in background seawater, sample, and end-member basement fluid, respectively.

Calcium concentrations were used to estimate the amount of basement fluid in samples SSF16-18 and SSF20 because no nitrate measurements were performed. Calcium was chosen over magnesium because of the insensitivity of the methods used to measure magnesium at low concentrations. The calcium concentration of venting fluids at 1025C (34.2 mM; Wheat *et al*., 2004) was used as the basement fluid end-member value to compute the percentage of basement fluid content in sample SSF9. For borehole U1301A samples, a calcium concentration previously reported from downhole sampling was used (55.6 mM; Wheat *et al*., 2010). Average calcium concentrations from borehole U1362A and U1362B samples with the lowest detectable nitrate were used as end-member values (54.1 mM and 55.5 mM, respectively).

Inference of unmeasured geochemical parameters

A subset of fluid samples was not sampled for the complete suite of geochemical parameters; the methods used to infer values for the missing measurements are described here. Most of the samples used for molecular analysis were from one of six sampling bags collected in sequence within the same 1-2 hours of sampling from a

single borehole. While magnesium and calcium concentrations were measured from fluids taken from each of the six bags, other measurements were not made for samples SSF16-18 and SSF20. Statistically indistinguishable (Q test, confidence level=95%) calcium concentrations were found between SSF17-18 and SSF20 and fluids collected in other bags from their same deployment series, suggesting that the chemical compositions of SSF17-18 and SSF20 are likely representable by the average values from fluid samples collected in other bags. However, the calcium concentration of SSF16 (53.3 mM) is significantly different (Q test, confidence level=95%) than those from fluids collected in other bags from their same deployment series, (55.5 ± 0.3 mM, n=7), likely due to seawater intrusion during sampling. Sample SSF16 is estimated to contain 95% of end-member basement fluid; nitrate, ammonium, sodium, potassium, alkalinity and phosphate were estimated based on the contribution of end-member basement fluid and seawater using a rearrangement of equation (SE1):

$$C_{i\_smp} = C_{i\_bf} \times P + C_{i\_sw} \times (1 - P) \qquad \text{(SE2)}$$

DNA extraction

To recover environmental DNA, most membrane filter samples from subsurface fluids and seawater were thawed to room temperature and extracted using the PowerSoil DNA isolation kit (MOBIO Laboratories, Carlsbad, CA, USA) following the manufacturer's recommended protocol. However, instead of the PowerSoil DNA isolation kit, samples SSF11, SSF12, SSF21-22, SSF23-24 were extracted as follows: a 40 µl solution of 50 mg ml^{-1} lysozyme (Sigma-Aldrich, St. Louis, MO, USA) in DNA lysis

buffer was added to the Sterivex filters and rotated for 45 min at 37°C. Proteinase K (Qiagen Corp., Valencia, CA, USA) was subsequently added to a final concentration of >0.55 μAU, SDS (ThermoFisher Scientific) was added to a final concentration of 1%, and the samples were rotated for an additional 2 h at 55°C. Lysates were transferred to 30 ml Oak Ridge tubes using a sterile syringe. An additional 1 ml of lysis buffer was added to each of the filters for washing at 55°C for 15 min and pooled with the initial lysates. A 3 ml volume of phenol:chloroform:isoamyl alcohol (25:24:1; pH 8.0) was added and the mixtures vortexed for 30 sec and centrifuged for 5 min at 2500 x g. The aqueous phase was subsequently transferred to new Oak Ridge tubes, 3 ml of chloroform:isoamyl alcohol (24:1) was added to each, the mixture was vortexed for 30 sec, and subsequently centrifuged for 5 min at 2500 x g. The aqueous phase was concentrated for 20 min by spin dialysis using Amicon Ultracel-30K filters (Millipore Corp.) and centrifugation at 1000 x g. Flow-throughs were decanted and the Amicon filters were spun again at 1000 x g for 20 min. A 1 ml volume of TE buffer [10mM Tris-HCl (pH 8.0), 1mM EDTA (pH 8.0)] was added to each of the Amicon filter membranes and spun at 1000 x g for 10 min; ~700 μl remained on each of the columns and was transferred to a new microcentrifuge tube. The filter columns were washed twice with 700 μl of TE buffer and pooled with the initial DNA concentrate. Finally, the resulting nucleic acids were concentrated using a vacuum centrifuge and resuspended in 50 μl of PCR-grade water.

After the transfer of lysates from the filter holders described above, the membrane filters were also manually excised and extracted using the PowerMax Soil DNA isolation kit (MOBIO Laboratories) following the manufacturer's specifications.

Environmental DNA from both extractions were pooled together and quantified via a Quant-iT™ dsDNA Assay High Sensitivity Kit (Life Technologies, Carlsbad, CA, USA).

All sediment samples were thawed to room temperature and environmental DNA was extracted using the PowerMax Soil DNA isolation kit (MOBIO Laboratories) as described previously (Jungbluth *et al*., 2013b). A negative DNA extraction consisting of only kit reagents was processed in parallel to sediment sample extractions (Jungbluth *et al*., 2013b).

SSU rRNA gene PCR amplification and Illumina sequencing

An Illumina sequencing approach (Caporaso *et al*., 2011; Caporaso *et al*., 2012) was used to characterize samples of borehole fluids, sediments, and seawater. Briefly, this approach involves the polymerase chain reaction (PCR)-mediated amplification of the V4 region of the small subunit ribosomal RNA (SSU rRNA) gene using oligonucleotide primers 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACVSGGGTATCTAAT-3') specific to *Bacteria* and *Archaea* and which are modified to include the Illumina flowcell adapter sequences (Bates *et al*., 2011). Reverse primer 806R contained an additional 12-bp barcode, which was used to assign individual sequences to samples. The taxonomic coverage of primer pair 515/806 was assessed using PrimerProspector (Walters *et al*., 2011) with the SILVA SSURef NR99 version 115 database (Pruesse *et al*., 2007) and found to be nearly universal, as described previously (e.g. Bates *et al*., 2011; Walters *et al*., 2011). Each 25 µl PCR reaction was prepared in 5Prime HotMasterMix (Eppendorf-5Prime Inc., Gaithersburg, MD, USA) and contained 0.5 U *Taq* DNA polymerase, 45 mM KCl, 2.5 mM Mg^{2+}, 200

μM of each of the four deoxynucleoside triphosphates (dNTPs), 200 nM of both forward and reverse primer, and 2 μl of genomic DNA template (equivalent to 1-20 ng of environmental genomic DNA). PCR cycling conditions consisted of an initial denaturation step at 94°C for 3 min, followed by 35 cycles of 94°C denaturation for 45 sec, 50°C annealing for 1 min, 72°C extension for 1.5 min, and a final extension step at 72°C for 10 min. Triplicate PCR reactions were pooled at equimolar concentrations and PCR cleanup was performed on the final pooled product using the UltraClean PCR clean up kit (MOBIO Laboratories). Sequencing was performed on an Illumina (San Diego, CA, USA) MiSeq sequencer (Instrument ID: M00517; run number: 41) at the University of Colorado BioFrontiers Institute.

Sequence read-processing and read-pairing

Sequence quality was assessed using FastQC (version 0.10.1) and indicated an overall high-quality run was executed, although, notably, read quality across all bases decreased more rapidly for the reverse sequencing reaction. Figure S2 provides an overview of the bioinformatic workflow used to process the resulting SSU rRNA Illumina sequencing reads. Demultiplexing of sequence data was performed using the QIIME pipeline (Caporaso *et al*., 2010) split_libraries_fastq.py script (version 1.7) with a maximum of 1.5 errors in the barcode, and the following quality-filtering parameters adapted from (Bokulich *et al*., 2013): r=3, p=0.75, q=3, and n=0. Forward reads (515F) were in higher abundance than reverse reads (806R) (Table S3). Sample metadata and the SSU rRNA sequence files used in this study have been submitted to the NCBI

BioSample and Sequence Read Archive databases and can be accessed using the BioProject identifier PRJNA266365.

Read pairing was performed using a variety of tools specialized to handle non-gapped sequencing errors typical of the Illumina sequencing platform and parameter values selected to control for differences in the pairing methods (i.e. base quality within the overlapping region) (Table S3). Where possible, parameter values for the length of overlapping regions were kept low in order to increase the chance of successful pairing following truncation of low quality bases, which were relatively more abundant in the overlapping portion of the reads. All methods utilized default parameters unless otherwise noted. Pairing using USEARCH (version 7.0.959; Edgar, 2010) fastq_mergepairs script used parameters: fastq_truncqual=3, fastq_minovlen=5 and fastq_filter script with fastq_maxee parameters=0.05, 1.0, and 5.0. Pairing using FLASH (version 1.2.7; Magoč and Salzberg, 2011) used parameters: m=5 and x=0.25. Pairing using PANDAseq (version 2.5; Masella *et al*., 2012) used parameters: o=5 and t=0.25. Pairing using merge_illumina_pairs script (Eren *et al*., 2013) used parameters: ignore-Ns, o=5 or 48, min-qual-score=3, and m/o=0.25. Multiple overlap sizes and stringency requirements were used in association with the merge_illumina_pairs script due to high sensitivity of the pairing to the specific length of the overlapping region and the number of mismatches in the overlap. In contrast to the other methods used, the merge_illumina_pairs script performed superiorly when the parameter for length of the overlapping region was near the full overlap length expected (o=48). A final quality check was implemented using the enforce-Q30-check (Minoche *et al*., 2011) in association with the merge_illumina_pairs script due to the quality score stripping

associated with this script. Read-pairing was attempted using four different tools (Table S3; Figure S2), but ultimately used the python script “merge-illumina-pairs” developed by Eren and colleagues (Eren *et al*., 2013) due to the relatively high numbers of reads successfully paired and a percentage of unique sequences retained – an index of read quality - that is consistent with the other methods tested (Table S4).

Paired-reads were demultiplexed using the QIIME pipeline (Caporaso *et al*., 2010) split_libraries_fastq.py script (version 1.7) with a maximum of 1.5 errors in the barcode, and the following quality-filtering parameters: r=10, p=0.75, q=3, and n=0. A range of values for parameter r - maximum number of low bases before read truncation - were tested due to the expected increased sensitivity for this parameter to the longer read lengths, which is due to the relatively low Phred scores in the overlapping portion of the reads. Parameter value r=10 was ultimately selected because it represents a balance between complete sequence retention (all pairs obtained) and no sequences obtained. Sequences generated using the merge_illumina_pairs script (Eren *et al*., 2013) were not processed this way due to the lack of quality scores; instead the sequences passing the enforce-Q30-check (Minoche *et al*., 2011) were used in subsequent analyses.

Sequences longer than 254 bases constituted a minor portion of the paired-reads using all methods except for the merge_illumina_pairs script with length of the overlapping region equal to 5. Queries against the NCBI non-redundant nucleotide database using BLAST (Altschul *et al*., 1990) revealed amplicons larger than 254 base pairs to have relatively few full-length sequence matches. In addition, expected fragment lengths were determined for the reconstructed SILVA SEED databases of

Archaea and *Bacteria* (Schloss, 2009), and a manually curated database of sequences >100 bp obtained from Juan de Fuca Ridge basement basalt (Cowen *et al*., 2003; Huber *et al*., 2006; Jungbluth *et al*., 2013a; Jungbluth *et al*., 2014; this study), sediment (Jungbluth *et al*., 2013b), and bottom seawater (Jungbluth *et al*., 2013a). Due to the relatively few numbers of reads expected and the incorrect read pairing observed, paired sequences longer than 254 bases were excluded from further analysis.

Chimeric sequences were first detected with USEARCH using the UCHIME (Edgar *et al*., 2011) *de novo* chimera function and excluded from further analysis. This method was followed by UCHIME chimera checking against the SILVA SEED reference database supplemented with the Juan de Fuca sequence database described above. Chimera-checked forward and paired reads were both used in further analyses.

OTU clustering – UCLUST/UPARSE

The QIIME open reference picking script (pick_open_reference_otus.py) was used to perform uclustref (Edgar, 2010) and *de novo* clustering using the Greengenes v13.5 and SILVA v111 databases clustered at 97% or 99% OTU similarity. Clustering performed using the USEARCH package involved sequence de-replication, abundance sorting, and clustering using UPARSE (Edgar, 2013). UPARSE clustering with reverse reads required 64-bit USEARCH (version 7.0.1090). Re-screening for chimeric OTUs was disabled using the command –parse_break -999 because chimera screening was already performed. Following clustering, reads were mapped onto OTU clusters using the –usearch_global script and a 97% sequence identity threshold.

OTU clustering – Average linkage

Average linkage clustering using mothur (Schloss *et al*., 2009) was also used (Kozich *et al*., 2013). Briefly, reads were aligned against the SILVA SEED database supplemented with the Juan de Fuca sequence database described above, given temporary (i.e. for clustering only) taxonomic assignment using the SILVA SEED *Bacteria* and *Archaea* databases combined, and finally clustered using the cluster.split command with taxonomic splitting at the Order level using a hard 0.1 genetic distance cutoff and the average-neighbor clustering algorithm.

OTU clustering – Distribution-based

Distribution-based clustering (Preheim *et al*., 2013) was also performed, using scripts associated with the software and in similar fashion to the protocol listed in the associated manual (https://github.com/spacocha/Distribution-based-clustering). USEARCH progressive clustering beginning at 98% sequence similarity level and iterating each single percentage down to 90% was used to generate a full fasta and sequence by library matrix to be used for downstream analysis. Sequences were aligned with mothur (Schloss *et al*., 2009) using the SILVA SEED database supplemented with the Juan de Fuca sequence database described above, and distance matrices were generated using FastTree (Price *et al*., 2010) with default options. Clustering was performed in parallel using the distribution_clustering.pl script with parameters: a=0, p=0.0005, and a range of distance parameters (d=0.1, 0.05, 0.03, 0.01) to assess differential OTU clustering that will result. Evaluation of the clustering results using the evaluate_parallel_results.pl script was not attempted; results were

evaluated using custom shell scripts to check for the criteria described in evaluate_parallel_results.pl. Final files were produced using the scripts included within clean_up_parallel_ultra.csh, which included custom perl and shell scripts and additional sequence alignment using mothur (Schloss *et al*., 2009) and the SILVA SEED database supplemented with the Juan de Fuca sequence database described above.

OTU clustering – Unique sequence clusters

All paired and unpaired reads were also analyzed using 100% cluster identities (Tikhonov *et al*., 2015). Unique sequence community matrices and fasta files were generated using perl scripts fasta2unique_table4.pl and OTU2lib_count_trans1.pl associated with the distribution-based clustering pipeline (Preheim *et al*., 2013).

Taxonomic assignment

Paired unique reads from the merge-illumina-pairs script (Eren *et al*., 2013) were selected for classification and further analysis due to the longer sequence lengths and the relatively high numbers of reads retained post processing. Sequences were aligned and taxonomy was assigned via the SINA aligner v1.2.11 (Pruesse *et al*., 2012) using the non-redundant SSURef_115 database pre-clustered with UCLUST at a 99% sequence similarity. Parameters used with the SINA tool included –lca-fields tax_slv to assign SILVA taxonomy and –search-min-sim 0.80 to expand searches for difficult-to-classify (i.e. divergent) sequences. Phylum names *Marinimicrobia* (SAR406), *Aminicenantes* (OP8), and *Aerophobetes* (BHI80-139) were adapted from Rinke *et al.*

(2013), while *Bathyarchaeota* (Miscellaneous Crenarcheotal Group; MCG) is from Meng *et al.* (2014).

Bathyarchaeota sequences identified by SILVA were further classified using the PhyloAssigner tool (Vergin *et al*., 2013), which includes pplacer (Matsen *et al*., 2010) and PhyML (Guindon *et al*., 2010), and used a manually curated *Bathyarchaeota* database that follows naming schemes described in Kubo *et al.* (2012). The *Bathyarchaeota* database was generated from sequences within the SILVA SSU Ref 99 version 115 base tree. The final tree was composed of 3160 sequences and included: (1) archaeal sequences that were visually selected to encompass all of the major archaeal lineages identified in the SILVA v115 database, (2) all non-chimeric unclassified archaeal sequences, (3) all sequences from groups *Bathyarchaeota*, C3, and THSCG, and (4) *Chloroflexus aggregans* (CP001337), *Vibrio vulnificus* (X76333), and *Thermotoga maritime* (M21774) as outgroups. Groups within the *Bathyarchaeota* were classified by transferring group names from the PhyML tree generated by Kubo and colleagues (2012) whenever monophyletic lineages were consistent between their analysis and the base tree generated in the present study. Two unique lineages of *Bathyarchaeota* that were recently recovered from Juan de Fuca ridge flank fluids (Jungbluth *et al*., 2013a) were added to the *Bathyarchaeota* base tree via the parsimony insertion tools in ARB (1301A08_240 and 1301A09_032 in Figure 6B). Two pairs previously recovered as discrete groups by Kubo and colleagues (2012) appear polyphyletic in the present analysis (MCG-11 and MCG-12, and MCG-15 and Group C3; Figure 6B). Results were plotted using R package gplots with the heatmap.2 and hclust

functions (Warnes *et al.* 2015). The *Bathyarchaeota* database used here is available on request.

Statistical analysis of sample groupings

Statistical analyses were performed using unpaired forward reads rarefied to an even sampling depth across all samples (n=6108 reads). Dissimilarity among community matrices and associated chemical metadata were explored using Mantel tests with 1000 replications performed in QIIME (version 1.8) using the scripts distance_matrix_from_mapping.py and compare_distance_matrices.py. Comparisons by sample type (seawater, sediment, subsurface fluids [SSF16 removed]) were also performed by using the compare_categories.py script in QIIME. Statistical tests included ANOSIM implemented through QIIME, and adonis, MRPP, PERMDISP, PERMANOVA, and db-RDA (with mixed and controls included) implemented through QIIME using the R package vegan (Oksanen *et al*., 2013). All analyses were based on 1000 permutations.

SSU rRNA gene sequencing of flow sorted single cells

A 1 ml cryopreserved aliquot of borehole U1362A fluid sample SSF19 was sent for single cell isolation and identification at the Bigelow Laboratory for Ocean Sciences Single Cell Genomics Center (http://scgc.bigelow.org) using fluorescent activated cell sorting in a 384-well plate. Briefly, cells were lysed, the genomic DNA was amplified using multiple displacement amplification as described previously (Swan *et al*., 2011), and the single cell whole genome amplified DNA was screened with bacterial (27F: 5'-

AGRGTTYGATYMTGGCTCAG-3'/907R_degen 5'-CCGTCAATTCMTTTRAGTTT-3') and archaeal (Arc_344F:5'-ACGGGGYGCAGCAGGCGCGA-3'/Arc_915R:5'-GTGCTCCCCCGCCAATTCCT-3') SSU rRNA gene primers. Paired-end Sanger sequences resulting from the sorted and genome-amplified single cells were assembled using default parameters in the Sequencher version 5.1 software package (Gene Codes Corp., Ann Arbor, MI, USA). A total of 74 paired reads and 9 unpaired reads representing the longer of the read pairs (unless quality score was less than 50%) were aligned and given taxonomic identifications using the SINA online aligner and SILVA classifier tool (version 1.2.11; Pruesse *et al*., 2012) using minimum identity with query sequence: 0.6, number of neighbors per query sequence: 1, reject sequences below identity: 50%, and default parameters. Manual inspection of sequences using BLAST queries (Altschul *et al*., 1990) revealed non-rRNA genes (n=9), which were removed from analysis.

SSU rRNA gene cloning and sequencing

Small subunit rRNA genes were amplified from nucleic acids extracted from selected environmental samples (SSF11, SSF12, SSF21-22, SSF23-24) using universal primer pair 519F (5'-CAGCMGCCGCGGTAATWC-3') and 1406R (5'-ACGGGCGGTGTGTRC-3') (Lane *et al*., 1985). The PCR amplification, cloning, and Sanger sequencing have been described previously (Jungbluth *et al*., 2013a). Taxonomic identifications using SINA/SILVA were performed identically as described above except the parameter "reject sequences below identity" was changed to 70%. All

non-redundant clone SSU rRNA gene sequences generated in this study have been deposited in GenBank under accession numbers KR072702-KR072893.

Phylogenetic analysis

Ribosomal RNA gene sequences resulting from the amplified single cells and clone libraries were manually curated using Sequencher software (Gene Codes Corp.). Curated sequences were first aligned using the online SINA tool version 1.2.11 (Pruesse *et al*., 2012) before importing into the ARB software package (Ludwig *et al*., 2004), where the multiple species alignment was manually curated and sequences classified taxonomically using version SSURef_115 of the SILVA ARB database clustered to a 99% level of similarity (Pruesse *et al*., 2007). Additional sequences that were highly similar to the SSU rRNA gene sequences obtained in this study were identified by BLAST search against the non-redundant nucleotide database (Altschul *et al*., 1990), and added to the ARB database. The nucleotide substitution model that best fit the near-full length sequence (>1200 nucleotide) alignment was determined using jModelTest version 2.1.1 (Darriba *et al*., 2012). Phylogenetic analyses were performed using near-full length sequences (>1200 nucleotides) with the RAxML maximum likelihood method using the GTR model of nucleotide substitution under the gamma- and invariable- models of rate heterogeneity (Stamatakis *et al*., 2006). The tree with the highest log likelihood score was selected from performing 100 iterations of the RAxML method. Sequences of short length were added to the maximum likelihood-derived phylogeny using the parsimony insertion tool in ARB in the following order: (i) cloned and single cell amplicon SSU rRNA genes, (ii) sequences derived from Genbank, and

(iii) Illumina tag sequences. Bootstrap analysis of the near-full length sequence alignment (i.e. prior to addition of sequences via parsimony) was determined by RAxML using the rapid boostrap analysis algorithm (1000 bootstraps) implemented within ARB (Stamatakis *et al*., 2008). The tree was visualized using iTOL (Letunic and Bork, 2007; Letunic and Bork, 2011).

Sample preparation for microscopy and fluorescence microscopy

Fluid samples for microscopy collected in 2011 were prepared in similar fashion to those collected in sampling years 2008-2010 and described previously (Jungbluth *et al*., 2013a). Briefly, 40 to 120 ml sub-samples were fixed with a final concentration of 3% of 0.2 µm-filtered formaldehyde for 2 to 4 hours at 4ºC, and subsequently filtered through 0.2 µm pore-sized polycarbonate membranes (Whatman, Maidstone, United Kingdom). After air-drying, membranes were stored desiccated at -80ºC until microscopic analysis.

Filter sections were prepared for fluorescence microscopy using a mix of Citifluor/VectaShield/PBS/DAPI as described previously (Jungbluth *et al*., 2013a). Stained filter sections were inspected with a Leica DM5000B epifluorescence microscope (Leica Microsystems, Wetzlar, Germany) (samples: SSF1-2, SSF4, MIX1-4, SW1-5, SW9-11, SW14-15) or an Eclipse 90i (Nikon Corp., Tokyo, Japan) epifluorescence microscope (all other samples). Both microscopes were equipped with 100x objectives and filter sets appropriate for DAPI fluorescence.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-410. doi: 10.1016/S0022-2836(05)80360-2

Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N (2011). Examining the global distribution of dominant archaeal populations in soil. *ISME J* **5:** 908-917. doi: 10.1038/ismej.2010.171

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R *et al.* (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* **10:** 57-59. doi: 10.1038/nmeth.2276

Brewer P, Spencer D (1971). Colorimetric determination of manganese in anoxic waters. *Limnol Oceanogr* **16:** 107-110. doi: 10.4319/lo.1971.16.1.0107

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335-336. doi: 10.1038/nmeth.f.303

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6:** 1621-1624. doi: 10.1038/ismej.2012.8

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108:** 4516-4522. doi: 10.1073/pnas.1000080107

Cowen JP, Copson DA, Jolly J, Hsieh C-C, Lin H-T, Glazer BT *et al.* (2012). Advanced instrument system for real-time and time-series microbial geochemical sampling

of the deep (basaltic) crustal biosphere. *Deep-Sea Res Pt I* **61:** 43-56. doi: 10.1016/j.dsr.2011.11.004

Cowen JP, Giovannoni SJ, Kenig F, Johnson HP, Butterfield D, Rappé MS *et al.* (2003). Fluids from aging ocean crust that support microbial life. *Science* **299:** 120-123. doi: 10.1126/science.1075653

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9:** 772. doi: 10.1038/nmeth.2109

Dickson AG, Sabine CL, Christian JR (eds) (2007). *Guide to best practices for ocean CO_2 measurements*, 191pp. (URL: http://cdiac.ornl.gov/oceans/Handbook_2007.html)

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461. doi: 10.1093/bioinformatics/btq461

Edgar RC (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10:** 996-998. doi: 10.1038/nmeth.2604

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194-2200. doi: 10.1093/bioinformatics/btr381

Eren AM, Vineis JH, Morrison HG, Sogin ML (2013). A filtering method to generate high quality short reads using Illumina paired-end technology. *PloS One* **8:** e66643. doi: 10.1371/journal.pone.0066643

Expedition 327 Scientists (2011a). Expedition 327 summary. In: Fisher AT, Tsuji T, Petronotis K, Expedition 327 Scientists (eds). *Proceedings of the Integrated*

Ocean Drilling Program. Integrated Ocean Drilling Program Management International, Inc.: College Station, TX. pp 1-32. doi: 10.2204/iodp.proc.327.101.2011

Expedition 327 Scientists (2011c). Methods. In: Fisher AT, Tsuji T, Petronotis K, Expedition 327 Scientists (eds). *Proceedings of the Integrated Ocean Drilling Program*. Integrated Ocean Drilling Program Management International, Inc.: College Station, TX. pp 1-42. doi: 10.2204/iodp.proc.327.102.2011

Expedition 327 Scientists (2011b). Site U1363. In: Fisher AT, Tsuji T, Petronotis K, Expedition 327 Scientists (eds). *Proceedings of the Integrated Ocean Drilling Program*. Integrated Ocean Drilling Program Management International, Inc.: College Station, TX. pp 1-32. doi: 10.2204/iodp.proc.327.106.2011

Gibbs C (1976). Characterization and application of ferrozine iron reagent as a ferrous iron indicator. *Anal Chem* **48:** 1197-1201. doi: 10.1021/ac50002a034

Grasshoff K, Kremling K, Ehrhardt M (eds) (1999) *Methods of seawater analysis*. Wiley: Weinheim. doi: 10.1002/9783527613984

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:** 307-321. doi: 10.1093/sysbio/syq010

Huber JA, Johnson HP, Butterfield DA, Baross JA (2006). Microbial life in ridge flank crustal fluids. *Environ Microbiol* **8:** 88-99. doi: 10.1111/j.1462-2920.2005.00872.x

Jones RD (1991). An improved fluorescence method for the determination of nanomolar concentrations of ammonium in natural-waters. *Limnol Oceanogr* **36:** 814-819. doi: 10.4319/lo.1991.36.4.0814

Jungbluth SP, Grote J, Lin H-T, Cowen JP, Rappé MS (2013a). Microbial diversity within basement fluids of the sediment-buried Juan de Fuca Ridge flank. *ISME J* **7:** 161-172. doi: 10.1038/ismej.2012.73

Jungbluth SP, Johnson LGH, Cowen JP, Rappé MS (2013b). Data report: microbial diversity in sediment near Grizzly Bare Seamount from Holes U1363B and U1363G. In: Fisher AT, Tsuji T, Petronotis K, Expedition 327 Scientists (eds). *Proceedings of the Integrated Ocean Drilling Program*. Integrated Ocean Drilling Program Management International, Inc.: Tokyo. doi: 10.2204/iodp.proc.327.201.2013

Jungbluth SP, Lin H-T, Cowen JP, Glazer BT, Rappé MS (2014). Phylogenetic diversity of microorganisms in subseafloor crustal fluids from boreholes 1025C and 1026B along the Juan de Fuca Ridge flank. *Front Microbiol* **5:** 119. doi: 10.2289/fmicb.2014.00119

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79:** 5112-5120. doi: 10.1128/AEM.01043-13

Kubo K, Lloyd KG, F. BJ, Amann R, Teske A, Knittel K (2012). Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in

marine sediments. *ISME J* **6:** 1949-1965. doi: 10.1038/ismej.2012.37

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82:** 6955-6959. doi: 10.1073/pnas.82.20.6955

Letunic I, Bork P (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23:** 127-128. doi: 10.1093/bioinformatics/btl529

Letunic I, Bork P (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39:** W475-478. doi: 10.1093/nar/gkr201

Lin H-T, Cowen JP, Olson EJ, Amend JP, Lilley MD (2012). Inorganic chemistry, gas compositions and dissolved organic carbon in fluids from sedimented young basaltic crust on the Juan de Fuca Ridge flanks. *Geochim Cosmochim Ac* **85:** 213-227. doi: 10.1016/j.gca.2012.02.017

Lin H-T, Hsieh C-C, Cowen JP, Rappé MS (2015). Data report: dissolved and particulate organic carbon in the deep sediments of Site U1363 near Grizzly Bare seamount. In: Fisher AT, Tsuji T, Petronotis K, Expedition 327 Scientists (eds). *Proceedings of the Integrated Ocean Drilling Program*. Integrated Ocean Drilling Program Management International, Inc.: Tokyo. doi: 10.2204/iodp.proc.327.202.2015

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32:** 1363-1371. doi:

10.1093/nar/gkh293

Magoč T, Salzberg SL (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27:** 2957-2963. doi: 10.1093/bioinformatics/btr507

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012). PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics* **13:** 31. doi: 10.1186/1471-2105-13-31

Matsen FA, Kodner RB, Armbrust EV (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11:** 538. doi: 10.1186/1471-2105-11-538

Meng J, Xu J, Qin D, He Y, Xiao X, Wang F (2014). Genetic and functional properties of uncultivated MCG Archaea assessed by metagenome and gene expression analysis. *ISME J* **8:** 650-659. doi: 10.1038/ismej.2013.174

Minoche AE, Dohm JC, Himmelbauer H (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* **12:** R112. doi: 10.1186/gb-2011-12-11-r112

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB *et al.* (2013). vegan: Community Ecology Package. (URL: http://cran.r-project.org/web/packages/vegan/index.html)

Phillips BM, Anderson BS, Hunt JW (1997). Measurement and distribution of interstitial and overlying water ammonia and hydrogen sulfide in sediment toxicity tests. *Mar Environ Res* **44:** 117-126. doi: 10.1016/S0141-1136(96)00087-6

Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79:** 6593-6603. doi: 10.1128/AEM.00342-13

Price MN, Dehal PS, Arkin AP (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5:** e9490. doi: 10.1371/journal.pone.0009490

Pruesse E, Peplies J, Glöckner FO (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28:** 1823-1829. doi: 10.1093/bioinformatics/bts252

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35:** 7188-7196. doi: 10.1093/nar/gkm864

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437. doi: 10.1038/nature12352

Schloss PD (2009). A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS One* **4:** e8230. doi: 10.1371/journal.pone.0008230

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537-7541. doi: 10.1128/AEM.01541-09

Sharp JH, Carlson CA, Peltzer ET, Castle-Ward DM, Savidge KB, Rinker KR (2002a). Final dissolved organic carbon broad community intercalibration and preliminary use of DOC reference materials. *Mar Chem* **77:** 239-253. doi: 10.1016/S0304-4203(02)00002-6

Sharp JH, Rinker KR, Savidge KB, Abell J, Yves Benaim J, Bronk DA *et al.* (2002b). A preliminary methods comparison for measurement of dissolved organic nitrogen in seawater. *Mar Chem* **78:** 171-184. doi: 10.1016/S0304-4203(02)00020-8

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690. doi: 10.1093/bioinformatics/btl446

Stamatakis A, Hoover P, Rougemont J (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* **57:** 758-771. doi: 10.1080/10635150802429642

Stookey LL (1970). Ferrozine- a new spectrophotometric reagent for iron. *Anal Chem* **42:** 779-781. doi: 10.1021/ac60289a016

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D *et al.* (2011). Potential for chemolithoautrotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333:** 1296-1300. doi: 10.1126/science.1203690

Tikhonov M, Leach RW, Wingreen NS (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* **9:** 68-80. doi: 10.1038/ismej.2014.117

Vergin KL, Beszteri B, Monier A, Thrash JC, Temperton B, Treusch AH *et al.* (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series

Study site by phylogenetic placement of pyrosequences. *ISME J* **7:** 1322-1332. doi: 10.1038/ismej.2013.32

Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R (2011). PrimerProspector: de novo design and taxonomic analysis of barcoded PCR primers. *Bioinformatics*. doi: 10.1093/bioinformatics/btr087

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T *et al.* (2015). gplots: Various R Programming Tools for Plotting Data. (URL: http://cran.r-project.org/package=gplots)

Wheat CG, Jannasch HW, Fisher AT, Becker K, Sharkey J, Hulme S (2010). Subseafloor seawater-basalt-microbe reactions: continuous sampling of borehole fluids in a ridge flank environment. *Geochem Geophy Geosys* **11**. doi: 10.1029/2010gc003057

Wheat CG, Jannasch HW, Kastner M, Plant JN, DeCarlo EH, Lebon G (2004). Venting formation fluids from deep-sea boreholes in a ridge flank setting: ODP Sites 1025 and 1026. *Geochem Geophy Geosys* **5**. doi: 10.1029/2004GC000710

Figure S1. Borehole locations and characteristics of CORKs used to access basalt-hosted deep subseafloor crustal fluids along the Juan de Fuca Ridge flank. (A) Map of CORK observatory sampling sites on the Juan de Fuca Ridge flank in the Northeast Pacific Ocean. (B) Three-dimensional view of basement relief at boreholes 1026B, U1301A, U1362A, and U1362B (Expedition 327 Scientists, 2011a). (C) Schematic diagrams showing multiple generations of CORK observatory installed in boreholes on the Juan de Fuca Ridge flank. Borehole 1025C contains a first generation CORK, boreholes 1026B and U1301A contain second-generation CORKs, and boreholes U1362A and U1362B contain third-generation CORKs. The fluid sampling lines used in this study are indicated with dashed purple colored lines (Expedition 327 Scientists, 2011a).

Figure S2. Bioinformatic workflow used to process SSU rRNA gene Illumina tag sequencing reads. Forward, reverse, and paired reads were chimera checked prior to clustering and taxonomic classification with SILVA. Dashed lines highlight the schemes used to process unpaired and paired reads that were used for primary (non-experimental) α- and β-diversity analyses.

Figure S3. Box-plot diagram representing 1000 replications of α-diversity metrics. Colors of the bars correspond to the different clustering methods or unique read analysis, while percentages listed in the legend refer to the OTU similarity cutoff. DBC, distribution-based clustering; d, maximum genetic distance used to assess clusters; gg, greengenes; slv, silva.

Figure S4. Procrustes PCoA biplot of forward unique read clusters and (A) reverse unique read clusters, (B) paired unique read clusters [merge-paired-reads; Eren et al. 2013], (C) forward read clusters [UCLUSTref with SILVA 99% OTU database], (D) forward read clusters [UPARSE], (E) forward read clusters [mothur; 97% OTUs], (F) forward read clusters [distribution-based clustering; d=0.1]. Green, subsurface fluids; yellow/brown, sediments; purple/blue, seawater and low-quality subsurface fluids. Clustering differences between samples are indicated using a single line with equal length parts red and white.

Figure S5. (A) Logarithmic best-fit model applied to scatterplot of all subsurface, mixed, and seawater samples with corresponding cell abundance and magnesium concentration data. (B) Linear best-fit model applied to scatterplot of all high-integrity subsurface samples with corresponding cell abundance and magnesium concentration data.

Figure S6. (A) Jackknifed β-diversity analysis of Bray-Curtis dissimiliarity indices using unique unpaired forward Illumina tag reads and rarefying to an even sequence depth (100 reads) across samples. Cluster stability is indicated using semi-transparent ovals encompassing opaque-colored sample midpoints. (B) Venn diagram of unique unpaired Illumina tag forward reads showing overlap in microbial communities from seawater, sediment, and subsurface fluids.

Figure S7. Taxonomic diversity and abundance of SSU rRNA gene clones from select samples of U1301A (SSF11, SSF12), U1362A (SSF21-22), and U1362B (SSF23-24) borehole fluids.

Table S1. Summary of samples used in this study

Sample ID	*Sample type*	*Location*[a]	*Sample depth*[a]	*Sample date*	*Collection method*	*Biological replicates*[b]	*Technical replicates*[c]	*Vol filtered (Liters)*	*Sanger sequence*	*Cells ml^{-1} ($x\ 10^3$)*
SSF1	basalt-hosted	U1301A	8-107 msb	30-Aug-09	LVWS		A	2.6	Y[c]	9.0[c]
SSF2	basalt-hosted	U1301A	8-107 msb	30-Aug-09	LVWS		A	2.5		9.0[c]
SSF3	basalt-hosted	U1301A	8-107 msb	19-Jun-10	LVWS	A		13.2		7.4
SSF4	basalt-hosted	U1301A	8-107 msb	19-Jun-10	LVWS	A		11.2	Y[c]	15.3[c]
SSF5	basalt-hosted	U1301A	8-107 msb	20-Jun-10	LVWS	B		2.4		16.1
SSF6	basalt-hosted	U1301A	8-107 msb	20-Jun-10	LVWS	B		5.3		4.0
SSF7	basalt-hosted	U1301A	8-107 msb	23-Jun-10	LVWS	C		14.9		50.7
SSF8	basalt-hosted	U1301A	8-107 msb	23-Jun-10	LVWS	C		3.2		30.5
SSF9	basalt-hosted	1025C	0-46 msb	27-Jun-10	LVWS			2.0	Y[e]	n.d.
SSF10	basalt-hosted	U1301A	8-107 msb	30-Jun-10	LVWS	D		21.1		11.2
SSF11	basalt-hosted	U1301A	8-107 msb	30-Jun-10	LVWS	D		29.2	Y	15.4
SSF12	basalt-hosted	U1301A	8-107 msb	30-Jun-10	*in situ*	D		~120	Y	n.d.
SSF13	basalt-hosted	U1301A	8-107 msb	04-Jul-11	LVWS	E		0.9		n.d.
SSF14	basalt-hosted	U1301A	8-107 msb	04-Jul-11	LVWS	E		1.3		n.d.
SSF15	basalt-hosted	U1301A	8-107 msb	04-Jul-11	MVBS	E		14.1		5.8
SSF16	basalt-hosted	U1362B	29-117 msb	08-Jul-11	LVWS	F		9.4		21.8
SSF17	basalt-hosted	U1362B	29-117 msb	08-Jul-11	LVWS	F		24.7		5.6
SSF18	basalt-hosted	U1362B	29-117 msb	10-Jul-11	MVBS	G		14.4		2.6
SSF23	basalt-hosted	U1362B	29-117 msb	10-Jul-11	*in situ*	G	B	~70	Y	n.d.
SSF24	basalt-hosted	U1362B	29-117 msb	10-Jul-11	*in situ*	G	B	~70	Y	n.d.
SSF19	basalt-hosted	U1362A	193-292 msb	12-Jul-11	MVBS	H		13.8		26.1
SSF20	basalt-hosted	U1362A	193-292 msb	12-Jul-11	MVBS	H		15.7		15.6
SSF21	basalt-hosted	U1362A	193-292 msb	12-Jul-11	*in situ*	H	C	~124	Y	n.d.
SSF22	basalt-hosted	U1362A	193-292 msb	12-Jul-11	*in situ*	H	C	~124	Y	n.d.
MIX1	mixed	U1301A	8-107 msb	07-Aug-08	GeoMicrobe			10.0		67.6
MIX2	mixed	U1301A	8-107 msb	07-Aug-08	GeoMicrobe			10.0		54.3
MIX3	mixed	U1301A	8-107 msb	07-Aug-08	LVWS	I		3.0		99.6
MIX4	mixed	U1301A	8-107 msb	07-Aug-08	LVWS	I		2.8		99.6
MIX5	mixed	1026B	0-48 msb	31-Aug-09	LVWS	J		4.9		41.2
MIX6	mixed	1026B	0-48 msb	31-Aug-09	LVWS	J		4.8		41.2
MIX7	mixed	U1301B	208-318 msb	03-Sep-09	LVWS	K		5.0		36.8
MIX8	mixed	U1301B	208-318 msb	03-Sep-09	LVWS	K		5.0		36.8
MIX9	mixed	U1301A	8-107 msb	04-Sep-09	LVWS	L		4.2		63.8
MIX10	mixed	U1301A	8-107 msb	04-Sep-09	LVWS	L		4.0		63.8
MIX11	mixed	U1301A	8-107 msb	04-Sep-09	LVWS	L		4.3		63.8
SW1	seawater	near U1301A	2644 m	08-Aug-08	CTD Niskin			2.6	Y[c]	87.8[c]
SW2	seawater	near U1301A	2515 m	08-Aug-08	CTD Niskin			3.9		72.2
SW3	seawater	near U1301A	~2650 m	04-Sep-09	CTD Niskin	M		3.8	Y[c]	95.0[c]
SW4	seawater	near U1301A	~2650 m	04-Sep-09	CTD Niskin	M		3.6		95.0[c]
SW5	seawater	near U1301A	~2650 m	04-Sep-09	CTD Niskin	M		3.7		95.0[c]
SW6	seawater	near 1025C	2571 m	24-Jun-10	LVWS	N		4.0		22.0

SW7	seawater	near 1025C	2571 m	24-Jun-10	LVWS	N	4.3		22.0
SW8	seawater	near 1025C	2571 m	24-Jun-10	LVWS	N	9.0		22.0
SW9	seawater	near U1301A	2648 m	28-Jun-10	CTD Niskin	O	5.0	Y[c]	76.0[c]
SW10	seawater	near U1301A	2648 m	28-Jun-10	CTD Niskin	O	5.0		76.0[c]
SW11	seawater	near U1301A	2648 m	28-Jun-10	CTD Niskin	O	5.0		76.0[c]
SW12	seawater	near U1301A	2575 m	28-Jun-10	CTD Niskin	P	5.0		84.2
SW13	seawater	near U1301A	2575 m	28-Jun-10	CTD Niskin	P	5.0		84.2
SW14	seawater	near U1301A	~2655 m	29-Jun-10	Jason Niskin		4.9	Y[c]	88.6[c]
SW15	seawater	near U1301A	~2655 m	30-Jun-10	Jason Niskin		4.9		88.6[c]
SW16	seawater	near U1301A	2661 m	04-Jul-11	Jason Niskin		5.2		120.0
SW17	seawater	near U1301A	2661 m	07-Jul-11	CTD Niskin	Q	3.9		120.0
SW18	seawater	near U1301A	2661 m	07-Jul-11	CTD Niskin	Q	6.1		120.0
SW19	seawater	near U1301A	2500 m	07-Jul-11	CTD Niskin	R	4.5		61.8
SW20	seawater	near U1301A	2500 m	07-Jul-11	CTD Niskin	R	5.3		61.8
SD1	sediment	U1363F	~32 mbsf	03-Sep-10	HPC		--		n.d.
SD2	sediment	U1363B	~38 mbsf	31-Aug-10	XCB		--	Y[d]	n.d.
SD3	sediment	U1363F	~10 mbsf	03-Sep-10	HPC		--		n.d.
SD4	sediment	U1363G	~2 mbsf	04-Sep-10	HPC		--	Y[d]	n.d.
SD5	sediment	U1363G	~18 mbsf	04-Sep-10	HPC		--	Y[d]	n.d.
SD6	sediment	U1363B	~2 mbsf	31-Aug-10	HPC		--	Y[d]	n.d.
SD7	sediment	U1363D	~225 mbsf	02-Sep-10	XCB		--		n.d.
SD8	sediment	U1363B	~45 mbsf	31-Aug-10	XCB		--		n.d.
SD9	sediment	U1363B	~1 mbsf	31-Aug-10	HPC		--	Y[d]	n.d.
SD10	sediment	U1363F	~25 mbsf	03-Sep-10	HPC		--		n.d.
SD11	sediment	U1363D	~205 mbsf	02-Sep-10	XCB		--		n.d.
C1	control	shipboard	--	25-Aug-09	distilled water system		1.9		n.d.
C2	control	--	--	--	--		--	Y[d]	--

[a]For seawater samples, depth is meters below the sea surface; msb - meters subbasement, mbsf - meters below seafloor, samples collected from a single borehole during a single dive are considered biological replicates.

[b]'SSF1' and 'SSF2' were filtered from the same sampling bag onto separate membranes. Two amplicon libraries each were generated from U1362A and U1362B large volume *in situ* filtered samples.

[c]Jungbluth *et al.*, 2013a

[d]Jungbluth *et al.*, 2013b

[e]Jungbluth *et al.*, 2014

--, not applicable; n.d. not determined

LVWS, large volume water sampler; MVBS, medium volume bag sampler; *in situ*, *in situ* filtration; GeoMicrobe, GeoMicrobe sled *in situ* filtration; CTDNiskin, CTD-Niskin rosette; JasonNiskon, ROV *Jason II* Niskin; HPC, hydraulic piston corer; XCB, extended piston corer

Table S2. Summary of basement fluid and seawater biogeochemistry and cellular abundances.

Sequence Sample IDs	*Location*	*pH*	*Ca^{2+} (mM)*	*Mg^{2+} (mM)*	*K^{+} (mM)*	*Na^{+} (mM)*	*Cl^{-} (mM)*	*Br^{-} (mM)*	*SiO_2 (µM)*	*NH_4^{+} (µM)*	*PO_4^{2-} (µM)*	*NO_2^{-} (µM)*	*NO_3^{-} (µM)*	*SO_4^{2-} (mM)*	*Fe_{aq} (µM)*	*Mn^{2+} (µM)*	*Dissolved H_2S*	*DOC (µM)*	*TDN (µM)*	*Alkalinity (meq/L)*	*Basaltic fluid content-based on [NO3]*	*Basaltic fluid content-based on [Ca]*	*Cells ml^{-1} (x 10^3)*
SSF1, SSF2	U1301A	7.4	54.1	2.0	6.0	462.0	547	0.83	1150	102.0	0.10	0.00	0.61	17.2	1.1	4.1	0.17	13	102	0.42	98%	97%	9.0[a]
SSF3	U1301A	7.6	54.3	2.7	7.0	470.8	558	0.88	1154	102.0	0.10	0.00	0.90	18.2	0.7	0.8	0.00	9	104	0.48	98%	97%	7.4
SSF4	U1301A	7.4	54.0	2.7	6.4	469.6	552	0.88	1153	102.0	0.10	0.00	0.90	18.0	0.5	0.8	0.00	15	102	0.48	98%	96%	15.3[a]
SSF5	U1301A	7.5	49.6	5.6	6.2	450.6	530	0.84	1051	91.0	0.22	0.00	2.69	17.9	0.5	0.8	0.00	26	102	0.57	93%	87%	16.1
SSF6	U1301A	7.4	53.2	3.7	6.4	468.8	553	0.89	1126	99.0	0.20	0.00	1.76	18.3	0.5	1.5	0.00	14	100	0.53	96%	95%	4.0
SSF7	U1301A	7.5	51.9	6.9	6.6	468.1	554	0.88	1060	92.0	0.43	0.06	3.18	19.0	0.2	4.1	0.00	12	103	0.65	92%	92%	50.7
SSF8	U1301A	7.5	51.9	5.7	6.5	463.6	546	0.88	1088	95.0	0.48	0.06	2.91	18.5	0.3	3.6	0.00	16	100	0.62	93%	92%	30.5
SSF9	1025C	7.9	30.4	28.9	9.4	468.7	539	0.85	590	43.0	0.05	0.00	6.40	26.2	0.1	0.8	0.00	22	49	0.88	--	84%	n.d.
SSF10	U1301A	7.3	51.9	4.0	6.3	474.8	566	0.91	1124	100.0	0.15	0.00	2.10	18.8	1.0	1.9	0.00	12	99	0.55	95%	92%	11.2
SSF11	U1301A	7.3	51.9	3.9	6.0	469.1	557	0.89	1128	99.0	0.16	0.00	2.00	18.4	0.5	1.4	0.00	10	101	0.56	95%	92%	15.4
SSF15	U1301A	7.4[b]	55.3	1.9	6.4	459.0	551	0.91	1149	98.0	0.09	0.00	0.08	18.1	0.8	n.d.	n.d.	13	104	0.43	100%	99%	5.8
SSF16	U1362B	7.3	53.3	4.0	6.6[c]	462.4[c]	548	0.84	1093	95.0[c]	0.20[c]	0.00	2.00[c]	19.1	1.4	n.d.	n.d.	11	100	0.60	--	95%	21.8
SSF17	U1362B	7.3[b]	55.4	2.2	6.4	462.6	549	0.88	1144	100.0	0.06	0.00	0.03[b]	18.8	1.7	n.d.	n.d.	12	105	0.51	--	100%	5.6
SSF18	U1362B	7.3[b]	55.4	2.5	6.4[b]	462.4[b]	547	0.85	1144	100.0[b]	0.06[b]	0.00	0.03[b]	18.6	1.3	n.d.	n.d.	12	104[b]	0.48	--	100%	2.6
SSF19	U1362A	7.5[b]	53.7	2.2	6.5	460.4	548	0.87	1071	98.0	0.12	0.00	0.03	18.8	2.4	n.d.	n.d.	15	102	0.59	100%	99%	26.1
SSF20	U1362A	7.5[b]	54.3	2.8	6.5[b]	461.4[b]	544	0.81	1067	98.0	0.11[b]	0.00	0.03[b]	18.7	1.9[b]	n.d.	n.d.	16	103[b]	0.60	--	100%	15.6
MIX3, MIX4	U1301A	7.6	15.4	47.4	9.8	455.6	537	0.78	284	16.0	2.53	0.00	35.40	26.9	0.0	n.d.	n.d.	n.d.	45	2.27	10%	11%	99.6
MIX5, MIX6	1026B	7.7	10.2	54.6	10.0	461.7	538	0.78	177	0.1	2.74	0.00	41.40	27.8	0.3	n.d.	n.d.	40	43	2.39	0%	0%	41.2
MIX7, MIX8	U1301B	7.6	10.5	52.4	10.1	463.8	539	0.79	181	0.1	2.85	0.00	41.40	27.8	0.0	n.d.	n.d.	39	42	2.38	0%	0%	36.8
MIX9, MIX10, MIX11	U1301A	7.8	13.0	49.9	9.9	466.2	542	0.83	235	6.2	2.69	0.00	39.00	27.3	0.0	n.d.	n.d.	37	45	2.30	1%	6%	63.8
SW1	above U1301A	7.7	10.4	53.4	10.0	445.3	517	0.76	162	0.0	2.93	0.00	40.70	27.5	0.0	n.d.	n.d.	n.d.	40	2.56	--	--	87.8[a]
SW3, SW4, SW5	above U1301A	7.8	10.4	53.6	9.9	468.2	543	0.78	183	0.0	2.95	0.00	41.30	28.0	0.0	n.d.	n.d.	40	43	2.42	--	--	95.0[a]
SW6, SW7, SW8	above 1025C	7.9	10.2	51.1	10.2	459.9	531	0.84	104	0.0	2.14	0.00	27.30	27.5	0.0	0.7	n.d.	48	32	2.39	--	--	22.0
SW9, SW10, SW11	above U1301A	7.8	10.5	53.3	10.5	473.6	545	0.84	174	0.0	2.80	0.00	40.30	28.2	0.0	0.0	0.00	40	44	2.46	--	--	76.0[a]
SW12, SW13	above U1301A	7.8	10.6	53.9	10.4	471.6	550	0.89	166	0.0	2.80	0.00	41.00	28.6	0.0	0.0	0.00	40	44	2.47	--	--	84.2
SW17, SW18	above U1301A	7.8	10.4	53.0	10.2	463.7	541	0.86	188	0.0	2.70	0.00	42.40	28.1	0.0	0.0	0.00	52	45	2.44	--	--	120.0
SW19, SW20	above U1301A	7.8	10.6	52.8	10.2	463.7	541	0.86	177	0.0	2.70	0.00	42.80	28.1	0.0	0.0	0.00	42	44	2.44	--	--	61.8
SD1	U1363F (~32m)	7.3	20.8	37.8	11.0	498.0	578	0.81	422	197.0	1.04	0.00	0.00	27.5	0.0	22.8	--	192	197	1.63	--	--	n.d.
SD2	U1363B (~38m)	7.4	13.3	47.6	11.9	498.0	608	0.83	493	1897.0	13.80	0.00	0.00	25.1	18.1	37.3	--	478	1897	5.79	--	--	n.d.
SD3	U1363F (~10m)	7.3	12.5	48.3	10.8	475.0	547	0.83	484	774.0	40.40	0.00	0.00	22.1	93.0	47.1	--	862	774	10.34	--	--	n.d.

SD4	U1363G (~2m)	7.3	12.6	48.7	11.9	479.0	555	0.85	456	166.0	15.00	0.00	0.00	28.3	51.3	44.1	--	336	166	3.76	--	--	n.d.
SD5	U1363G (~18m)	7.2	21.1	38.4	11.0	492.0	568	0.84	433	166.0	6.83	0.00	0.00	26.0	49.6	44.6	--	294	166	3.29	--	--	n.d.
SD6, SD9	U1363B (~1m)	7.6	10.4	52.8	12.1	488.0	592	0.79	439	110.0	15.20	0.00	0.00	28.6	8.3	34.1	--	403	110	4.23	--	--	n.d.
SD7	U1363D (~225m)	7.5	35.8	31.3	6.1	478.0	561	0.84	340	89.0	0.30	0.00	0.00	27.6	8.7	88.9	--	103	89	1.25	--	--	n.d.
SD8	U1363B (~45m)	7.2	20.3	38.9	11.2	491.0	596	0.80	529	172.0	3.27	0.00	0.00	25.4	19.6	46.1	--	349	172	2.47	--	--	n.d.
SD10	U1363F (~25m)	7.5	18.1	39.2	10.4	467.0	541	0.85	498	586.0	2.46	0.00	0.00	24.8	26.0	24.5	--	613	586	3.01	--	--	n.d.
SD11	U1363D (~205m)	7.2	30.7	34.4	5.5	462.0	544	0.82	404	147.0	0.30	0.00	0.00	26.6	9.8	65.6	--	122	147	1.51	--	--	n.d.

[a]Jungbluth *et al.*, 2013a

[b]Averaged values of samples in other bags collected within an hour from the same hole.

[c]Calculated value based on basement fluid end-member and seawater mixing.

n.d., not determined; --, not applicable

Table S3. Illumina read statistics

	Non-paired quality filtering		*Read pairing and quality filtering*[a]						
	Qiime default: forward (515F) read	*Qiime default: reverse (806R) read*	*USEARCH*			*FLASH*	*Pandaseq*	*Eren merge-illumina-pairs*	
			e=5.0	*e=1.0*	*e=0.05*			*m/o=0.25, o=48*	*m/o=0.25, o=5*
Total reads (all read lengths)	1734012 (100%)	1513711 (87.3%)	912101 (52.6%)	664404 (38.3%)	35152 (2.0%)	767046 (44.2%)	194913 (11.2%)	965443 (55.7%)	56545 (3.2%)
Mode read length (bp)	~150	~150	~250	~250	~250	~250	~250	~250	~290
Total paired-reads ≤254bp[b]	--	--	849375 (49.0%)	609512 (35.2%)	35132 (2.0%)	758504 (43.7%)	193399 (11.2%)	965443 (55.7%)	17405 (1.0%)
UCHIME_denovo non-chimeras	1723431 (99.4%)	1511248 (87.2%)	837419 (48.3%)	600094 (34.6%)	34270 (2.0%)	748440 (43.2%)	189437 (10.9%)	952375 (54.9%)	n.p.
UCHIME_ref non-chimeras (total reads)	1701960 (98.2%)	1490603 (86.0%)	804762 (46.4%)	577257 (33.3%)	33189 (1.9%)	720495 (41.6%)	181747 (10.5%)	917211 (52.9%)	n.p.
Ave reads per sample	25029	21921	11835	8489	488	10596	2673	13488	n.p.
Min reads per sample	6108	5317	2832	1950	107	2584	711	3250	n.p.
Max reads per sample	43629	38261	19158	14083	864	17195	4575	23098	n.p.
Standard deviation per sample	9169	8080	4292	3165	196	3841	1020	4895	n.p.

[a]1,734,012 overlapping read pairs were used for pairing

[b]Expected merged-pair read length is ~253-254 bases excluding primers

n.p., not performed; --, not applicable

Table S4. Illumina sequence clustering statistics

		Non-paired quality filtering		*Read pairing and quality filtering*					
		QIIME default: forward (515F) read	*QIIME default: reverse (806R) read*	*USEARCH*[c] *e=5.0*	*USEARCH*[c] *e=1.0*	*USEARCH*[c] *e=0.05*	*FLASH*	*Pandaseq*	*Eren merge-illumina-pairs*
Total Reads (no. of unique sequences; percentage)	Similarity/ Cluster Threshold	1701960 (335762; 19.7%)	1490603 (1394851; 93.6%)	804762 (283058; 35.2%)	577257 (180619; 31.3%)	33189 (13045; 39.3%)	720495 (280670; 39.0%)	181747 (74206; 40.8%)	917211 (303095; 33.0%)
Unique read clusters		77055	21956	44893	33226	2336	40753	11721	51239
Uclustref + de novo (GreenGenes)[a]	0.03	15806	148819	9529	7389	1595	9528	4243	11693
Uclustref + de novo (GreenGenes)[a]	0.01	16390	149505	10451	8300	1798	10284	4790	10552
Uclustref + de novo (SILVA)[a]	0.03	15871	150403	9727	7716	1709	9664	4468	11411
Uclustref + de novo (SILVA)[a]	0.01	16869	150735	10723	8527	1826	10523	4977	10320
UPARSE		21145	670280	11277	7336	1429	11571	4283	12463
mothur	0.03	23155	n.p.	10737	8043	1481	14610	5103	10923
mothur	0.01	50897	n.p.	23025	15963	1809	22344	6720	23001
Distribution-based clustering	0.01	37589	n.p.	n.p.	n.p.	1647	n.p.	n.p.	20879
Distribution-based clustering	0.03	18910	n.p.	n.p.	n.p.	1348	n.p.	n.p.	10539
Distribution-based clustering	0.05	15381	n.p.	n.p.	n.p.	1132	n.p.	n.p.	8167
Distribution-based clustering	0.1	12477	n.p.	n.p.	n.p.	877	n.p.	n.p.	5918

[a]Either GreenGenes or SILVA reference databases were used for UCLUST reference-based clustering

n.p., not performed

Table S5. Results of Mantel correlation tests

	All samples		*Subsurface fluids*	
Variable	*r statistic*	*p-value*	*r statistic*	*p-value*
pH	0.412	0.001***	0.184	0.155
Calcium	0.548	0.001***	0.338	0.002**
Magnesium	0.543	0.001***	0.318	0.002**
Potassium	0.470	0.001***	0.246	0.063*
Sodium	-0.067	0.473	0.135	0.257
Chloride	0.154	0.129	0.073	0.540
Bromide	0.033	0.669	0.024	0.809
Silicate	0.559	0.001***	0.140	0.270
Ammonium	0.535	0.001***	0.314	0.004**
Phosphate	0.513	0.001***	0.118	0.372
Nitrite	0.275	0.008**	0.027	0.872
Nitrate	0.583	0.001***	0.436	0.001***
Sulfate	0.531	0.001***	0.311	0.006**
Dissolved Iron	0.552	0.001***	0.467	0.001***
DOC	0.577	0.001***	0.179	0.152
TDN	0.558	0.001***	0.290	0.024*
Alkalinity	0.488	0.001***	0.335	0.003**

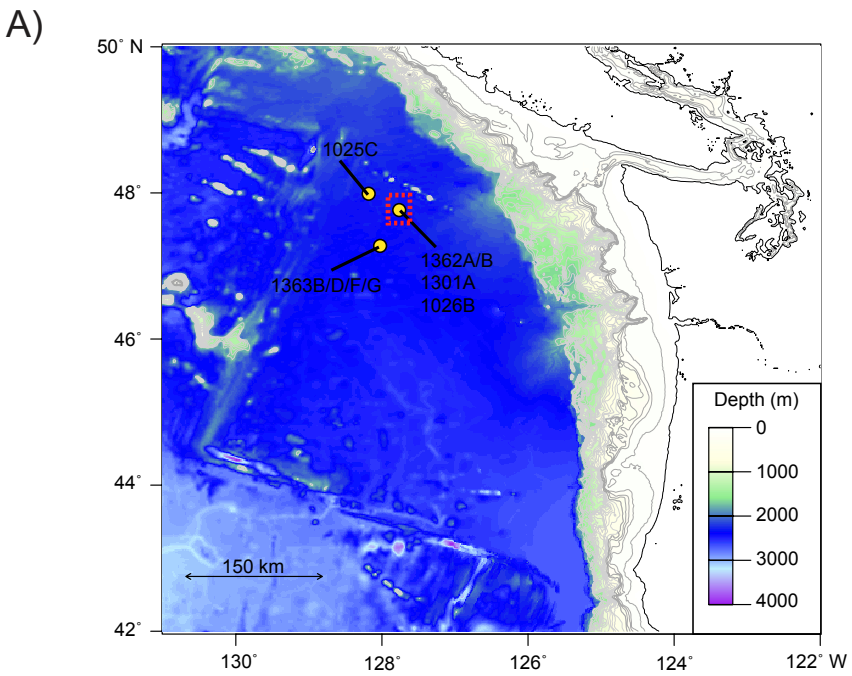
Significance values: ***, ($p \leq 0.001$); **, ($p \leq 0.01$); *, ($p \leq 0.1$)

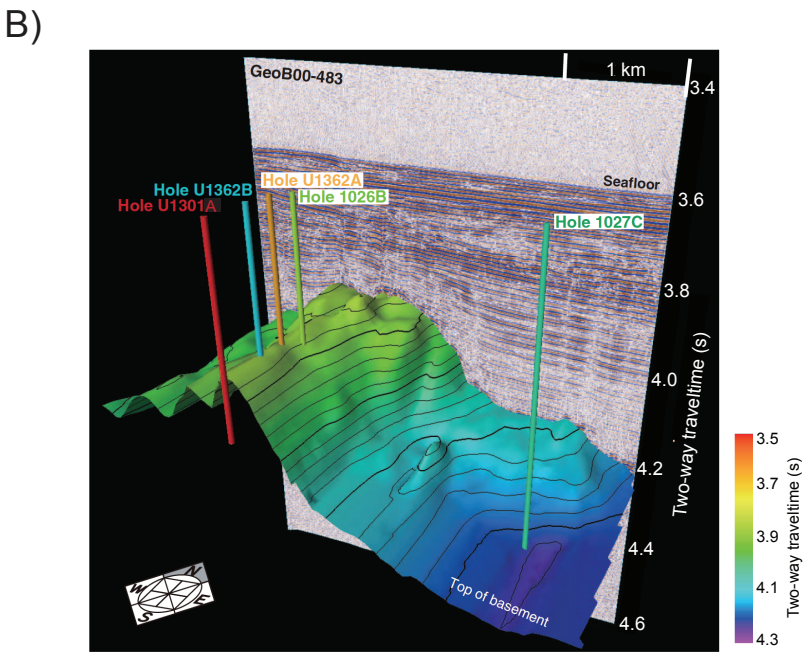
Table S6. Results of PERMDISP statistical test after grouping samples by type[1]

	seawater	sediment	subsurface
seawater	--	0.091*	0.002**
sediment	0.090*	--	0.081*
subsurface	<0.001***	0.083*	--

[1]Observed p-value below diagonal, permuted p-value above diagonal.

Figure S1


A)
50° N
48°
46°
44°
42°
130°
128°
126°
124°
122° W
1025C
1362A/B
1301A
1026B
1363B/D/F/G
150 km
Depth (m)
0
1000
2000
3000
4000


B)
GeoB00-483
1 km
Hole U1362B
Hole U1301A
Hole U1362A
Hole 1026B
Hole 1027C
Seafloor
Top of basement
3.4
3.6
3.8
4.0
4.2
4.4
4.6
Two-way traveltime (s)
3.5
3.7
3.9
4.1
4.3
Two-way traveltime (s)

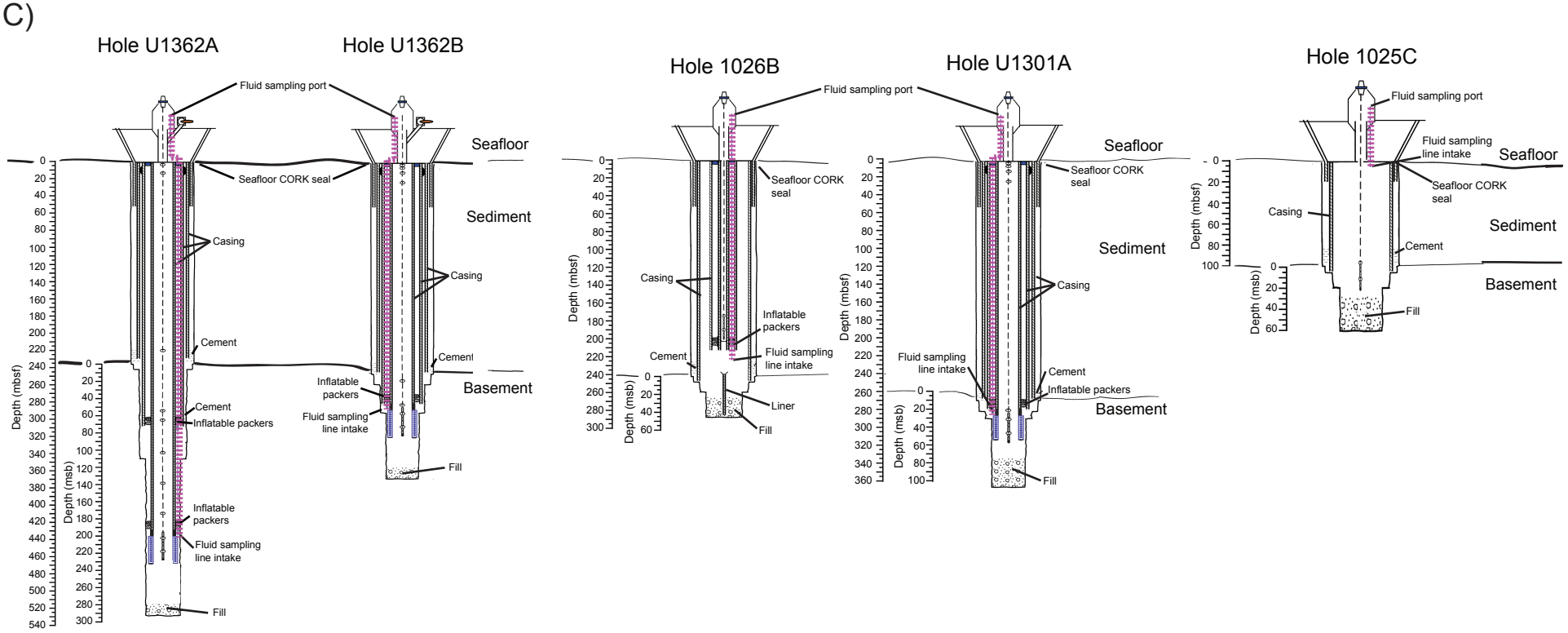

C)
Hole U1362A
Hole U1362B
Hole 1026B
Hole U1301A
Hole 1025C
Fluid sampling port
Seafloor
Seafloor CORK seal
Sediment
Casing
Cement
Basement
Inflatable packers
Fluid sampling line intake
Liner
Fill
Depth (mbsf)
Depth (msb)

Figure S2

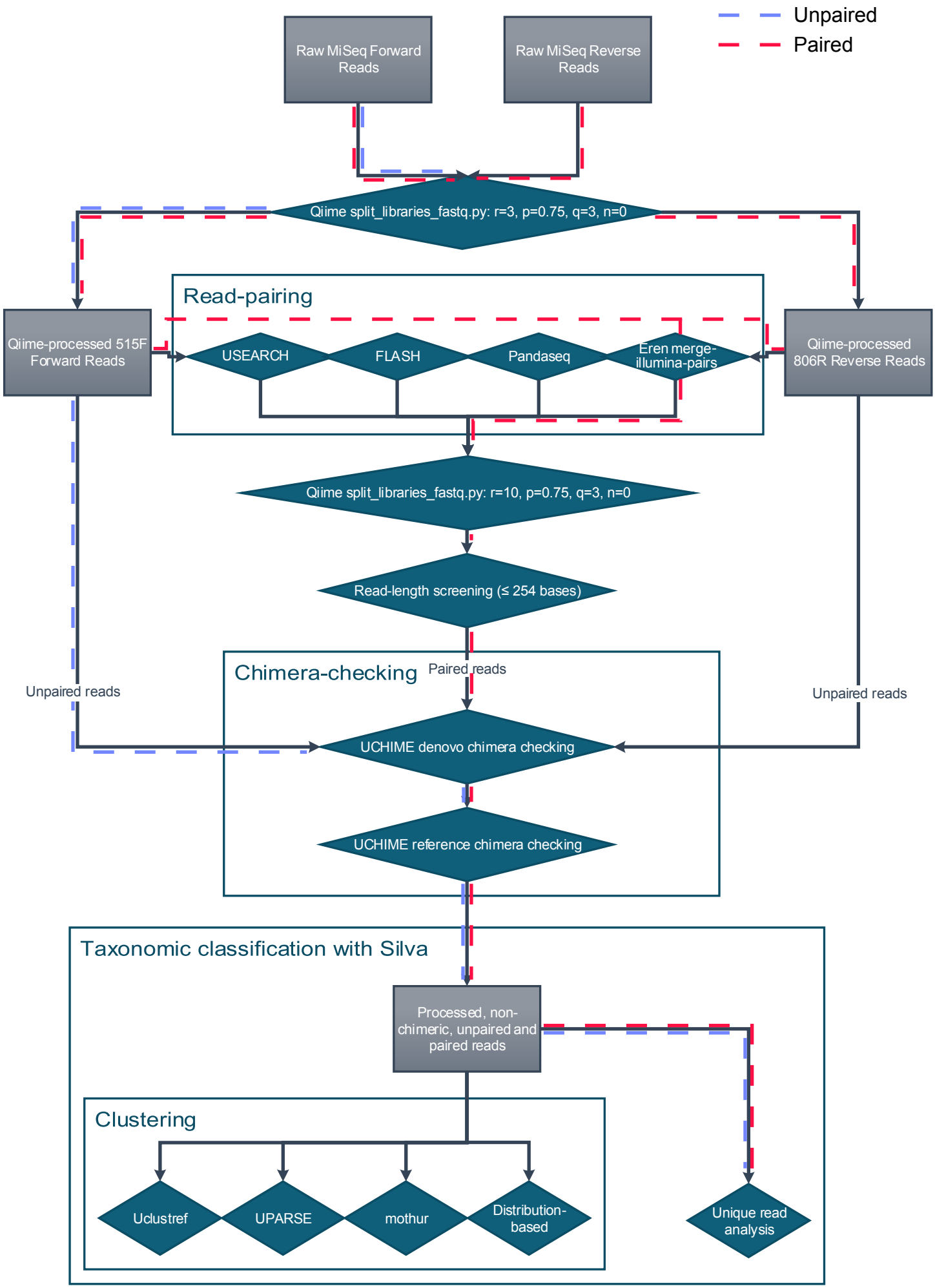

Unpaired
Paired
Raw MiSeq Forward Reads
Raw MiSeq Reverse Reads
Qiime split_libraries_fastq.py: r=3, p=0.75, q=3, n=0
Read-pairing
Qiime-processed 515F Forward Reads
USEARCH
FLASH
Pandaseq
Eren merge-illumina-pairs
Qiime-processed 806R Reverse Reads
Qiime split_libraries_fastq.py: r=10, p=0.75, q=3, n=0
Read-length screening (≤ 254 bases)
Chimera-checking
Paired reads
Unpaired reads
Unpaired reads
UCHIME denovo chimera checking
UCHIME reference chimera checking
Taxonomic classification with Silva
Processed, non-chimeric, unpaired and paired reads
Clustering
Uclustref
UPARSE
mothur
Distribution-based
Unique read analysis

Figure S3

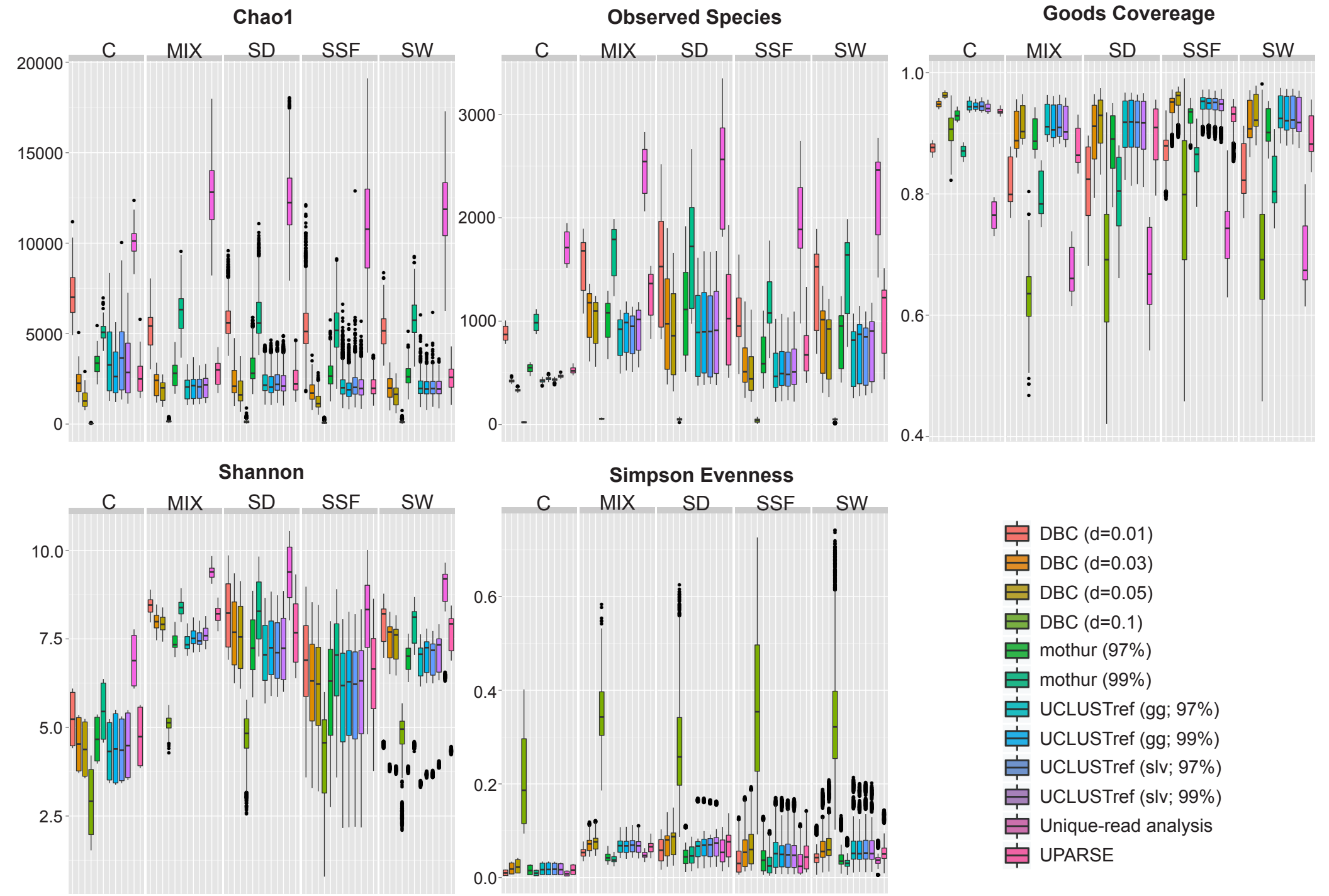

Chao1
Observed Species
Goods Covereage
Shannon
Simpson Evenness
C
MIX
SD
SSF
SW
DBC (d=0.01)
DBC (d=0.03)
DBC (d=0.05)
DBC (d=0.1)
mothur (97%)
mothur (99%)
UCLUSTref (gg; 97%)
UCLUSTref (gg; 99%)
UCLUSTref (slv; 97%)
UCLUSTref (slv; 99%)
Unique-read analysis
UPARSE

Figure S4

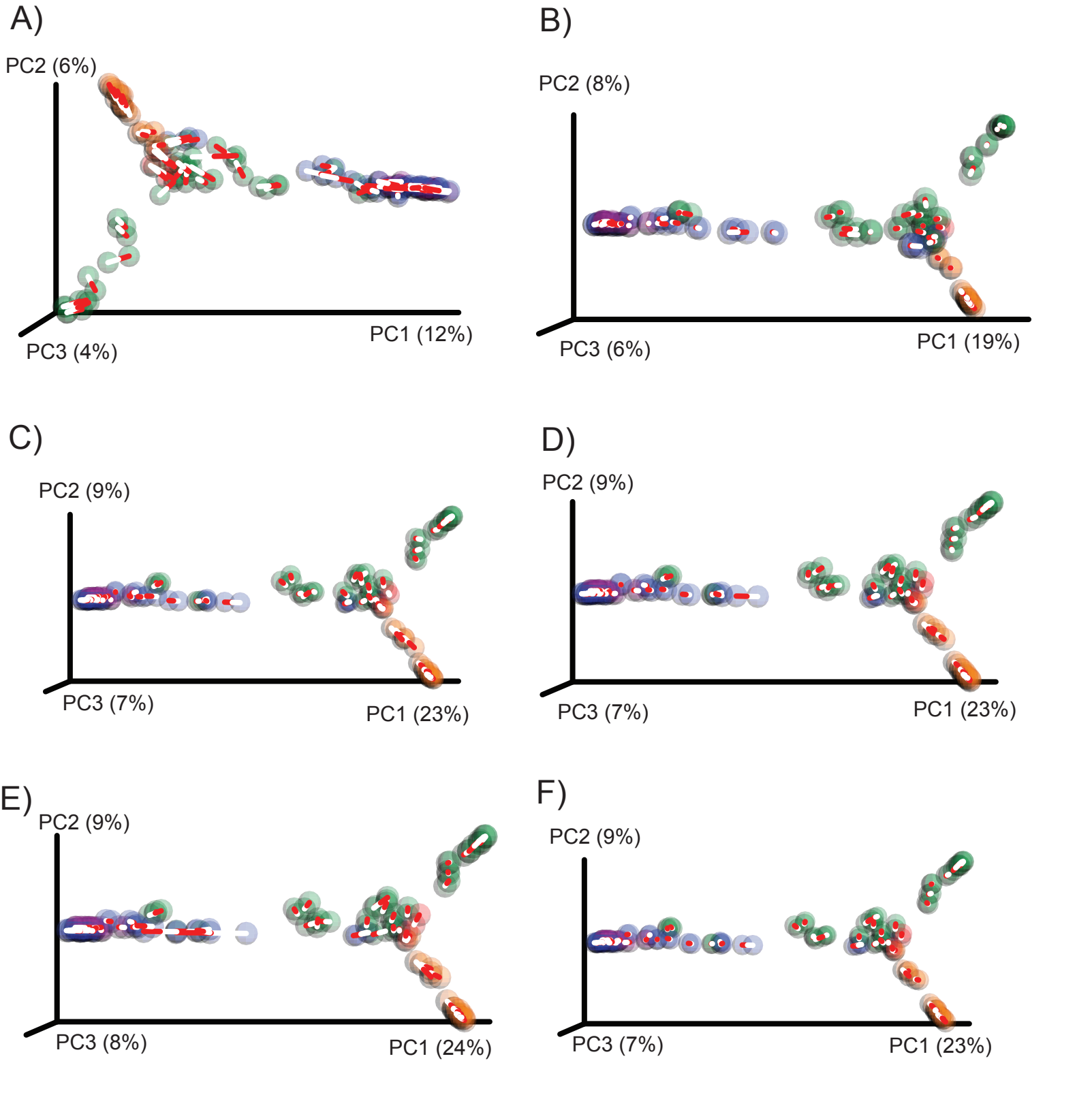
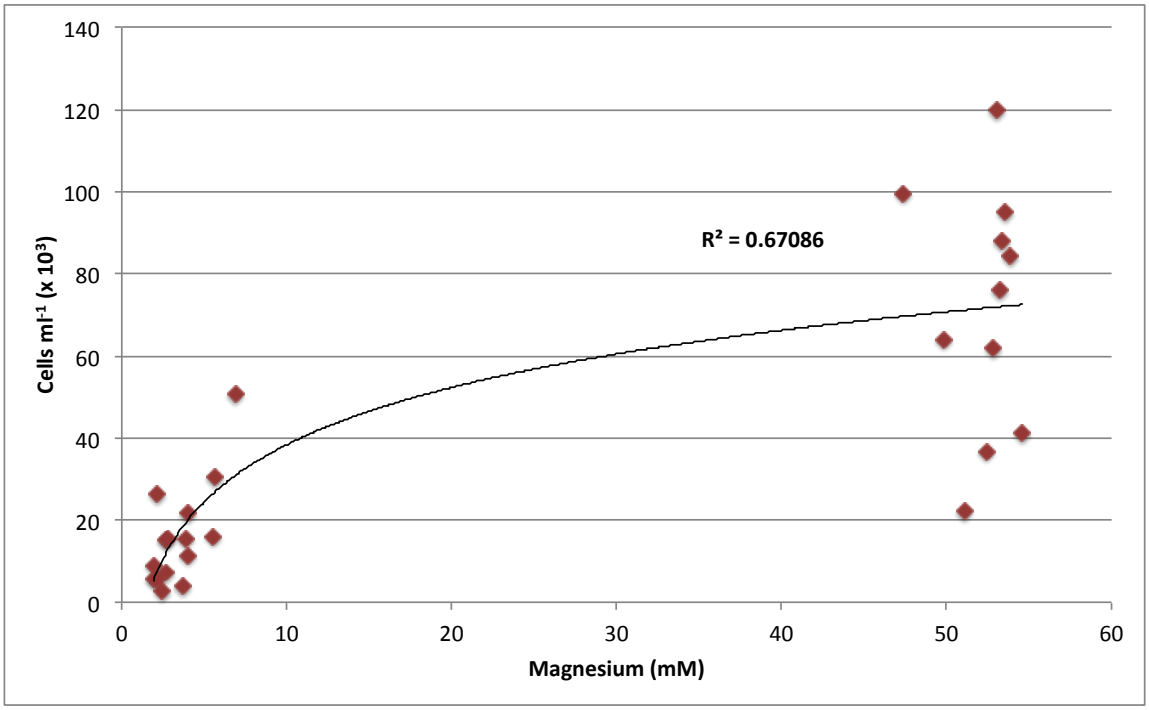

A)
PC2 (6%)
PC3 (4%)
PC1 (12%)
B)
PC2 (8%)
PC3 (6%)
PC1 (19%)
C)
PC2 (9%)
PC3 (7%)
PC1 (23%)
D)
PC2 (9%)
PC3 (7%)
PC1 (23%)
E)
PC2 (9%)
PC3 (8%)
PC1 (24%)
F)
PC2 (9%)
PC3 (7%)
PC1 (23%)

Figure S5

A)


140
120
100
80
60
40
20
0
Cells ml-1 (x 10³)
R² = 0.67086
0
10
20
30
40
50
60
Magnesium (mM)

B)

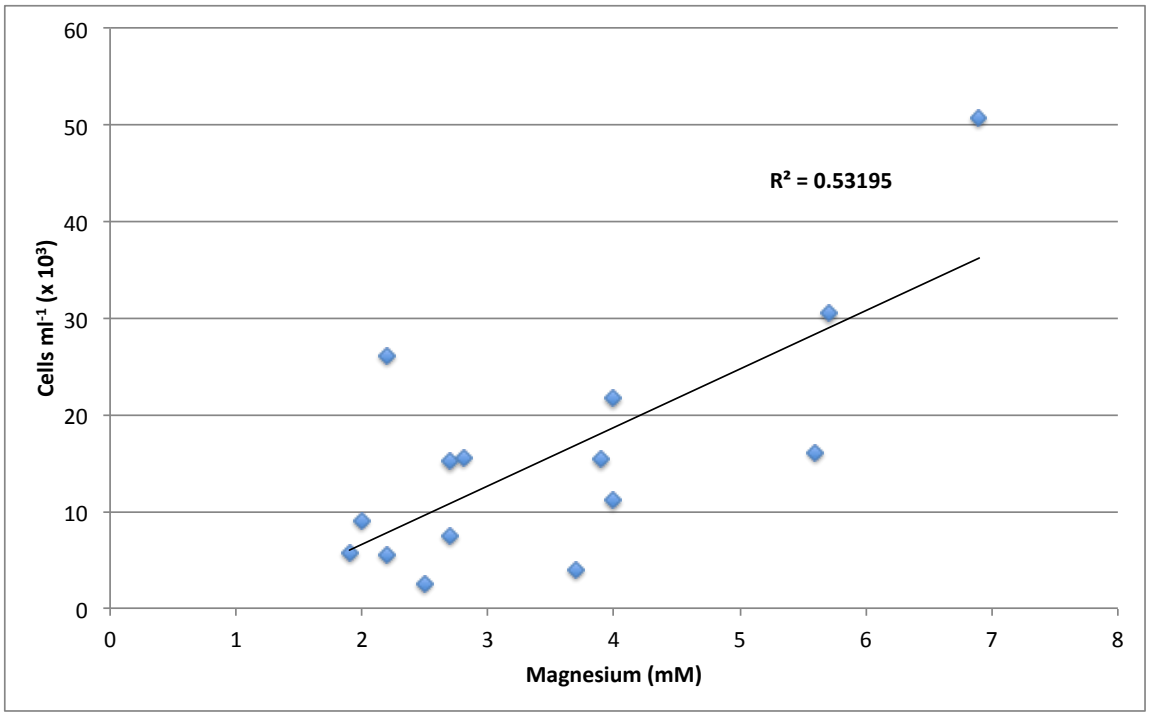
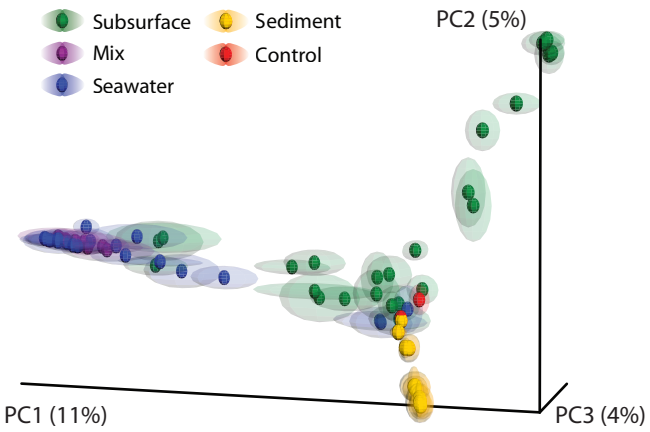

60
50
40
30
20
10
0
Cells ml-1 (x 10³)
R² = 0.53195
0
1
2
3
4
5
6
7
8
Magnesium (mM)

Figure S6

A)


Subsurface
Mix
Seawater
Sediment
Control
PC2 (5%)
PC1 (11%)
PC3 (4%)

B)

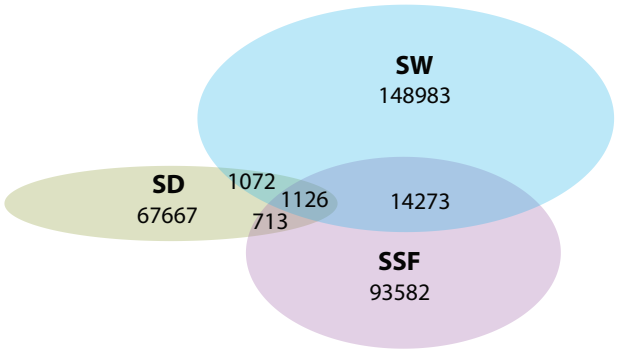
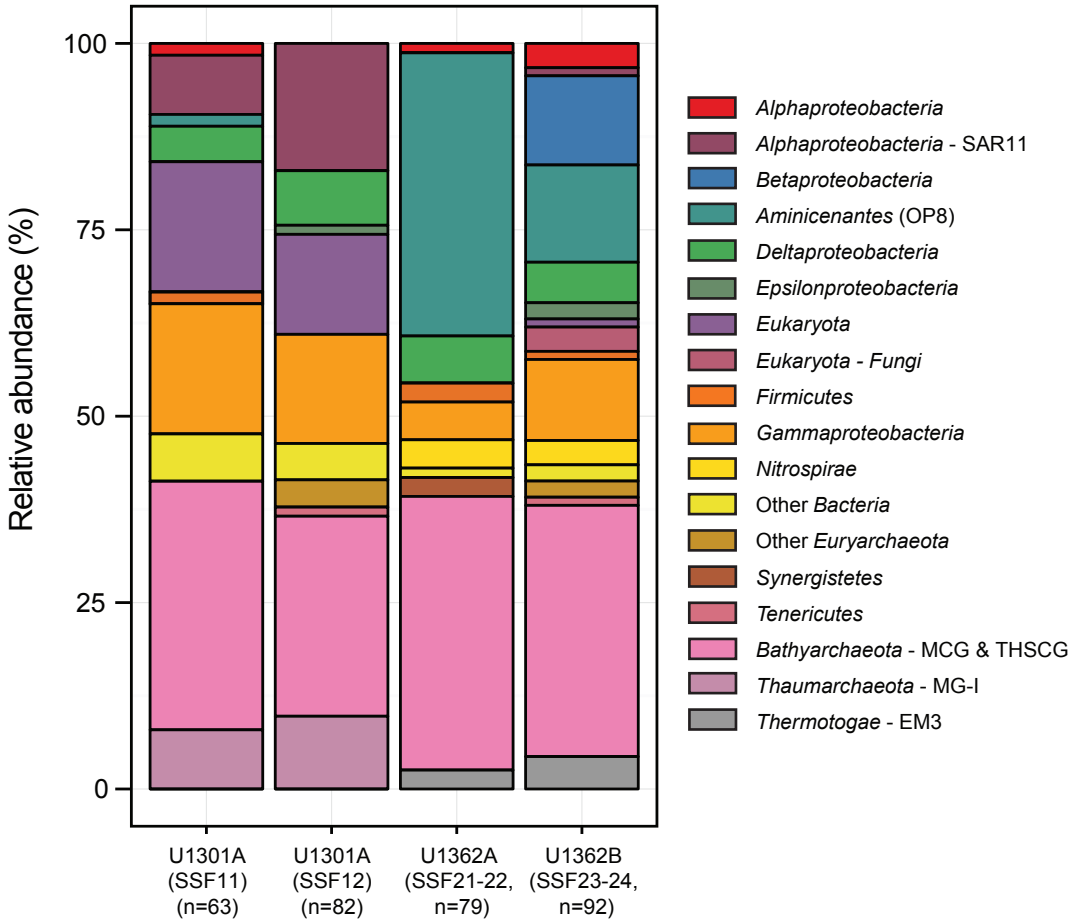

SW
148983
SD
67667
1072
1126
713
14273
SSF
93582

Figure S7


Relative abundance (%)
100
75
50
25
0
U1301A (SSF11) (n=63)
U1301A (SSF12) (n=82)
U1362A (SSF21-22, n=79)
U1362B (SSF23-24, n=92)
Alphaproteobacteria
Alphaproteobacteria - SAR11
Betaproteobacteria
Aminicenantes (OP8)
Deltaproteobacteria
Epsilonproteobacteria
Eukaryota
Eukaryota - Fungi
Firmicutes
Gammaproteobacteria
Nitrospirae
Other Bacteria
Other Euryarchaeota
Synergistetes
Tenericutes
Bathyarchaeota - MCG & THSCG
Thaumarchaeota - MG-I
Thermotogae - EM3