Supplemental material for:
Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm by D. S. Jones, H. L. Albrecht, K. S. Dawson, I. Schaperdoth, K. H. Freeman, Y. Pi, A. Pearson, and J. L. Macalady

Supplemental Methods

Field site, sample collection, and geochemistry

Snottite sample RS24 was collected in August, 2005 from Ramo Sulfureo site RS2 in the Frasassi cave system, Italy (Figure S1). We collected approximately two grams of biofilm from one square meter of cave wall. Within five hours of collection, samples for fluorescent *in situ* hybridization were fixed with 4% paraformaldehyde as in Macalady et al. (2007), and samples for DNA and lipid extractions were preserved in RNAlater (Ambion, USA). Biofilm pH was measured in the field with pH paper (range 0-2.5), and was between 0 and 1 for snottites in the collection area. Concentrations of $CO_2(g)$, $H_2S(g)$, $NH_3(g)$, and $NO_2(g)$ were measured using Dräger tubes (Dräger Safety, Germany). $H_2S(g)$ concentrations were between 8 to 24 parts-per-million by volume (ppmv) (4 measurements), $CO_2(g)$ was 3300 ppmv (2 measurements), and $NH_3(g)$ and $NO_2(g)$ were below 0.25 and 0.1 ppmv respectively. $CH_4(g)$ concentration at site RS2 was measured in August, 2006. Replicate glass flasks with Glass Expansion™ valves were flushed 10x with a hand pump. $CH_4(g)$ measurements were made using standard gas chromatographic techniques with a flame ionization detector (Agilent™ 5890 Series II) (Sowers et al., 2005). Data are reported on the new NOAA 04 gravimetric scale (Dlugokencky et al., 2009) with an overall uncertainty of ±0.02 ppmv. $CH_4(g)$ results on the two flasks were 1.9 and 2.2 ppmv, which is in good agreement with atmospheric $CH_4(g)$ concentrations at the nearest NOAA/ERSL continuous monitoring station located at Hegyhatsal, Hungary (46°N, 16°E, 248 meters above sea level) (Dlugokencky et al., 2009). $SO_2(g)$ at site RS2 was measured on several occasions in 2006 through 2008 with a MX-2100 portable gas sensor (Enmet Corp., USA), and was always below detection (lower detection limit 0.1 ppmv).

Full cycle rRNA methods

Fluorescent *in situ* hybridization (FISH) was performed using the probes and cell counting technique described in Macalady et al. (2007). Environmental DNA was extracted as described in Bond et al. (2000), after first diluting the RNAlater-preserved sample with 3 parts phosphate-buffered saline (PBS) to 1 part sample. To remove excess polysaccharides from the final extract, we re-precipitated the DNA under high salt concentrations as follow: the pellet was resuspended in 200 uL Tris (200 mM, pH 8.0), 100 uL NaCl (5 M), and 600 uL ethanol (100%), incubated at -20°C for 30 minutes, and pelleted by centrifugation for 20 minutes at 4°C.

A 16S rDNA clone library was constructed from sample RS24 using archaeal specific PCR primers. Each reaction contained 50 µL of DNA template (1-150 ng), 1.25 units high-affinity ExTaq DNA polymerase, 1x buffer (TaKaRa Bio Inc., Shiga, Japan), 0.2 mM each dNTP, 0.2 µM forward primer 344f (ACGGGGYGCAGCAGGCGCGA) (Raskin et al., 1994), and 0.2 µM reverse primer deg1392r (ACRGGCGGTGTGTRC). Primer deg1392r is a modification of primer 1392r (Lane, 1991). Thermal cycling was as follows: 5 minutes initial denaturation at 94°C, followed by 25 cycles of denaturation (45 seconds at 94°C), annealing (45 seconds at 50°C), and elongation (2 minutes at 72°C), and 20 minutes final elongation at 72°C.

Cloning, transformation, and colony PCR were performed using the materials and procedure described by Macalady et al. (2008). Clones were sequenced at the Penn State University Biotechnology Center using plasmid-specific primers T3 and T7. Twenty-eight archaeal 16S rRNA gene sequences were obtained from sample RS24. Sequences were manually curated and assembled with CodonCode Aligner v.2.0.4 (CodonCode Corporation, USA). The library was checked for chimeric sequences using Bellerophon (Huber et al., 2004), and no chimeric sequences were detected.

16S rRNA sequences were first aligned with the NAST aligner (DeSantis et al., 2006) and manually refined in ARB (Ludwig et al., 2004). Alignments were trimmed so that all sequences were of equal length, and nucleotide positions with less than 40% base-pair conservation were masked. Only sequences greater than 1300 base pairs were used for phylogenetic analyses of bacterial sequences. Archaeal alignments were trimmed to 1018 characters, and the only sequences with missing information were shorter 'DSJA' sequences from Macalady et al. (2007). Maximum likelihood (ML) analyses were performed in PAUP* v.4b10 (Swofford, 2000) using the general time-reversible (GTR) model, five random-addition-sequence replicates, and tree bisection-reconnection (TBR) branch swapping. Base frequency, substitution rates, and shape parameter for the gamma distribution were estimated from the data. Tree topology for our ML analyses was identical to ML analyses performed with parameters selected via MODELTEST (Posada and Crandall, 1998). Bayesian phylogenetic analyses were performed with MrBayes v.3.1 (Ronquist and Huelsenbeck, 2003) using the GTR model and gamma-distributed rate variation. Analyses were run for 500,000 generations with 4 independent chains, trees were produced every 100 generations, and posterior clade probabilities were calculated from a consensus tree computed after discarding the first 20% of trees. Maximum parsimony and neighbor joining analyses were performed in PAUP*. Parsimony analyses were performed by heuristic search with 100 random-addition-sequence replicates, TBR branch swapping, and 2000 bootstrap replicates. Neighbor joining was performed with Jukes-Cantor (JC) corrected distance matrix and 2000 bootstrap replicates.

Partial squalene-hopene cyclase (SHC) genes were amplified from RS24 DNA extract as described in Pearson et al. (2007). PCR reactions were prepared with Sigma RedTaq™ ReadyMix™, with forward primer SHCF and reverse primers SHCCyR, SHCPrBR, and SHCPr_R as described previously. Separate clone libraries were constructed from SHCF/SHCCyR, SHCF/SHCPrB, and SHCF/SHCPr_R primer combinations. Cloning reactions were performed as described in Pearson et al. (2007). Sequencing in both directions was done at the Dana-Farber/Harvard Cancer Center DNA Resource Core (http://dnaseq.med.harvard.edu/) using primers M13F and M13R, and data were curated manually. An additional partial SHC sequence was obtained by direct sequencing a SHCF/SHCPr_R amplicon produced by nested amplification of a 700 base pair product from the original SHCF/SHCPr_R amplicon. This clone (SDPr_DirF) gave a clean read only in middle of the amplicon and yielded a short translated length of 126 amino acids. Sequences were translated into open reading frames (ORFs) by the ExPASy translate tool (http://ca.expasy.org/tools/dna.html) and preliminary alignments were obtained in ClustalW (http://align.genome.jp/). Clones were classified as identical if they had more than 99% amino acid identity.

Lipid extraction and analysis

For total lipid extraction from RS24 biofilm, the sample was first rinsed three times with PBS to remove RNAlater, and freeze-dried until further analysis. Lipids were separated by a

modified Bligh-Dyer extraction as described by Talbot et al. (2003), with dichloromethane substituted for chloroform.

Analysis of ether lipids followed the procedure developed by Hopmans et al. (2000), using an Agilent 6310 high pressure liquid chromatograph/mass spectrometrometer (HPLC/MS^n) with an atmospheric pressure chemical ionization (APCI) chamber. Separation was achieved on a Prevail Cyano column (2.1 x 150 mm 3-µm; Alltech, Deerfield, IL), maintained at 30°C with a flow rate of 0.2 ml/min. Isoprenoid glycerol dialkyl glycerol tetraether lipids (GDGTs) were eluted isocratically with 99% hexane: 1% isopropanol for 5 min, followed a linear gradient to 98.4% hexane: 1.6% isopropanol over 40 min. After each analysis the column was cleaned with 90% hexane: 10% isopropanol for 10 min, followed by a 10 minute reequilibration to 99% hexane: 1% isopropanol. Instrument conditions were as follows: corona = 5000 nA, capillary = 3500 V, nebulizer = 60 psi, dry gas = 6 l/min, dry temperature = 200°C, and vaporizer temperature = 400°C. Isoprenoid glycerol dialkyl glycerol tetraether lipids (GDGT's) were identified based upon comparison to extracted lipids from a laboratory culture of *Thermoplasma acidophilum*, with GDGT's containing zero to five cyclopentane rings.

Analysis of bacteriohopanepolyols ('hopanoids') was performed on the HPLC/MS^n described above using the procedure and conditions from Talbot et al. (2003) with minor modifications. First, lipid extracts were acetylated by reaction of 200 µl pyridine and 200 µl acetic anhydride at 60ºC for 1 hr. Reversed-phase HPLC was performed using a Phenomenex Gemini 5µm C18 column (150mm x 3.0mm i.d.) and a 5µm pre-column of the same material. Instrument conditions were as follows: positive APCI scanning from 150-1300 m/z, 'smart' fragmentation with corona voltage = 8000 nA, nebulizer pressure = 60 psi, dry gas = 5 l/min, dry temperature = 350°C, and vaporizer temperature = 490°C. Three target mass segments were created: 0-10 min 285m/z, 10-17 min 1002 m/z, 17-50 min 655 m/z, all with optimization normal. Auto MS^n settings with 2 precursor ions were as follows: absolute threshold 100,000, relative threshold 5%, exclude after 2 ions, and release after 0.5 min. The acquisition parameter has a fragment amplitude 1.0 V, with an average of 5, and isolation width of 3.0 m/z. Auto MS(n>2) settings with 1 precursor ion are as follows: absolute threshold = 1000, relative threshold = 5%, and fragment amplitude = 1.0V.

Metagenomic sequencing

Environmental DNA from sample RS24 was pyrosequenced at the Pennsylvania State University Center for Genomic Analysis with a GS20 Sequencing System (454 Life Sciences, USA) (Marguiles et al., 2005) and analyzed with updated GS20 analytical software version 1.1.01.20.

Estimation of metagenomic sequence coverage

We estimated the expected fraction of genome coverage for each member of the RS24 snottite community as described by Whitaker and Banfield (2006) (Figure S7). First, average genome coverage (c) of each taxon i in the community is determined by equation S1:

$$C_i = \frac{G_i A_i}{\sum_{i=1}^{n} G_i A_i} \quad \text{(S1)}$$

where g_i is the genome size of species i, a_i is the abundance of species i in the community determined by FISH, and n is the total number of taxa in the community (Whitaker and Banfield, 2006). Genome size of each phylotype was assumed to be roughly that of their closest relative.

Based on the following genome sizes—*Acidithiobacillus ferrooxidans*, 3 Mb (J. Craig Venter Institute CMR: http://cmr.jcvi.org/), *Acidimicrobium ferrooxidans*, 2.1 Mb, *Ferroplasma acidarmanus*, 1.9 Mb, and *Picrophilus torridus*, *Thermoplasma acidiphilum*, *Thermoplasma volcanum*, 1.6 Mb each (http://img.jgi.doe.gov/mer)—we assume genome sizes of 2 Mb for the *Acidimicrobiaceae* and *G-plasma*, and 3 Mb for *Acidithiobacillus* and rare organisms. Proportion of the genome of each organism in the RS24 metagenomic dataset was estimated based on a Poisson distribution of base pair coverage (equation S2; Sokal and Rohlf, 1996):

(S2) $Y = e^{-C}$

where Y is the proportion of base pairs not sequenced in the dataset, and c is the average coverage of that organism (equation S1). Proportion of the total genome present in the dataset is simply 1-Y.

In order to confirm expected coverage calculations, we measured coverage of the snottite *Acidithiobacillus* population by blasting the metagenome against single copy genes from all three *Acidithiobacillus spp.* for which complete genome sequences are available*: At. caldus* ATCC 51756, *At. ferrooxidans* str. ATCC 23270, and *At. ferrooxidans* str. ATCC 53993. We used 18 of the 31 universal sequences from Ciccarelli et al. (2006): COG0016, COG0048, COG52, COG0081, COG0087, COG0091, COG0092, COG0093, COG0096, COG0097, COG0098, COG0099, COG0102, COG0103, COG0184, COG0197, COG0256, and COG0533. This subset represents all the sequences that Ciccarelli et al. identified as single copy for every organism tested.

To further evaluate how our metagenome represents the genetic potential of major snottite populations, we simulated random shotgun sequencing of three complete genomes. Then, we calculated the resulting proportion of the genome and total gene complement represented for different amounts of sequencing. We used complete genomes of *Thermoplasma acidophilum* str. DSM 1728*, Bdellovibrio bacteriovorus* str. HD100*,* and *Thiomicrospira crunogena* str. XCL-2, with accompanying annotations from the UCSC Archaeal Genome Browser (http://archaea.ucsc.edu/), available through Galaxy (http://galaxy.psu.edu/). To simulate shotgun sequencing by a 454 GS20 platform, we used a script to randomly select fragments averaging 100 bp in length (with standard deviation of 10 bp), until a designated coverage was achieved. Coverage is defined as (total nucleotides sequenced)/(genome size). For each simulation, sequencing was replicated 100 times and the average and standard deviation of the total nucleotide positions and total genes represented were recorded. A gene was considered sequenced if at least 50 bp were represented by a simulated read. Results are summarized in Figure S8. Sequencing to 0.75x coverage represents only 53% of the total nucleotide positions in a genome, but over 90% of genes represented by at least two reads. This approach is unique because it addresses gene content separately from nucleotide positions in a genome, and shows that small amounts of sequencing are sufficient to represent the gene content of an organism. The python script used here to simulate random shotgun sequencing is freely available upon request.

Additionally, we estimated how well the RS24 metagenome represents the gene content of each snottite organism by extracting and quantifying reads matching tRNA synthase genes (Table 1, main text), using the COG annotation described below.

Annotation, taxonomic classification, and binning of metagenomic reads

All metagenome reads were compared via BLASTX (Altschul et al., 1997) to the NCBI non-redundant (nr) database. Prior to analysis, identical read copies were removed from the RS24 dataset. Putative 16S and 23S rRNA sequences were identified with BLASTN (e-value

cutoff 1e-20) against ARB-SILVA databases (Pruesse et al., 2007) and excluded from BLASTX analyses. The (final) nr database used in these analyses was downloaded on January 10th, 2010. Before analysis, protein sequences from the 'Thermoplasmatales archaeon Gpl' bin from the Acid Mine Drainage microbiome (Tyson et al., 2004; available at http://img.jgi.doe.gov/m) were added to the nr database.

A binning approach based on phylogenetic identity was used to assign metagenomic reads to taxonomic groups. Binning was performed using MEGAN (Husen et al., 2007). BLASTX output was input to MEGAN v.3.2.1, and reads were assigned to taxa using a bit score threshold of 40 (min score filter, with min support of 1), retaining hits within 10% of the top bit score (top percent filter). We chose these parameters based the following criteria: (1) We randomly extracted 11,000 reads from the RS24 database (10% of the total reads), randomized the nucleotide sequence of each read as described by Biddle et al. (2008), and blasted the randomized dataset against the NCBI nr database. There was only one random match above bit score 40 (Figure S13). (2) We tested false assignment rate using two simulated datasets provided with MEGAN software: one dataset generated from *Bdellovibrio bacteriovorus* HD100, and one generated from *Escheria coli* K12. Each dataset contained 2000 reads with average length of 100 bp. The *E. coli* K12 dataset was blasted (by Husen et al., 2007) against a database from which all *E. coli* K12 sequences had been removed. The combination of parameters described above was chosen because it produced no false assignments with the *Bdellovibrio* dataset, and only 1.1% and 0.16% assignments outside *Escherichia* and Gammaproteobacteria with the *E. coli* dataset. The MEGAN parameters we chose for binning are more stringent than those recommended for GS20 pyrosequencing data by Husen et al. (2007). Annotation and binning with MEGAN was tested further with additional simulated datasets, described below.

We created bins for the three most abundant organisms in the metagenome. Reads assigned to *Gammaproteobacteria* are considered to have come from *Acidithiobacillus*, *Actinobacteria* from *Acidimicrobiaceae*, and all archaea from 'G-plasma.' Such an approach is reasonable because snottite sample RS24 has low biodiversity, and each of these high-level clades is represented predominantly by a single snottite population. Each of these bins is expected to have minor contributions from rare taxa, such as C- and E-plasma sequences in the G-plasma bin (Figure S6). In some cases, functions were assigned to *Acidithiobacillus* based on copy number. Because of its high sequence coverage (Figure S7), *Acidithiobacillus* is also assigned functions with at least 20 matching reads, which corresponds to at least 2x coverage of a 1000 bp gene.

Reads were annotated to functional categories identified in the Clusters of Orthologous Groups (COG) of proteins classification system (Tatusov et al., 1997; Tatusov et al., 2003). Nucleotide sequences of reads that significantly matched the nr database, as described above, were compared against the pre-compiled database of COG position specific scoring matrices (PSSM) using reverse position-specific (RPS) BLAST (Marchler-Bauer et al., 2002; Marchler-Bauer et al., 2009). The decision to assign COGs using RPSBLAST of nucleic acid sequences was based on a comparison of false COG assignments using simulated GS20 pyrosequencing datasets constructed from full-length protein sequences (see below).

In order to identify specific metabolic capabilities that are not represented by COG categories: First, sequences corresponding to each gene of interest were downloaded from the REFSEQ database in genbank, or from literature sources (e.g. sources cited for C-fixation genes, below). Matching metagenome sequences were identified by BLASTX of the RS24 metagenome against each sequence dataset using the same BLASTX criteria described above. Successful

matches to each gene of interest were checked against the top blast hits of that sequence versus the complete nr database, and were not considered if that read had higher homology to unrelated proteins.

To identify carbon-dioxide fixation pathways in the metagenome, we specifically targeted enzymes that are diagnostic for each known pathway. Specifically, we searched for evidence of the reductive pentose phosphate pathway (RuBisCO and phosphoribulokinase; Shively et al., 1998), the reductive acetyl-CoA pathway (CO dehydrogenase/acetyl-CoA synthases; Ragsdale and Pierce, 2008), the rTCA cycle (citrate lyase; Hügler et al., 2007), the 3-hydroxypropionate cycle (propionyl-CoA synthetase and malonyl-CoA reductase from organisms known to possess the 3-hydroxypropionate cycle; Klatt et al., 2007; Zarzycki et al,. 2009), and the 3-hydroxypropionate/4-hydroxybutyrate and dicarboxylate/4-hydroxybutyrate cycles (4-hydroxybutyryl-CoA dehydratase; Berg et al., 2010).

Simulated datasets for COG assignment

In order to assign COG categories to short metagenomic reads in the most rigorous way possible, we tested several methods using simulated GS20 pyrosequencing datasets. Simulated datasets were generated with a python script that randomly selects short nucleic acid sequences from a fasta file of full-length protein sequences. In order to mimic GS20 pyrosequencing data, the average length of simulated nucleic acid sequences was 100 bp (normally distributed, with standard deviation ±10 bp). Simulated reads are each generated from a random protein in the input file, and from a random starting point in that protein.

To test accuracy in COG assignments, we used two simulated datasets composed of 5000 simulated reads generated from *Acidithiobacillus ferrooxidans* ATCC 23270 and *Bdellovibrio bacteriovorus* HD100. Protein sequences and COG assignments are available at the JGI IMG system (Markowitz et al., 2010; http://img.jgi.doe.gov). It is important to note that neither of these genomes are included in the COG database (Tatusov et al., 2003). We assigned COGs to reads in the simulated dataset using several techniques, and compared the resulting COG assignments of each simulated read to the COG assignment for the full length protein that the simulated read was generated from. We assigned COGs using four techniques:

i. BLASTX of nucleic acid sequences against the myva sequence database (myva is a database of all sequences used to generate COGs).
ii. BLASTP of amino acid translations of each sequence against the myva sequence database. Translations were generated by first comparing nucleic acid sequences with BLASTX against the complete non-redundant (nr) database, and extracting the amino acid sequences produced by BLASTX local alignment.
iii. RPSBLAST of nucleic acid sequences against the COG database.
iv. RPSBLAST of amino acid translations against the COG database.

Error in COG assignments using each technique is reported in Table S2. Proteins that are assigned multiple COG categories in IMG system (Markowitz et al., 2010; http://img.jgi.doe.gov) were excluded from analyses. RPS-BLAST produced a lower error rate than BLASTX. RPS-BLAST of amino acid translations produced a slightly lower error rate than RPS-BLAST of nucleic acid sequences. However, analyses comparing nucleic acid sequences produced substantially more COG assignments than analyses using translations, so we chose to use RPS-BLAST of nucleic acid sequences for COG assignments. The python script used to generate simulated datasets is freely available upon request.

Simulated datasets for testing rates of read assignment and binning

Despite the low microbial diversity of the snottite community, we could only confidently assign 40.5% of the metagenome reads (Figure 2a, main text). Although an annotation rate of 40.5% is high compared with other metagenomic studies that use unassembled 454 GS20 sequences (c.f. Biddle et al., 2008; Dinsdale et al., 2008), it is low for datasets of full length gene sequences. This is because for shorter reads, we cannot use *ab initio* gene predictors and must assign function based on similarity between the short metagenome reads and full-length sequences in public databases.

To illustrate this phenomenon and to evaluate our annotation rate, we compared BLASTX and MEGAN analyses of simulated datasets. We simulated 10,000 random shotgun reads, as described earlier, from three genomes: *Sulfurovum sp.* str. NBC37-1, *Picrophilus torridus* str. DSM 9790, and *Acidithiobacillus ferrooxidans* str. ATCC 53993. We also simulated a metagenome with 20,000 reads, in which 75%, 18.5%, and 6.5% of the sequences were from *At. ferrooxidans* str. ATCC 53993, *P. torridus* str. DSM 9790, and *Sulfurovum. sp.* str. NCB37-1, respectively. Then, we blasted each simulated dataset by BLASTX against the most recent version of the GenBank non-redundant (nr) protein database. First, datasets were first blasted against the complete nr database, and then again against a modified nr from which all sequences corresponding to *Sulfurovum sp.* str. NBC37-1, *P. torridus* str. DSM 9790, and all *At. ferrooxidans* strains were removed. (This includes sequences from *At. ferrooxidans* str. ATCC 23270 in addition to sequences from *At. ferrooxidans* str. ATCC 53993.) Additionally, we blasted the simulated *At. ferrooxidans* and metagenome datasets against a second modified nr from which only *At. ferrooxidans* str. ATCC 53993 were removed (along with *Sulfurovum sp.* str. NBC37-1, *P. torridus* str. DSM 9790).

These three organisms were chosen to represent different 'amounts' of database bias. *Acidithiobacillus ferrooxidans* str. ATCC 53993 is from a clade (*Gammaproteobacteria*) that is generally well represented in the nr database. The nr database we used includes complete genome sequences from two strains of *At. ferrooxidans* and one of *At. caldus*. *Picrophilus torridus* str. DSM 9790 is from the archaeal clade *Thermoplasmatales* that includes complete genome sequences of multiple members of the genera *Thermoplasma* and *Ferroplasma*. *Sulfurovum sp.* str. NBC37-1 is the only member with a complete genome sequence from the *Sulfurovumales* clade (*Epsilonproteobacteria*). Therefore, we predict that the modified nr database should be most biased against *Sulfurovum* and least biased against *At. ferrooxidans* str. ATCC 53993. Likewise, in the RS24 metagenome, we expect bias to be strongest against *Acidimicrobium*, and increasingly less for the G-plasma and *Acidithiobacillus* populations (Figures S4, S5, and S6).

Results from these simulations are summarized in Table S3. Results from simulations show that for the MEGAN parameters we used for binning (40 bit score threshold, 1 min support, 10% top percent), false assignment rates were less than 5% for all analyses. When sequences from *P. torridus*, *Sulfurovum sp.* str. NBC37-1, and *At. ferrooxidans* strains were removed from the nr database, only between 48.1 and 55.2% of the simulated datasets and 53.6% of simulated metagenome matched nr above bit score 40.

Calculation of overrepresented functions

In order to determine functions that are overrepresented in the RS24 metagenome, we compared COG assignments for RS24 against that of eight publicly available metagenomes

generated with GS20 pyrosequencing technology. A maximum of 200,000 reads were used from each metagenome. Reads for other metagenomes were assigned to COGs exactly as for RS24: first, identical sequence copies, if present, were removed prior to analysis. Nucleotide sequences from each metagenome were compared against the NCBI nr database with BLASTX, and significant matches to nr were determined using the same criteria as for RS24. COGs were assigned by comparing all nucleotide sequences that significantly matched nr against the COG database with RPSBLAST. Although we did not remove rRNA sequences from these other metagenomes prior to analysis, no rRNA sequences were mistakenly translated and assigned to COGs when these same criteria were used for RS24 COG assignments.

To determine overrepresented functions, the COG assignments for each metagenome were organized into a matrix of COG assignments (*j* columns) by sample (*i* rows). Raw numbers of reads assigned to each COG were first converted to proportions (X_{ij}) of the total COG assignments for each metagenome (row totals). Then, each matrix element X_{ij} were standardized to zero mean and unit variance (Z_{ij}; 'z-score') by equation S3 (McCune and Grace, 2002):

$$\text{(S3)} \qquad Z_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j}$$

where $\overline{X}_j$ and S_j are the average and standard deviation of each column. COGs with positive Z_{ij} scores in a particular metagenome are overrepresented in that metagenome, and COGs with negative Z_{ij} scores are underrepresented (Figure S14, S15).

Analyses were performed using the *decostand* function available with the vegan package (Oksanen et al., 2007) in R version 2.5.1 (R core development team, 2007; http://www.r-project.org/). For this analysis, we used several other publicly available metagenomes generated by GS20 pyrosequencing technology: a marine sediment metagenome (sample "1 original") from Biddle et al. (2008), a metagenome from "red" environments from the Soudan iron mine in Minnesota (Edwards et al., 2006), and several metagenomes from Dinsdale et al. (2008): Highborne Cay stromatolite, Line Islands Tabuaeran and Kingman Reef, Salton Sea, cow rumens 80F6, and healthy striped bass gut. RS24 sequences were only compared against other unassembled GS20 metagenomes in order to maintain consistency among analyses and avoid biases due to gene size (i.e. larger genes have more reads in unassembled data compared to assembled data).

Supplementary references

Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Berg I, Ramos-Vera W, Petri A, Huber H, Fuchs G (2010). Study of the distribution of autotrophic CO2 fixation cycles in *Crenarchaeota*. *Microbiology* **156:** 256-269.

Biddle J, Fitz-Gibbon S, Schuster S, Brenchley J, House C (2008). Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci USA* **105:** 10583-10588.

Bond P, Smriga S, Banfield J (2000). Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. *Appl Environ Microbiol* **66:** 3842-3849.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72:** 5069-5072.

Dinsdale E, Edwards R, Hall D, Angly F, Breitbart M, Brulc J *et al* (2008). Functional metagenomic profiling of nine biomes. *Nature* **452:** 629-632.

Dlugokencky E, Bruhwiler L, White J, Emmons L, Novelli P, Montzka S *et al* (2009). Observational constraints on recent increases in the atmospheric CH_4 burden. *Geophys Res Lett* **36:** L18803.

Dlugokencky E, Myers R, Lang P, Masarie K, Crotwell A, Thoning K *et al* (2005). Conversion of NOAA atmospheric dry air CH_4 mole fractions to a gravimetrically prepared standard scale. *Journal of Geophysical Research* **110:** D18306.

Druschel G, Baker B, Gihring T, Banfield J (2004). Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochem Trans* **5:** 13-32.

Edwards R, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson D *et al* (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7:** 57.

Hopmans E, Schouten S, Pancost R, van der Meer M, Damstè J (2000). Analysis of intact tetraether lipids in archaeal cell material and sediments by high performance liquid chromatography/atmospheric pressure chemical ionization mass spectrometry. *Rapid Communications in Mass Spectrometry* **14:** 585-589.

Huber T, Faulkner G, Hugenholtz P (2004). Bellerophon; a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20:** 2317-2319.

Hügler M, Huber H, Molyneaux S, Vetriani C, Sievert S (2007). Autotrophic CO2 fixation via the reductive tricarboxylic acid cycle in different lineages within the phylum *Aquificae*: evidence for two ways of citrate cleavage. *Environ Microbiol* **9:** 81-92.

Huson D, Auch A, Qi J, Schuster S (2007). MEGAN analysis of metagenomic data. *Genome Res* **17:** 377-386.

Klatt C, Bryant D, Ward D (2007). Comparative genomics provides evidence for the 3-hydroxypropionate autotrophic pathway in filamentous anoxygenic phototrophic bacteria and in hot spring microbial mats. *Environ Microbiol* **9:** 2067-2078.

Lane DJ (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systematics*. Wiley: New York. pp 115-175.

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.

Macalady JL, Dattagupta S, Schaperdoth I, Jones DS, Druschel GK, Eastman D (2008). Niche differentiation among sulfur-oxidizing bacterial populations in cave waters. *ISME J* **2:** 590-601.

Macalady JL, Jones DS, Lyon EH (2007). Extremely acidic, pendulous microbial biofilms from the Frasassi cave system, Italy. *Environ Microbiol* **9:** 1402-1414.

Marchler-Bauer A, Anderson J, Derbyshire M, DeWeese-Scott C, Gonzales N, Gwadz M *et al* (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **35:** D237-240.

Marchler-Bauer A, Panchenko A, Shoemaker B, Thiessen P, Geer L, Bryant S (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* **30:** 281-283.

Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L *et al* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Markowitz V, Ivanova N, Szeto E, Palaniappan K, Chu K, Dalevi D *et al* (2007). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36:** D534-538.

McCune B, Grace JB (2002). *Analysis of Ecological Communities*. MjM Software Design: Gleneden Beach, OR.
Oksanen J, Kindt R, Legendre P, O'Hara R, Simpson GL, Stevens MHH (2008). Vegan: Community Ecology Package, R package version 1.11-0.
Pearson A, Page S, Jorgenson T, Fischer W, Higgins M (2007). Novel hopanoid cyclases from the environment. *Environ Microbiol* **9:** 2175-2188.
Posada D, Crandall K (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.
Pruesse E, Quast C, Knittel K, Fuchs B, 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.
R Core Development Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria.
Ragsdale S, Pierce E (2008). Acetogenesis and the Wood-Ljungdahl pathway of CO2 fixation. *Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics* **1784:** 1873-1898.
Raskin L, Stromley J, Rittmann B, Stahl D (1994). Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. *Appl Environ Microbiol* **60:** 1232-1240.
Ren Q, Kang K, Paulsen I (2004). TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* **32:** D284-D288.
Ronquist F, Huelsenbeck J (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.
Shively J, Van Keulen G, Meijer W (1998). Something from almost nothing: carbon dioxide fixation in chemoautotrophs. *Annu Rev Microbiol* **52:** 191-230.
Sokal R, Rohlf F (1996). Biometry: The Principles and Practice of Statistics in Biological Research. 3rd edn.: 865 p. WH Freeman & Co., New York.
Sowers T, Bernard S, Aballain O, Chappellaz J, Barnola J, Marik T (2005). Records of the $\delta^{13}C$ of atmospheric CH_4 over the last 2 centuries as recorded in Antarctic snow and ice. *Global Biogeochem Cy* **19:** GB2002.
Swofford DL (2000). PAUP*: Phylogenetic analysis using parsimony and other methods (software). Sinauer Associates: Sunderland, MA.
Talbot H, Squier A, Keely B, Farrimond P (2003). Atmospheric pressure chemical ionisation reversed-phase liquid chromatography/ion trap mass spectrometry of intact bacteriohopanepolyols. *Rapid Communications in Mass Spectrometry* **17:** 728-737.
Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E *et al* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.
Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P *et al* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43.
Whitaker R, Banfield J (2006). Population genomics in natural microbial communities. *Trends Ecol Evol* **21:** 508-516.
Zarzycki J, Brecht V, Müller M, Fuchs G (2009). Identifying the missing steps of the autotrophic 3-hydroxypropionate CO2 fixation cycle in *Chloroflexus aurantiacus*. *Proc Natl Acad Sci USA* **106:** 21317-21322.

Supplemental Figures and Tables

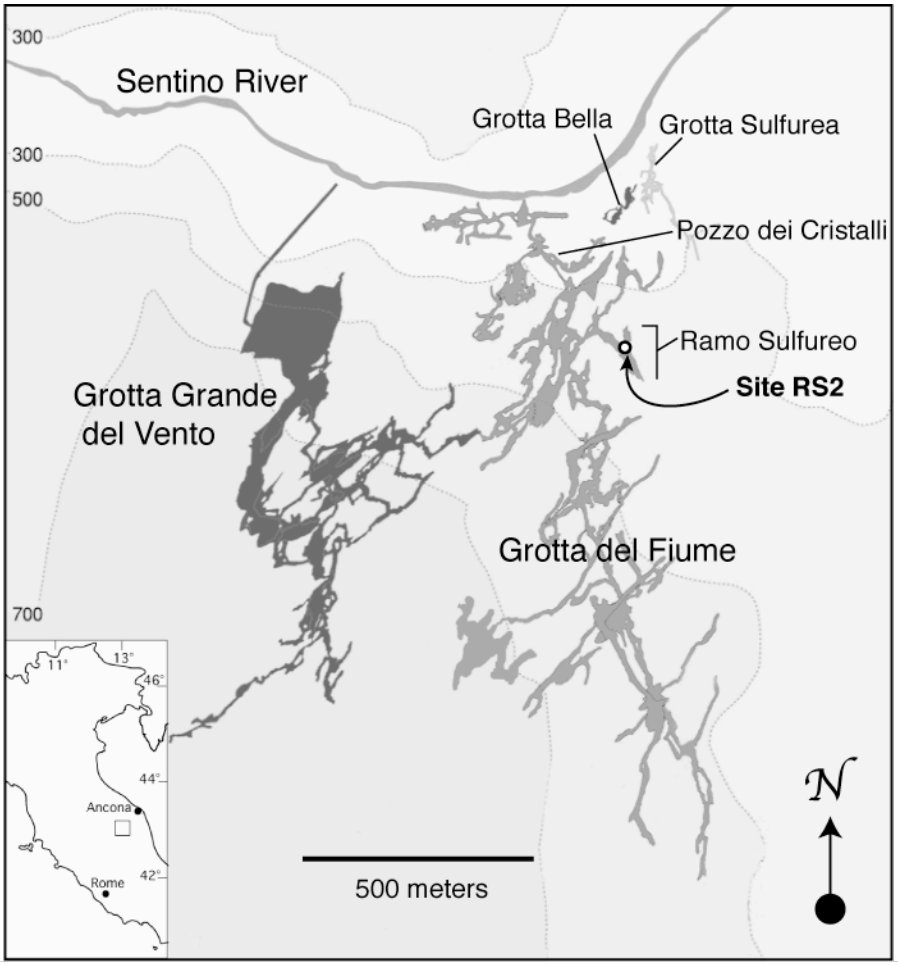


Figure S1. Map of the Grotta Grande del Vento-Grotta del Fiume (Frasassi) cave complex, with sample site RS2 indicated. Base map provided by the Gruppo Speleologico CAI di Fabriano.

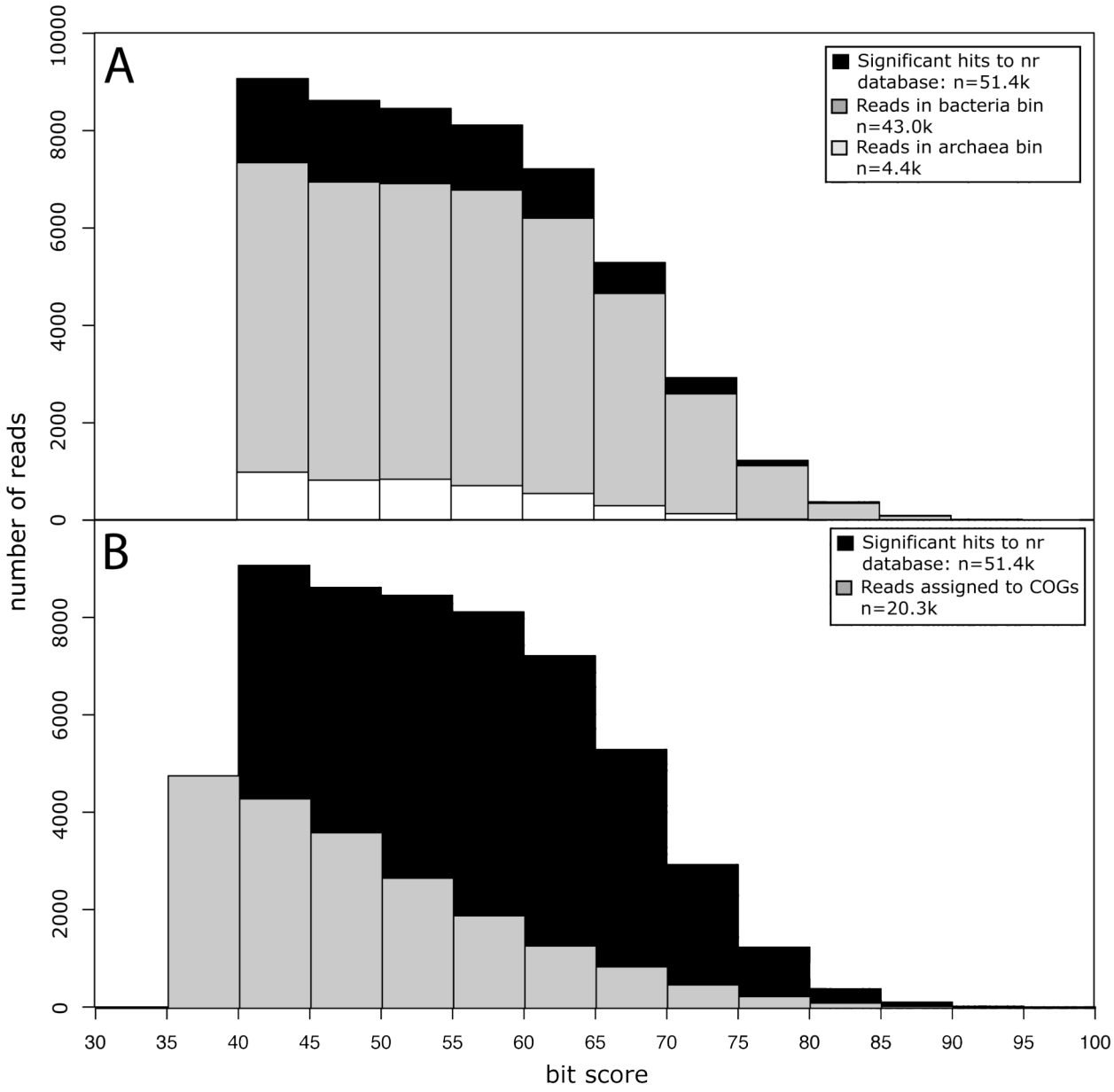


Figure S2. Histograms that depict the distribution of bit scores for BLAST analyses of metagenome reads. (A) Histogram from BLASTX analysis of all metagenome reads against the NCBI nr database, compared to similar BLASTX analyses of only those reads in the bacteria and archaea bins. Only reads matching the nr database at bit score above 40 were considered in metagenomic analyses. (B) Histogram from BLASTX analysis of all metagenome reads against the NCBI nr database, compared to RPS-BLAST of all reads against the COG database. For RPS-BLAST, only reads matching the COG database at bit scores above 35 were considered.

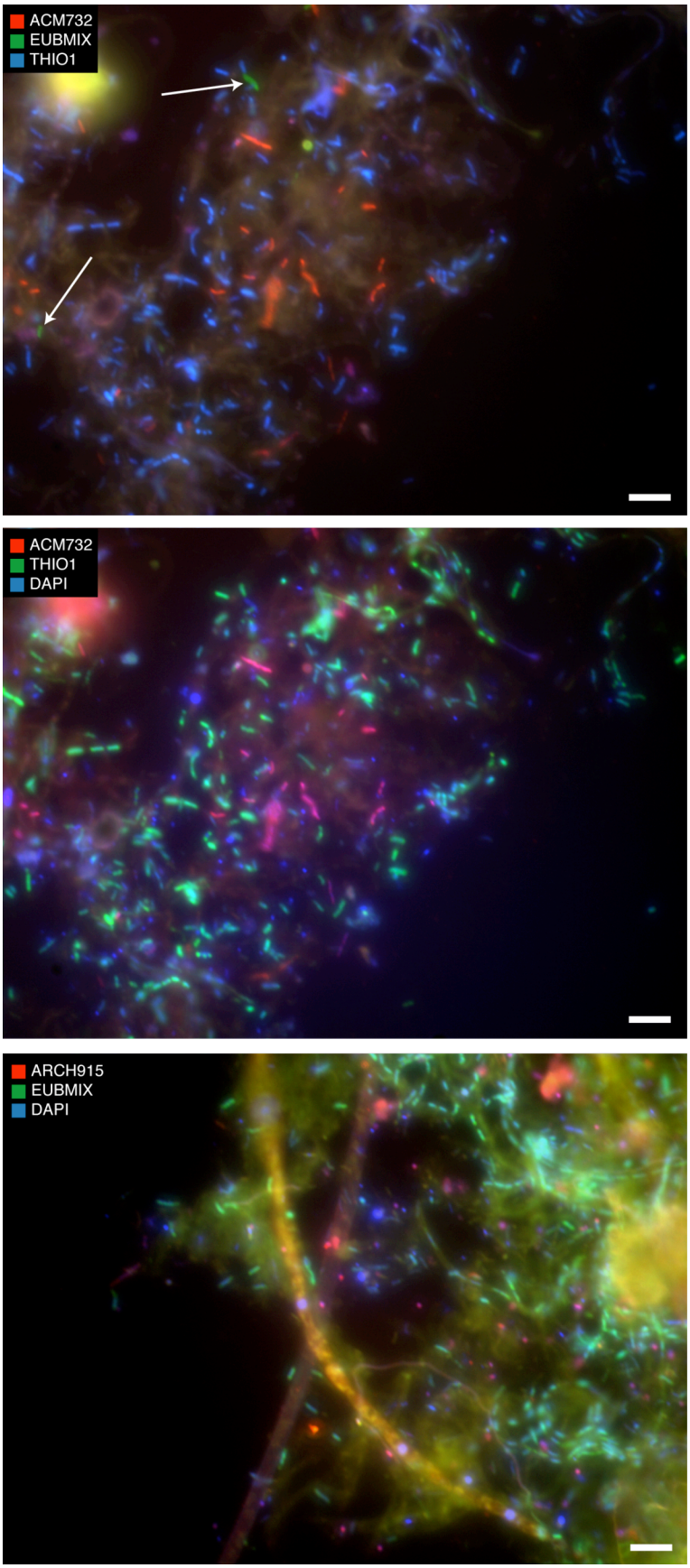


Figure S3. Fluorescent *in situ* hybridization (FISH) photomicrographs of sample RS24. Scale bars are 5 μm, and FISH probes used are listed in the upper left. FISH probe specificity is as follows: EUBMIX = most bacteria, ARCH915 = most archaea, THIO1 = all *Acidithiobacillus*, ACM732 = all *Acidimicrobiaceae*. White arrows indicate bacteria not labeled by ACM732 or THIO1, which make up 1% of total cells. Complete information on FISH probes is published in Macalady et al. (2007).

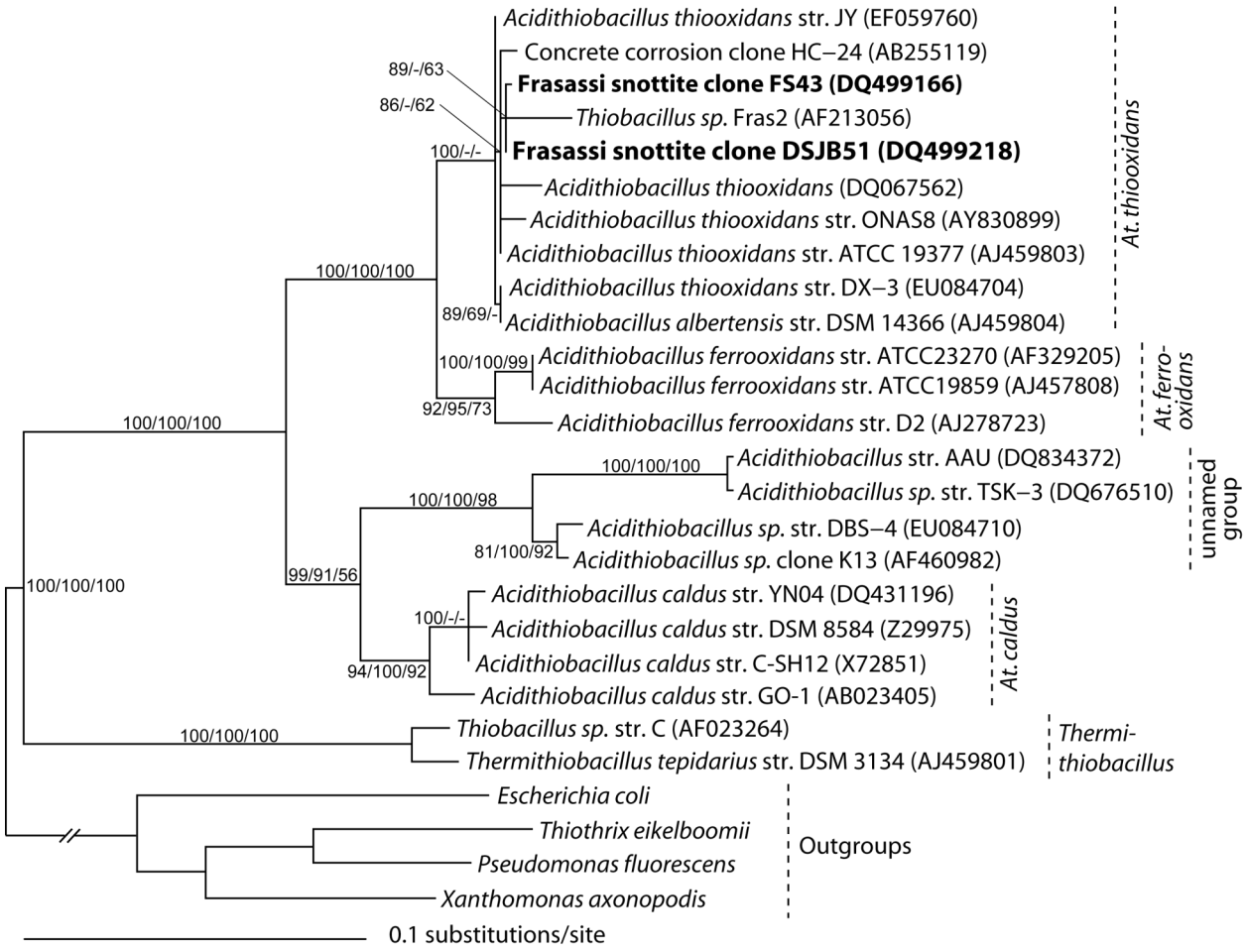


Figure S4. Maximum likelihood phylogram of 16S rRNA sequences from the genus *Acidithiobacillus* in the *Gammaproteobacteria*. Bootstrap values from Bayesian, neighbor joining, and maximum parsimony analyses (in that order) are shown for each node. Frasassi snottite clones, sequenced by Macalady et al. (2007), are shown in bold.

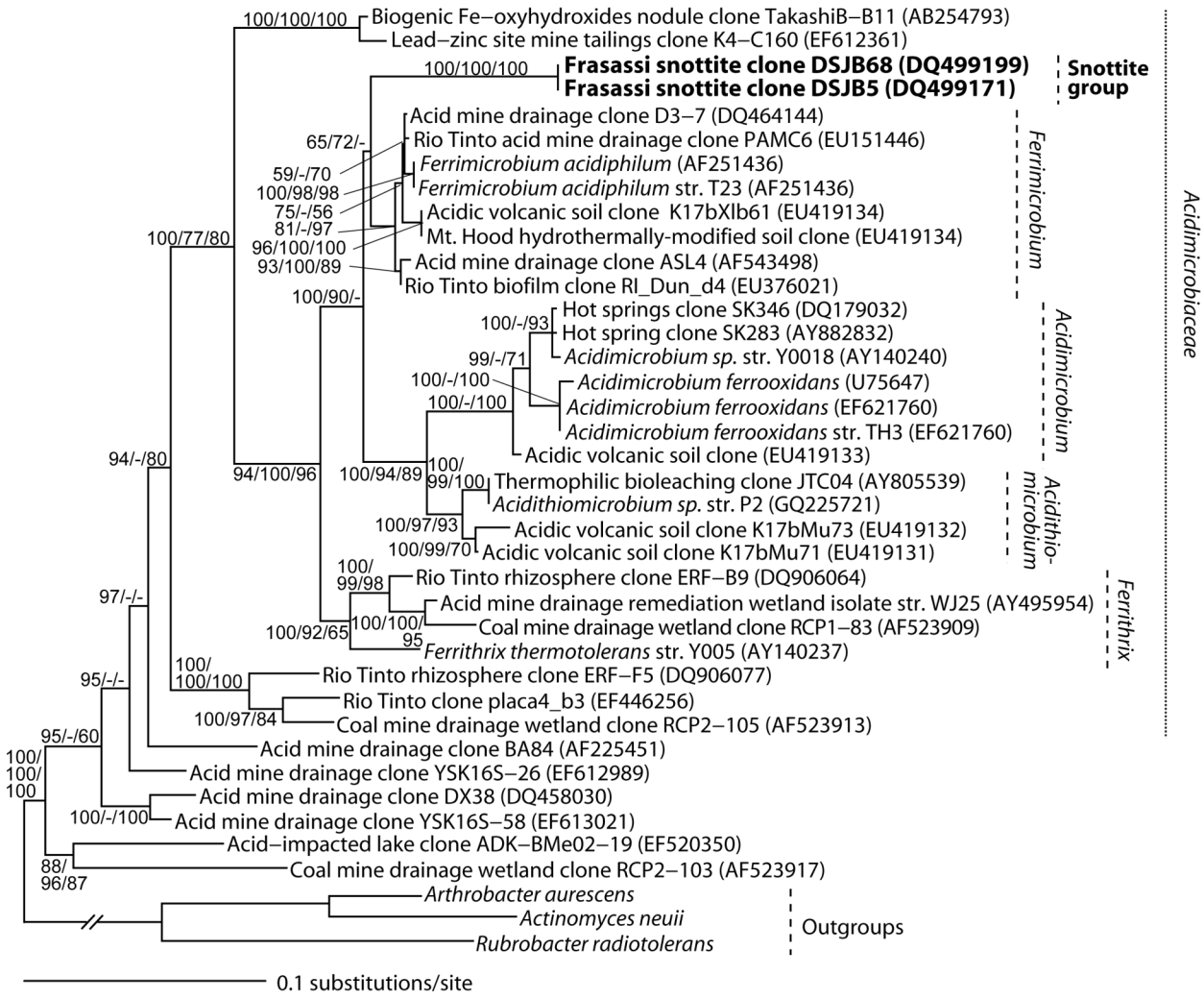


Figure S5. Maximum likelihood phylogram of 16S rRNA sequences from the *Acidimicrobiaceae* family in the *Actinobacteria*. Bootstrap values from Bayesian, neighbor joining, and maximum parsimony analyses (in that order) are shown for each node. Frasassi snottite clones, sequenced by Macalady et al. (2007), are shown in bold.

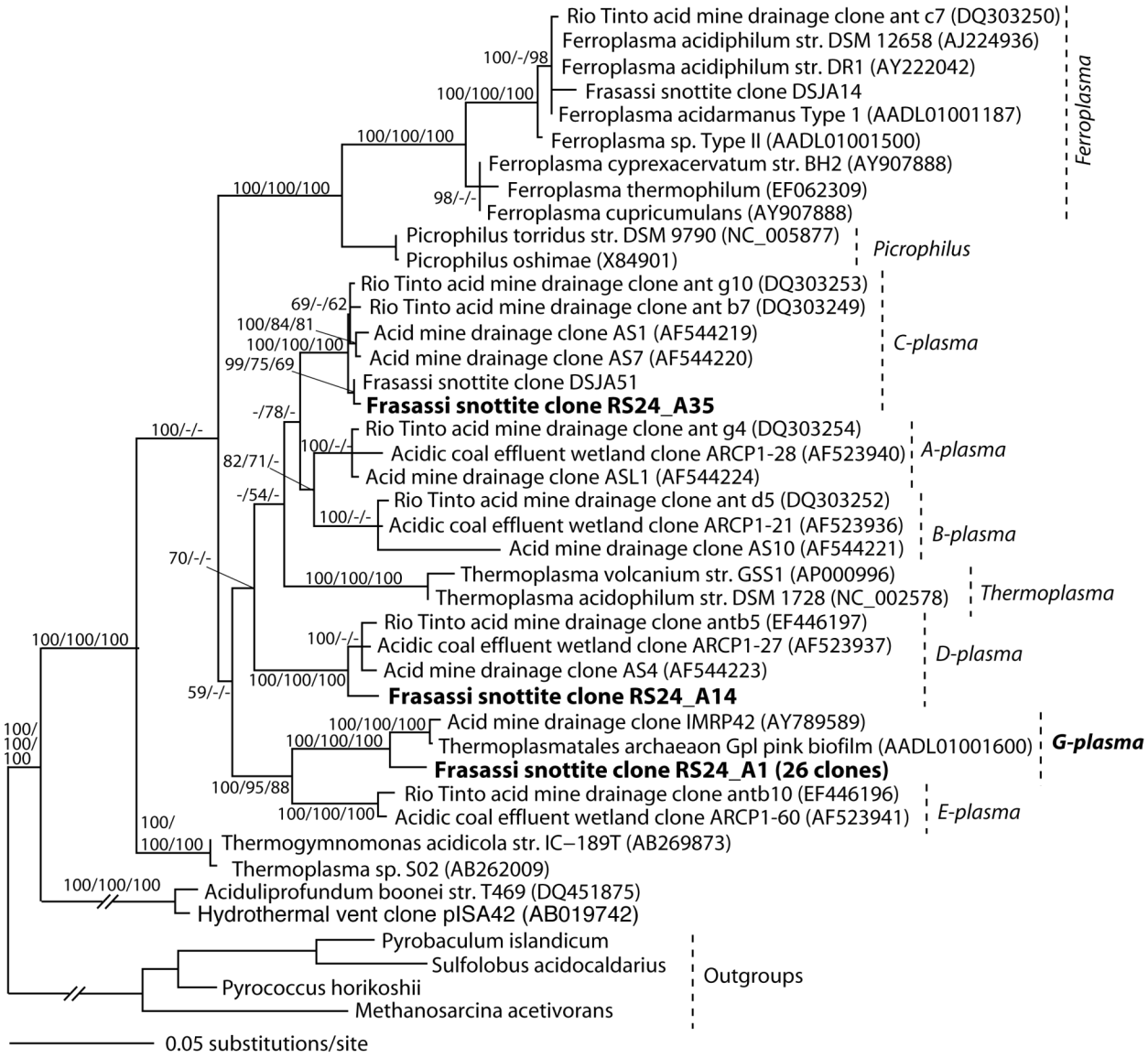


Figure S6. Maximum likelihood phylogram of 16S rRNA sequences from the *Thermoplasmatales* family in the *Euryarchaeota*. Bootstrap values from Bayesian analysis, neighbor joining, and maximum parsimony analyses (in that order) are shown for each node. Frasassi snottite clones from this study are shown in bold, and numbers in parentheses indicate the number of clones sequenced. Other Frasassi snottite clones sequenced by Macalady et al. (2007) are included in the tree. Group names for the 'alphabet plasma' clades are from Druschel et al. (2004).

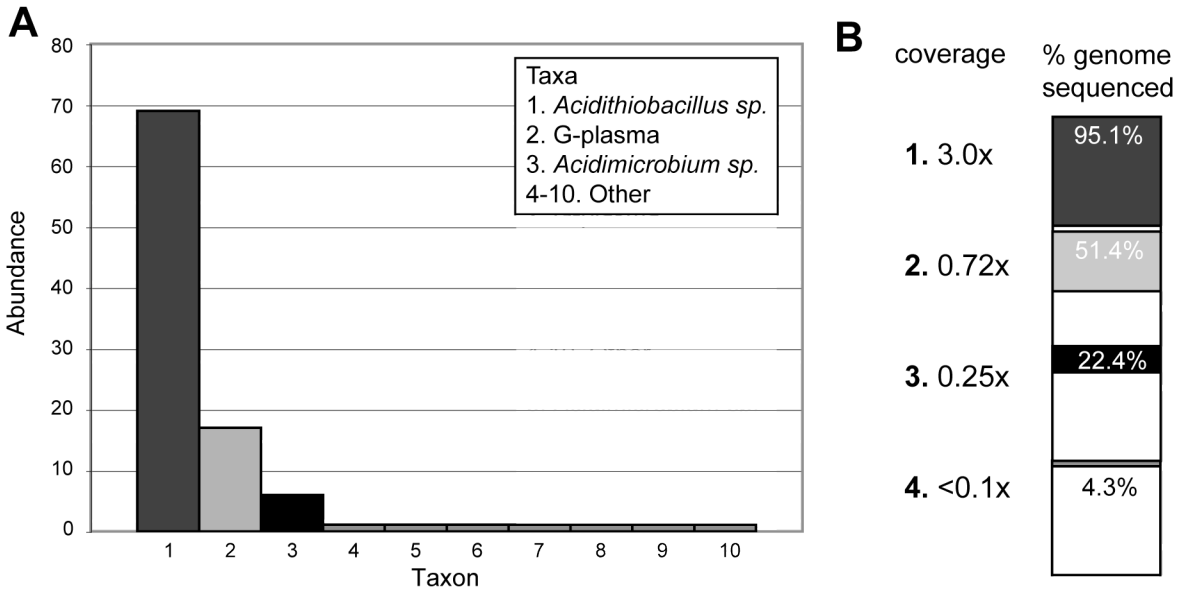


Figure S7. Expected metagenomic sequence coverage of taxa in the RS24 sample. (A) Community composition of RS24 based on FISH cell counts. 'Other' represents low-abundance organisms that compose the 'other Bacteria' category in Figure 2f. (B) For each taxon in A, numbers represent the average genome coverage, and bars represent the fraction of the total genome present in the metagenome, assuming a Poisson distribution of sequence coverage.

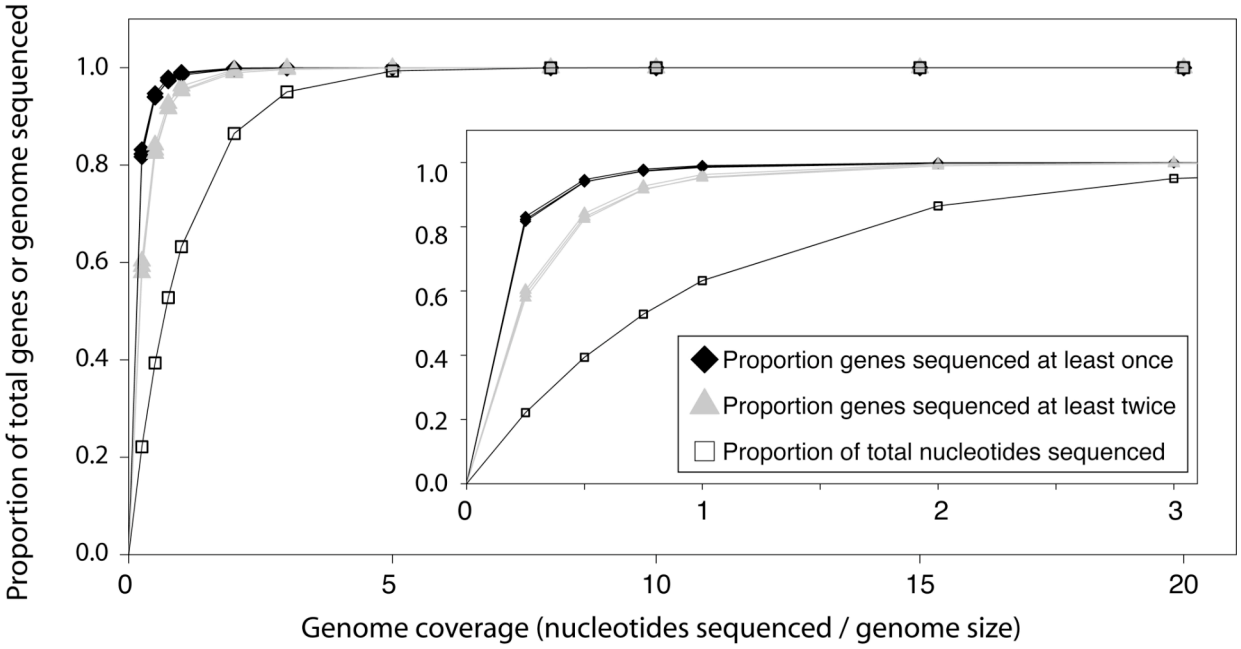


Figure S8. Results from simulated 454 GS20 sequencing showing the total percentage of genes and genome (total nucleotide positions) represented for different amounts of sequencing. More than 8x coverage is necessary to consistently sequence every nucleotide position in the genome, while only little more than 1x coverage is necessary to represent every gene with at least two reads. Results from simulated sequencing of three genomes of different sizes are shown: *Bdellovibrio bacteriovorus* str. HD100, *Thermoplasma acidophilum* str. DSM 1728, and *Thiomicrospira crunogena* str. XCL-2. Data points closely overlap. For each data point, standard deviations of the percentage of genes and nucleotides sequenced, based on 100 replicates, are smaller than the height of each point.

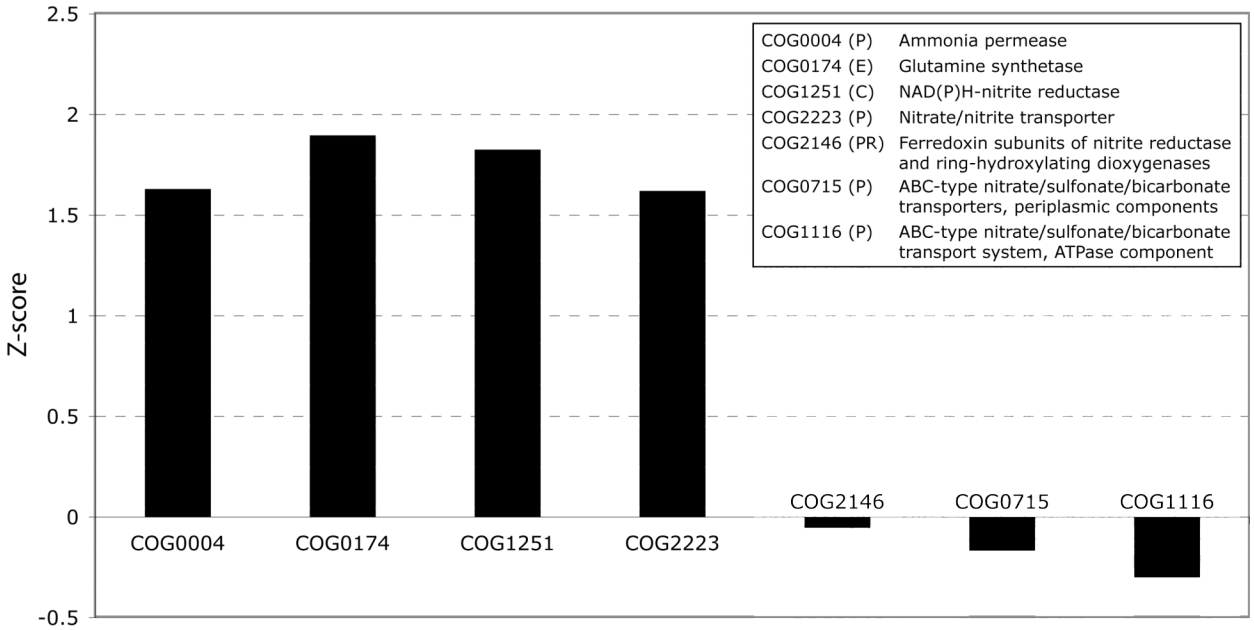


Figure S9. Standardized abundance scores of COGs involved in ammonia and nitrate uptake. The COG id (COG category) COG description is provided in the legend.

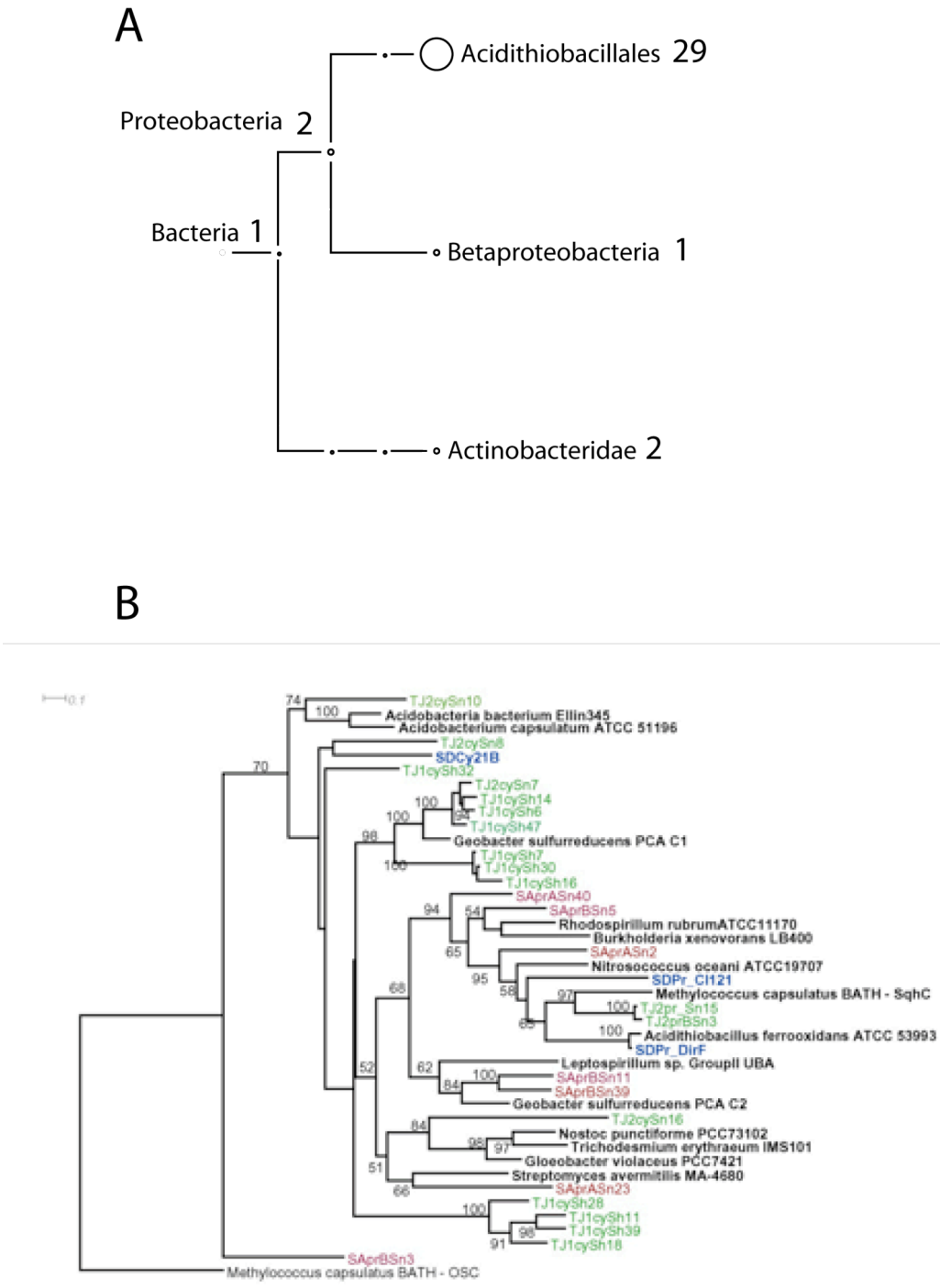


Figure S10. Taxonomic classification of squalene hopene cyclase (SHC) sequences in snottite sample RS24. (A) Taxonomic distribution of SHC homologues in the metagenome. Figure created in MEGAN (Husen et al., 2007). Numbers indicate reads assigned to each taxonomic group. (B) Maximum parsimony tree based on full-length SHC sequences. Bootstrap values (100 replicates) are shown for each node. Sequences in blue are from sample RS24. Included in the tree are environmental SHC sequences from Mishwam Lake (green) and Bahamas soil (red) (Pearson et al., 2007; Pearson et al., 2009).

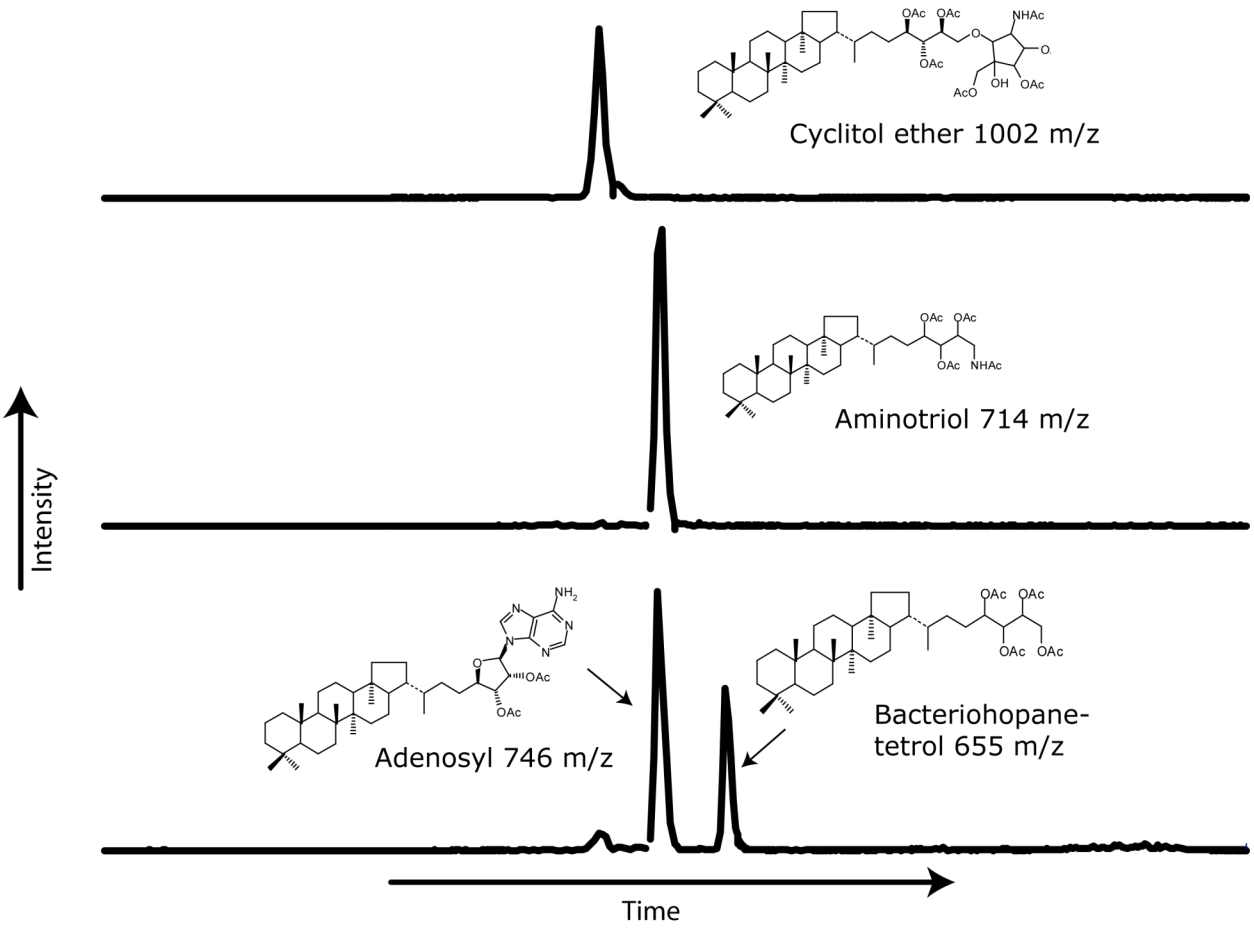


Figure S11. Extracted ion chromatograms showing hopanoid peaks in the snottite sample RS24 total lipid extract. Hopanoid structures corresponding to each peak are indicated.

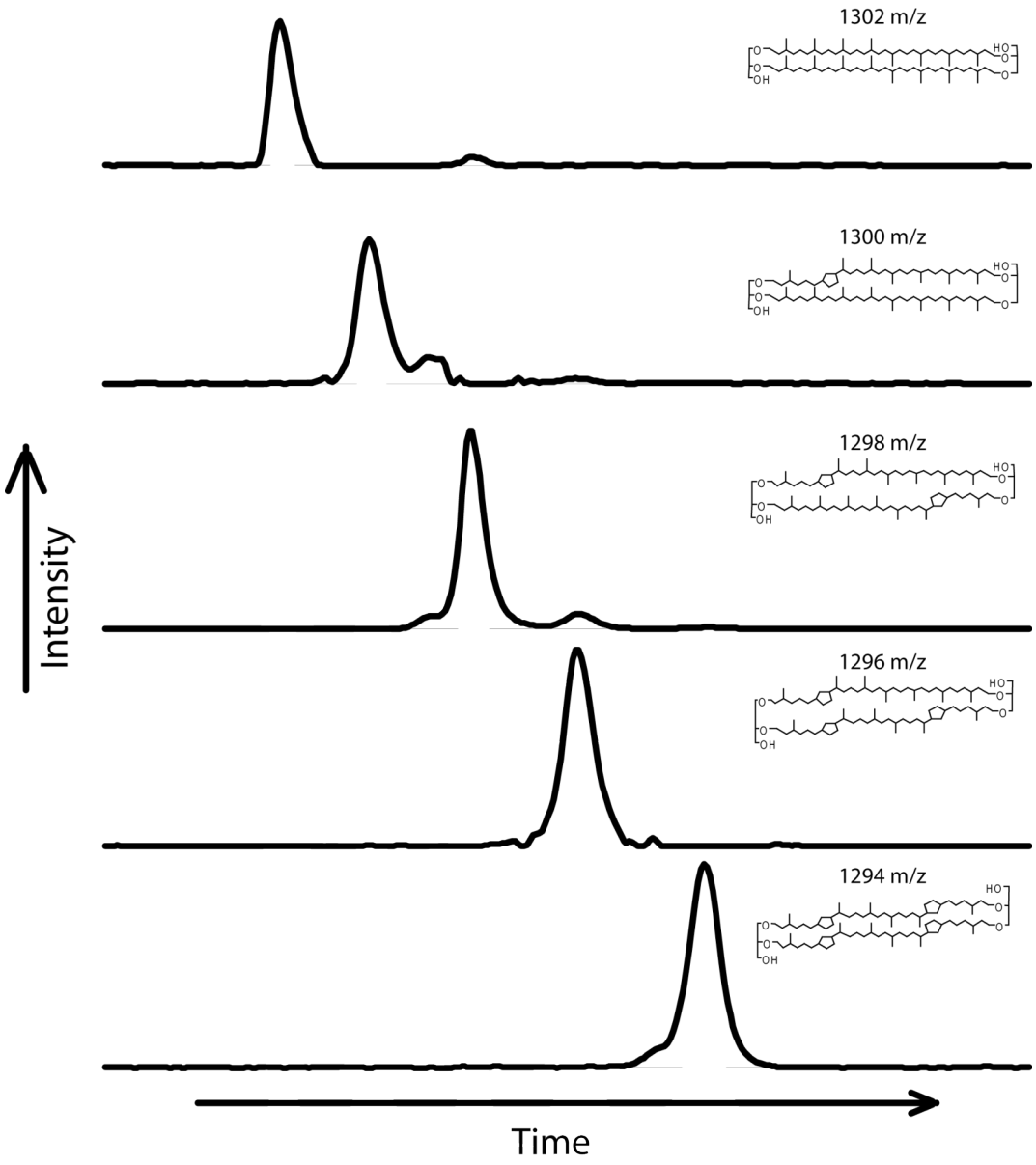


Figure S12. Extracted ion chromatograms showing isoprenoid glycerol dialkyl glycerol tetraether (GDGT) lipid peaks in the RS24 total lipid extract. GDGT structures corresponding to each peak are indicated.

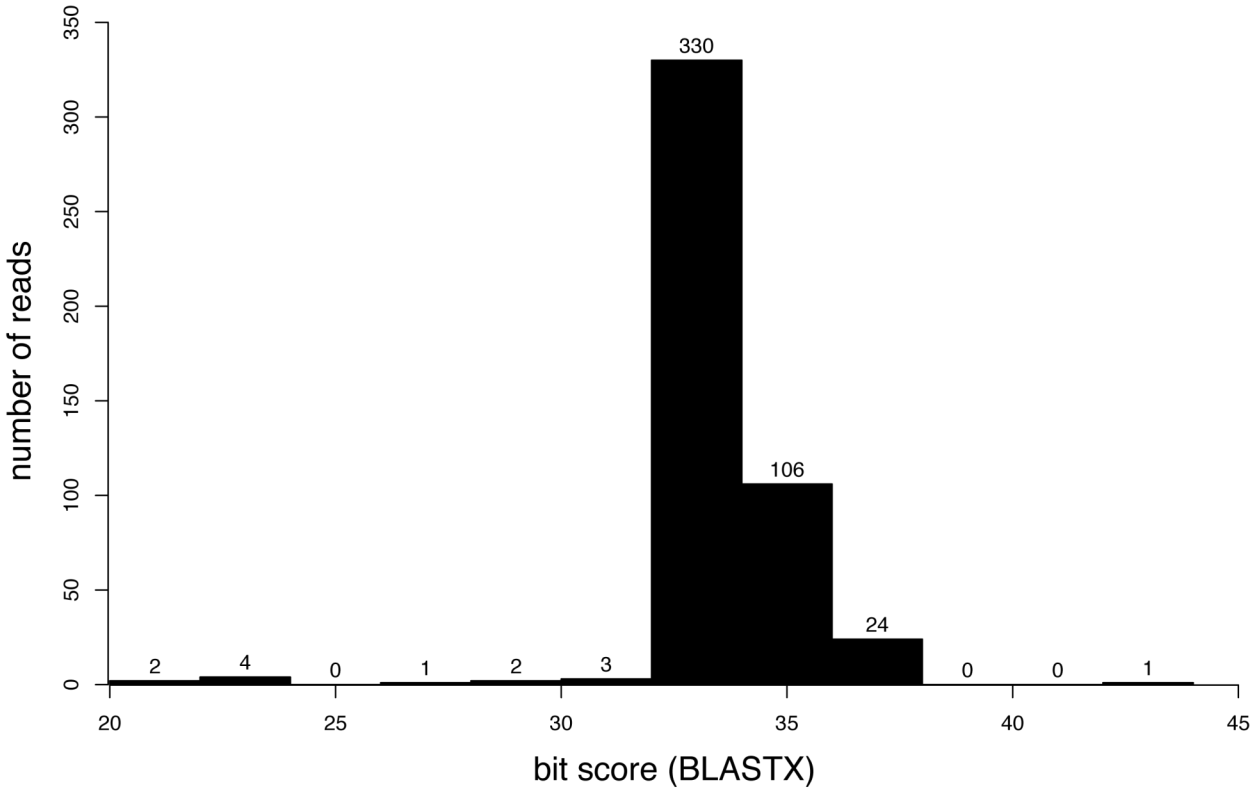


Figure S13. Results from BLASTX analysis of 11,000 randomized RS24 metagenome reads against the NCBI non-redundant (nr) database. Only 4.3% of randomized reads matched nr below the BLASTX default e-value of 10.

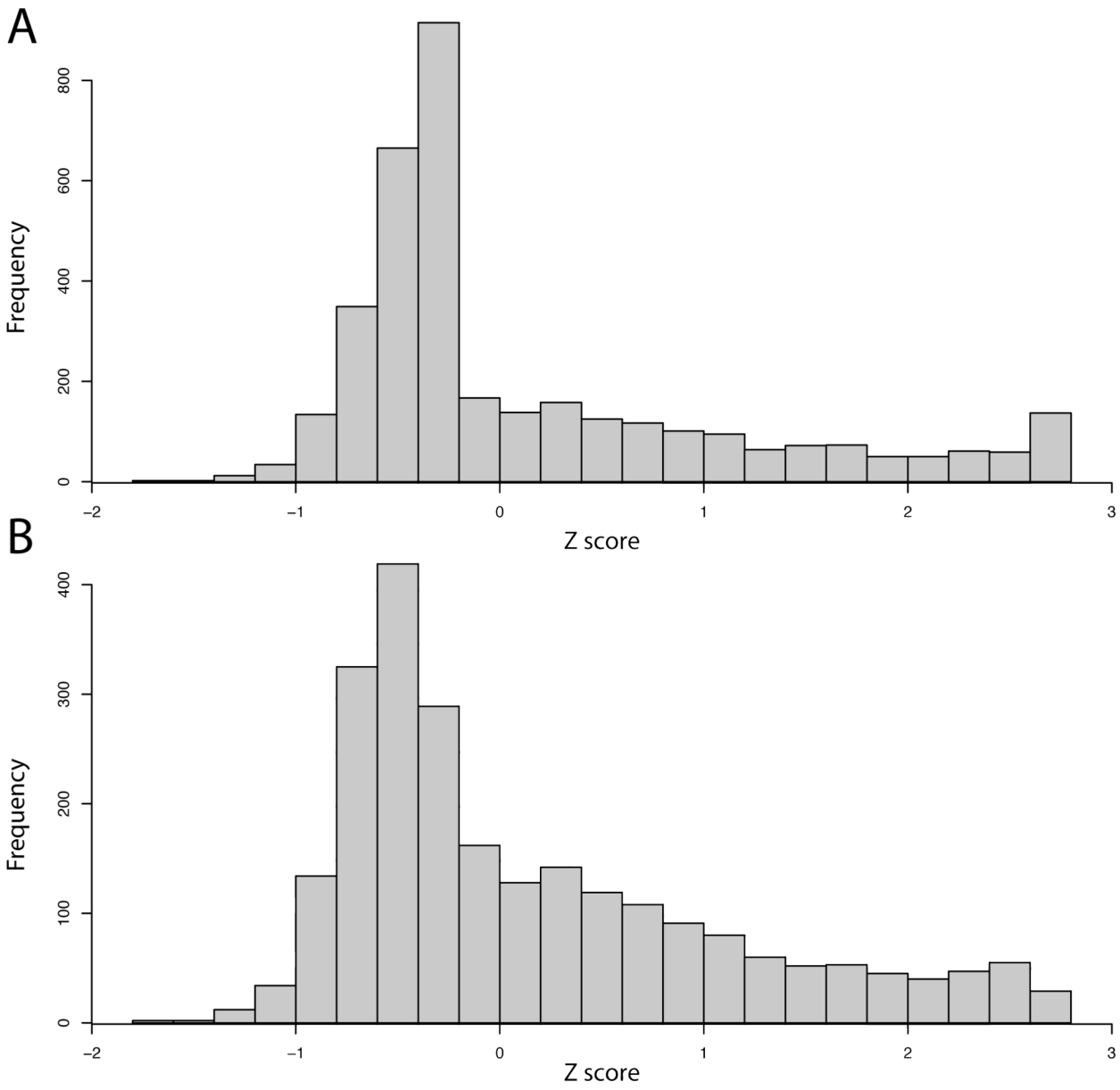
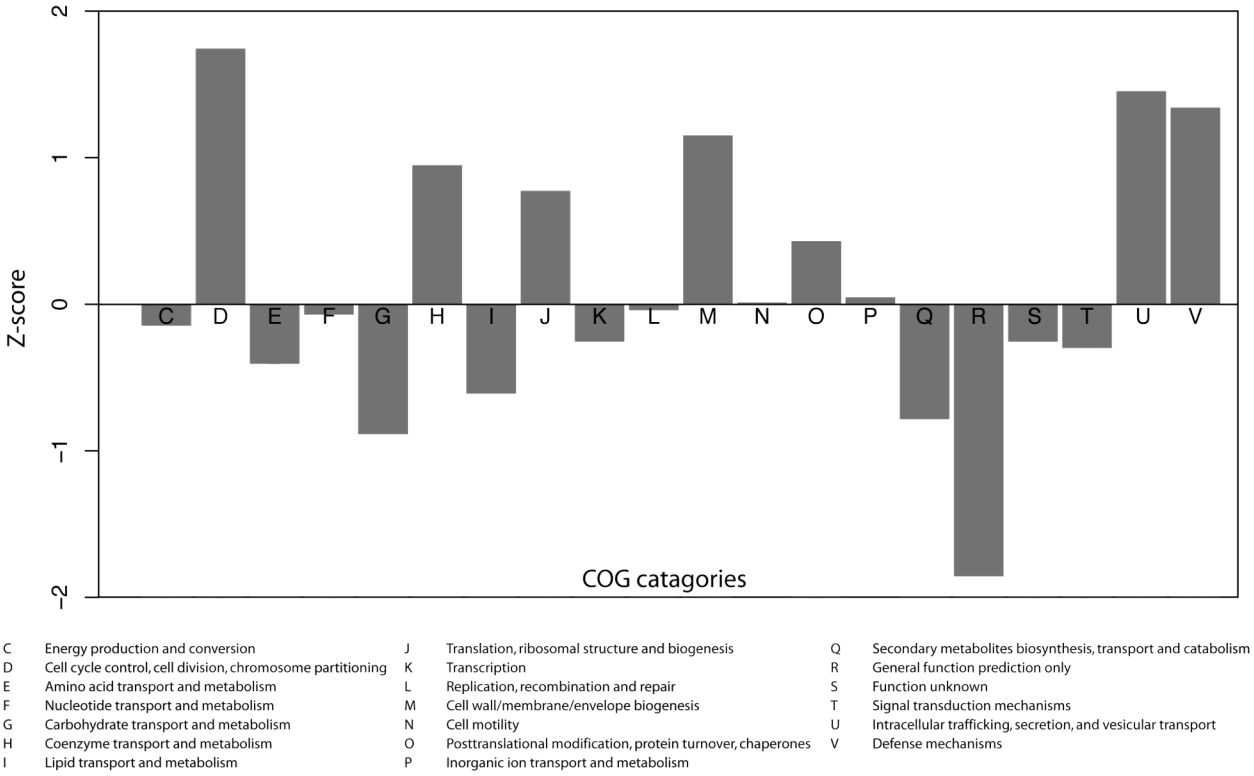


Figure S14. Histograms illustrating our z-score calculations of over- and underrepresented COG functions in the snottite metagenome. (A) Histogram including all COG categories present in the RS24 metagenome. (B) Histogram including only more abundant COG categories, defined as COG categories with at least five total reads assigned (sum of total reads assigned for all metagenomes used in the z-score calculation).



C	Energy production and conversion	J	Translation, ribosomal structure and biogenesis	Q	Secondary metabolites biosynthesis, transport and catabolism
D	Cell cycle control, cell division, chromosome partitioning	K	Transcription	R	General function prediction only
E	Amino acid transport and metabolism	L	Replication, recombination and repair	S	Function unknown
F	Nucleotide transport and metabolism	M	Cell wall/membrane/envelope biogenesis	T	Signal transduction mechanisms
G	Carbohydrate transport and metabolism	N	Cell motility	U	Intracellular trafficking, secretion, and vesicular transport
H	Coenzyme transport and metabolism	O	Posttranslational modification, protein turnover, chaperones	V	Defense mechanisms
I	Lipid transport and metabolism	P	Inorganic ion transport and metabolism		

Figure S15. Z-scores of over- and underrepresented COG categories in the snottite metagenome. Corresponds to Figure 3a in the main text.

Table S1. Genes for the breakdown of organic compounds in the *Acidimicrobiaceae* bin

No. of reads	*Enzyme name*	*Function*
4	Enoyl-CoA hydratase	fatty-acid oxidation
2	Extradiol ring-cleavage dioxygenase	aromatic compound breakdown
2	Homogentisate 1,2-diooxygenase	amino acid degredation
2	Chitinase	chitin breakdown
1	Phenylacetate-CoA oxygenase	amino acid degredation
1	Diterpenoid dioxygenase	growth on terpenoids
1	Haloalkane dehalogenase	breakdown of halogenated compounds
2	5-carboxymethyl-2-hydroxymuconate delta-isomerase	amino acid degredation
1	Pentachlorophenol monooxygenase	breakdown of phenols
3	alpha-amylase	polysaccharide breakdown
2	beta-glucosidase	polysaccharide breakdown
2	beta-galactosidase	polysaccharide breakdown
1	alpha glucosidase	polysaccharide breakdown
2	Alpha-galactosidase	polysaccharide breakdown
1	alpha-L-fucosidase	polysaccharide breakdown

Table S2. Number of false COG assignments using simulated datasets

Analysis	false assignments/total COG assignments	
	sim 1 (*A. ferrooxidans)*	sim 2 (*B. bacteriovorus*)
blastp	80/1852	48/928
translated dataset	4.32%	5.17%
blastx	81/1717	48/938
original dataset	4.72%	5.12%
rpsblast	31/1143	28/670
translated dataset	2.71%	4.17%
rpsblast	46/1578	39/942
original dataset	2.92%	4.14%

Table S3. Assignment and binning rates of simulated datasets

	Original nr dataset (all sequences included)		**Modified nr 1** (sequences from *P. torridus* str. DSM 9790, *Sulfurovum sp.* str. NBC37-1, and all *At. ferrooxidans* strains removed prior to analysis)		**Modified nr 2** (sequences from *P. torridus* str. DSM 9790, *Sulfurovum sp.* str. NBC37-1, and only *At. ferrooxidans* str. ATCC 53993 removed prior to analysis)	
Bit score cutoff[a]	35	40	35	40	35	40
***Acidithiobacillus ferrooxidans* ATCC 53993**						
% matching to nr	89.4	86.3	64.6	55.2	74.9	67.4
% Gammaproteobacteria	82.6	80.3	25.6	22.7	44.9	41.8
% false assignments (other phyla)[b]	0.4	0.2	6.2	4.7	4.0	3.1
% Proteobacteria	86.1	83.4	43.6	38.5	58.6	54.1
% Bacteria	88.1	83.4	57.9	50.0	69.6	63.2
***Picrophilus torridus* DSM 9790**						
% matching to nr	88.5	85.8	58.4	48.1	-	-
% Thermoplasmatales	87.0	84.6	37.1	31.9	-	-
% false assignments (other phyla)[b]	0.2	0.1	4.9	3.4	-	-
% Euryarchaeota	87.5	85.0	40.5	34.8	-	-
% Archaea	88.0	85.4	47.6	40.3	-	-
***Sulfurovum sp.* str. NBC37-1**						
% matching to nr	88.7	86.0	61.0	52.1	-	-
% Epsilonproteobacteria	85.3	83.0	37.2	32.6	-	-
% false assignments (other phyla)[b]	0.2	0.2	3.7	2.9	-	-
% Proteobacteria	86.4	84.0	44.1	38.5	-	-
% Bacteria	88.3	85.7	57.2	49.0	-	-
Simulated metagenome[c]						
% matching to nr	89.2	85.8	63.2	53.6	70.7	62.3
% Gammaproteobacteria, Thermoplasmatales, & Epsilonproteobacteria	83.7	81.0	29.6	26.5	43.3	39.9
% false assignments (other phyla)[b]	0.4	0.1	5.7	3.9	4.0	2.8

[a]Other MEGAN parameters are min support = 1 and top percent = 10%

[b]These are assignments outside the Gammaproteobacteria, Thermoplasmatales, or Epsilonproteobacteria.

[c]Simulated metagenome is composed of 75% sequences from *At. ferrooxidans* ATCC 53993, 18.5% *P. torridus* DSM 9790, and 6.5% *Sulfurovum sp.* str. NBC37-1)