Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla

Joanne B. Emerson,[1*†] Brian C. Thomas,[1] Walter Alvarez[1] and Jillian F. Banfield[1,2]

Departments of [1] Earth and Planetary Science and [2] Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA 94720-4767, USA

Summary

Research on geologic carbon sequestration raises questions about potential impacts of subsurface microbiota on carbon cycling and biogeochemistry. Subsurface, high-CO_2 systems are poorly biologically characterized, partly because of difficulty accessing high-volume, uncontaminated samples. CO_2-driven Crystal Geyser (CG, Utah, USA), an established geologic carbon sequestration analogue, provides high volumes of deep (~ 200–500 m) subsurface fluids. We explored microbial diversity and metabolic potential in this high-CO_2 environment by assembly and analysis of metagenomes recovered from geyser water filtrate. The system is dominated by neutrophilic, iron-oxidizing bacteria, including 'marine' *Mariprofundus* (*Zetaproteobacteria*) and 'freshwater' *Gallionellales*, sulfur-oxidizing *Thiomicrospira crunogena* and *Thiobacillus*-like *Hydrogenophilales*. Near-complete genomes were reconstructed for these bacteria. CG is notably populated by a wide diversity of bacteria and archaea from phyla lacking isolated representatives (candidate phyla) and from as-yet undefined lineages. Many bacteria affiliate with OD1, OP3, OP9, PER, ACD58, WWE3, BD1-5, OP11, TM7 and ZB2. The recovery of nearly 100 genes encoding ribulose-1,5 bisphosphate carboxylase-oxygenase subunit proteins of the Calvin cycle and AMP salvage pathways suggests a strong biological role in high-CO_2 subsurface carbon cycling. Overall, we predict microbial impacts on subsurface biogeochemistry via iron, sulfur, and complex carbon oxidation, carbon and nitrogen fixation, fermentation, hydrogen metabolism, and aerobic and anaerobic respiration.

Introduction

The capture and subsurface storage of anthropogenic CO_2, known as geologic carbon sequestration, has been proposed as a means of mitigating human-induced climate change. Research efforts are underway to assess the viability of this process, including evaluating the engineering feasibility of capture and storage, as well as characterizing and modelling geological and geochemical processes in the subsurface to ensure that carbon remains stored for a minimum of 1000 years (Schrag, 2007). Microbial communities have been found in a variety of terrestrial deep subsurface systems (e.g. Szewzyk *et al*., 1994; Boone *et al*., 1995; Chandler *et al*., 1998; Kovacik *et al*., 2006; Feng *et al*., 2007; Sahl *et al*., 2008; Brown and Balkwill, 2009), and

a recent biogeochemical modelling study suggested that the energetics in geologic carbon sequestration reservoirs can be conducive to microbial metabolism under certain conditions (West *et al*., 2011). While the 'Microbial Carbon Pump' is an established concept for marine open ocean environments (Jiao *et al*., 2010), relatively few geologic carbon sequestration studies have considered biological effects on subsurface sequestration.

Biological data related to geologic CO_2 sequestration have been reported or predicted from studies of isolates under high-CO_2 conditions and/or through geochemical modelling (e.g. Mitchell *et al*., 2010; West *et al*., 2011); however, data at the level of microbial communities in these systems is quite limited. DNA fingerprinting methods applied to samples collected from the Ketzin CO_2SINK pilot site in Germany have demonstrated the resilience of most naturally occurring subsurface populations after long-term exposure to CO_2 sequestration conditions in the laboratory (Wandrey *et al*., 2011a,b). At the same site, after deep subsurface CO_2 injection, *in situ* cell counts recovered to pre-injection levels after 5 months and the number of active microbial cells increased from one fourth to three fourths of the total number of cells, indicating a strong microbial response to high CO_2 conditions (Morozova *et al*., 2010; 2011). A 16S rRNA gene amplicon survey of 1.4 km-deep samples from the Otway Basin geologic carbon sequestration pilot site in Australia reported a shift in dominant bacteria from *Firmicutes* to *Proteobacteria* following supercritical CO_2 injection (Mu *et al*., 2014), and a new 16S rRNA gene amplicon study indicates an abundance of iron- and sulfur-oxidizing bacteria, including *Sideroxydans* and *Thiobacillus*, following the injection of CO_2 (mixed with H_2S and H_2) into the CarbFix pilot geologic carbon sequestration site in Iceland (Trias *et al*., 2014; B. Menez, pers. comm.). Although these analyses indicate that microorganisms and microbial consortia are capable of adapting to CO_2 sequestration conditions, further studies are needed to determine the diversity and metabolic potential of these microorganisms, to see how community composition varies across a range of geological and geochemical reservoir conditions, and to determine the long-term stability of these ecosystems.

Metagenomic analyses can yield insight into microbial diversity, phylogeny and metabolic potential in natural systems (e.g. Tyson *et al*., 2004; Venter *et al*., 2004). However, few metagenomic studies of the deep subsurface have been conducted, in part because of difficulties associated with accessing sufficiently large volumes of subsurface fluid for these analyses (for recent reviews of deep subsurface microbiology, see Edwards *et al*., 2012; Colwell and D'Hondt, 2013). An additional concern for subsurface microbial sample collection, including from CO_2 sequestration pilot sites, is the potential for contamination from the drilling process. Natural CO_2 sequestration analogue sites provide a means to mitigate some of these issues and are an excellent target for metagenomic studies.

Geysers, which rapidly catapult subsurface water to the surface, allow for easy access to the large volumes of water required for metagenomic analyses. Most geysers erupt hot water driven by expansion of steam, but a few cold-water geysers are driven by CO_2 degassing. Crystal Geyser (CG), located on the east bank of the Green River, about 6 km south of the town of Green River, Utah, is one such geyser. It is a cold (17°C), circumneutral, iron-rich geyser that erupts because of pressure from soluble and free-phase CO_2 accumulation in a subsurface aquifer, and it is an established natural analogue for geologic carbon sequestration (Shipton *et al.*, 2004; Bickle and Kampman, 2013). Although CO_2 has been leaking to the surface through springs in the vicinity of CG for ~400 000 years (Burnside *et al.*, 2013), CG came into existence after a potential oil trap, recognized from surface mapping, was tested for oil in 1935 with the Glen Ruby #1-X well (Shipton *et al.*, 2004). The well did not yield hydrocarbon, but instead pierced a reservoir holding CO_2-pressurized water that began erupting as a geyser.

CG eruption duration and frequency was studied over 76 days in 2005, during which the geyser erupted one to three times per day (average 1.8), with eruptions exhibiting a bimodal temporal pattern, lasting either 7–32 min or 98–113 min (Gouveia and Friedmann, 2006). This bimodal pattern was confirmed by Han and colleagues (2013) but was affected by drilling near the geyser in 2012 (Kampman *et al.*, 2014); drilling occurred after samples were collected for this study. Isotopic analyses indicate that the geyser water is primarily derived from the meteoric Navajo Aquifer (defined by Naftz *et al.*, 1997) at 200–500 m depth, with approximately 10% sourced from a deeper Paleozoic Aquifer (> 1500 m deep) that is in contact with the Paradox Salt Formation (Wilkinson *et al.*, 2009). These aquifers, particularly the Paleozoic Aquifer, are also the source of the geyser's salinity (~ 11 000–14 000 p.p.m). The CO_2 is primarily derived from crustal sources (i.e. the dissolution of limestones from the Paradox and/or Navajo Sandstone Formations), with a ~ 20% contribution from mantle degassing (Wilkinson *et al.*, 2009). Drilling near CG revealed *in situ* CO_2 concentrations close to saturation (> 900 mmol l^{-1}) near the base of the Navajo Sandstone Formation (~ 330 m) and decreasing but still quite high (~ 700 mmol l^{-1}) towards the top of the formation (~ 220 m), presumably because of mixing with meteoric water (Kampman *et al.*, 2013). For additional descriptions of CG, the reader is referred to Baer and Rigby (1978), Barnes (1996) and Waltham (2001); for a stratigraphic column of this area, see chart 67 in Hintze (1988), and for a geologic map, see Doelling (2002).

Here we use metagenomic techniques to investigate microbial community composition and metabolic potential in the reservoir that supplies CG. To complement previous work at pilot geologic carbon sequestration sites, which have measured more immediate microbial community responses to CO_2 injection, we present a metagenomic study of a system, in which bacterial and archaeal consortia have presumably had decades or longer (possibly hundreds of thousands of years or more; Burnside *et al.*, 2013) to

adapt to subsurface, high-CO_2 conditions. We hypothesized that CG would contain a metabolically diverse and potentially largely autotrophic community, and that metagenomic analyses would yield insight into how subsurface microbial communities impact biogeochemical cycling and potentially influence geologic carbon sequestration on long timescales.

Results and discussion

Using 1.3 Gbp of metagenomic data from CG2_0.2A (sample CG2, 0.2 µm filter, sequencing run A; Table 1), we reconstructed three near-complete (> 90% complete) and seven partial (>50% complete) genomes from CG bacteria and archaea, including predicted autotrophic iron- and sulfur-oxidizing *Proteobacteria* and bacteria and archaea from candidate phyla (CP). To place these genomes in the context of the larger CG microbial community and to enable the phylogenetic analysis of less abundant populations, additional sequencing via one Illumina HiSeq2000 run was conducted on the same library, which we call metagenome CG2_0.2B (sample CG2, 0.2 µm filter, sequencing run B), and via another HiSeq run on a library generated from the > 3.0 µm size fraction from the same water sample, which we call CG2_3.0 (sample CG2, 3.0 µm filter). In total, 91.3 Gbp of metagenomic sequence was generated (Table 1).

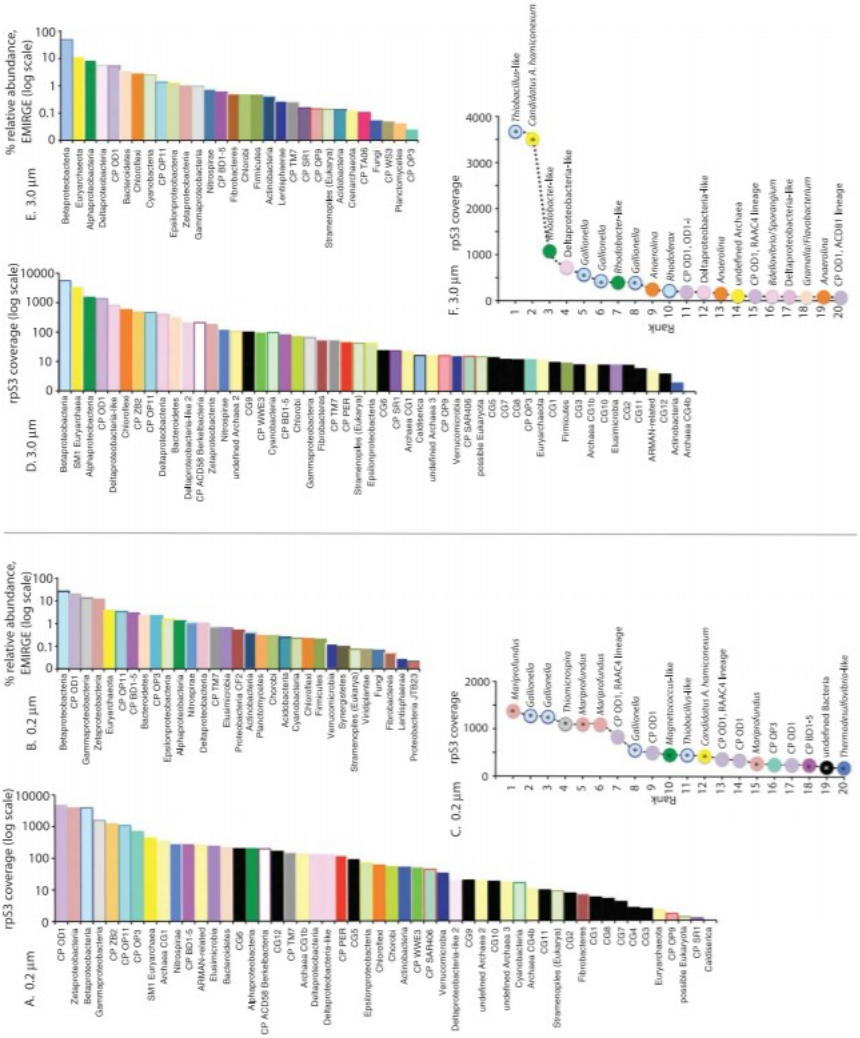

Figure 1. Rank-abundance curves from CG2_0.2B and CG2_3.0. The relative abundances of taxonomic groups in CG2_0.2B (left panel, A–C) and CG2_3.0 (right panel, D–F) are depicted. A, C, D and F are rank-abundance curves based on the average depth of coverage of the contig containing the ribosomal protein S3 (*rpS3*) gene sequence, calculated as follows: the number of reads that mapped to the contig × 100 bp (the read length)/the length of the contig in bp. Taxa are listed in vertical order of their ranked abundance. A and D are in log scale and summed at the phylum level, except for *Proteobacteria*, which are summed at the class level; C and F are the top 20 OTUs for each size fraction (C: 0.2 µm filter, F: 3 µm filter), with asterisks indicating OTUs with partially reconstructed genomes from CG2_0.2A metagenomic data. B and E are rank-abundance curves (log scale) summed at the phylum level based on EMIRGE calculations of 16S rRNA gene relative abundances (see Miller *et al*., 2011). Colouring is consistent across all six graphs. Taxa detected in this study that have no known affiliation at the phylum level are in black (bacteria) or light yellow (archaea). All archaea are shades of yellow. The data used to generate the *rpS3* graphs are in Table S5.

Overview of community structure

We profiled the community structure of the 0.2–3 µm and > 3 µm size fractions, using *de novo* assembled metagenomic data from CG2_0.2B and CG2_3.0, and we corroborated our results with Expectation Maximization Iterative Reconstruction of Genes from the Environment (EMIRGE) 16S rRNA

gene sequence analyses on the same metagenomes (EMIRGE results from the CG2_0.2A metagenome can be found in Table S1). In the *de novo* assemblies, bacteria and archaea were identified primarily using ribosomal protein S3 (*RpS3*) sequences (protein sequences predicted from gene sequences; Fig. 1 and Table S2), and, where possible, bacteria were also identified using DNA gyrase subunit A (GyrA) predicted protein sequences (Fig. 2). Taxonomic classifications were based primarily on a phylogenetic analysis that included 392 CG RpS3 sequences and 205 database sequences. We identified 311 unique operational taxonomic units (OTUs) from 47 phyla, including 12 different CP and 16 previously unknown lineages (i.e. RpS3 sequences from potentially phylum-level groups that are as-yet unidentified outside CG; Table S3). Rank-abundance curves were constructed for the CG2_0.2B and CG2_3.0 metagenomes, using these phylogenetic classifications and abundance information obtained from read coverage depth of the contig carrying the rpS3 gene (Fig. 1; Table S2). This read mapping-based approach allowed us to quantify the abundance of over 300 different taxa with a sensitivity as low as ~0.004% of the community, as long as the contig encoding the ribosomal protein block was reconstructed from at least one of the two filter size fractions. We also determined the overall phylum- (or class-, for *Proteobacteria*) level representation of the two size fractions, based on both RpS3 data and EMIRGE-calculated 16S rRNA gene relative abundance (Miller *et al*., 2011) (Fig. 1). The RpS3 and EMIRGE phylum-level analyses are largely in agreement, in terms of predicting the most abundant phyla in each size fraction. These analyses reveal that, although specific proteobacterial populations and a *Candidatus* Altiarchaeum hamiconexum SM1 euryarchaeal population dominate the system at the population level, there are a wide variety of CP bacteria that, when grouped at the phylum level, account for a significant portion of the community. For example, at moderate abundance are OD1, ZB2 and OP11 populations retained on the 0.2 and 3 µm filters, but, in aggregate, OD1 accounts for more cells than any other group on the 0.2 µm filter, as inferred from coverage of the contig containing the *rpS3* gene. EMIRGE relative abundance calculations suggest a similarly high abundance of OD1 on 0.2 µm filters, eclipsed only by *Betaproteobacteria*. ZB2, OP11 and OP3 are also relatively well-represented phyla, particularly in the 0.2–3.0 µm size fraction.

Table 1. Crystal Geyser sample metadata.

Sample	Date	Time	T (°C)	pH	TDS (p.p.m.)	Sample type(s)	CG2_0.2A[a] sequencing	CG2_0.2B[a] sequencing	CG2_3.0[a] sequencing
CG1	6 November 2009	7:32 p.m.	17	7.5	10880	Geochemical	N/A	N/A	N/A
CG2	8 November 2009	3:40 a.m.	17.5	7.6	11520	Metagenomic, Geochemical	1.3 Gbp	45 Gbp	45 Gbp

a. These metagenome names refer to the sample (CG2), filter size (0.2 or 3.0 µm) and, for the CG2 0.2 µm filter, the sequencing run (A: Illumina GAIIx, B: Illumina HiSeq2000).
Gbp, gigabase pairs; TDS, total dissolved solids.

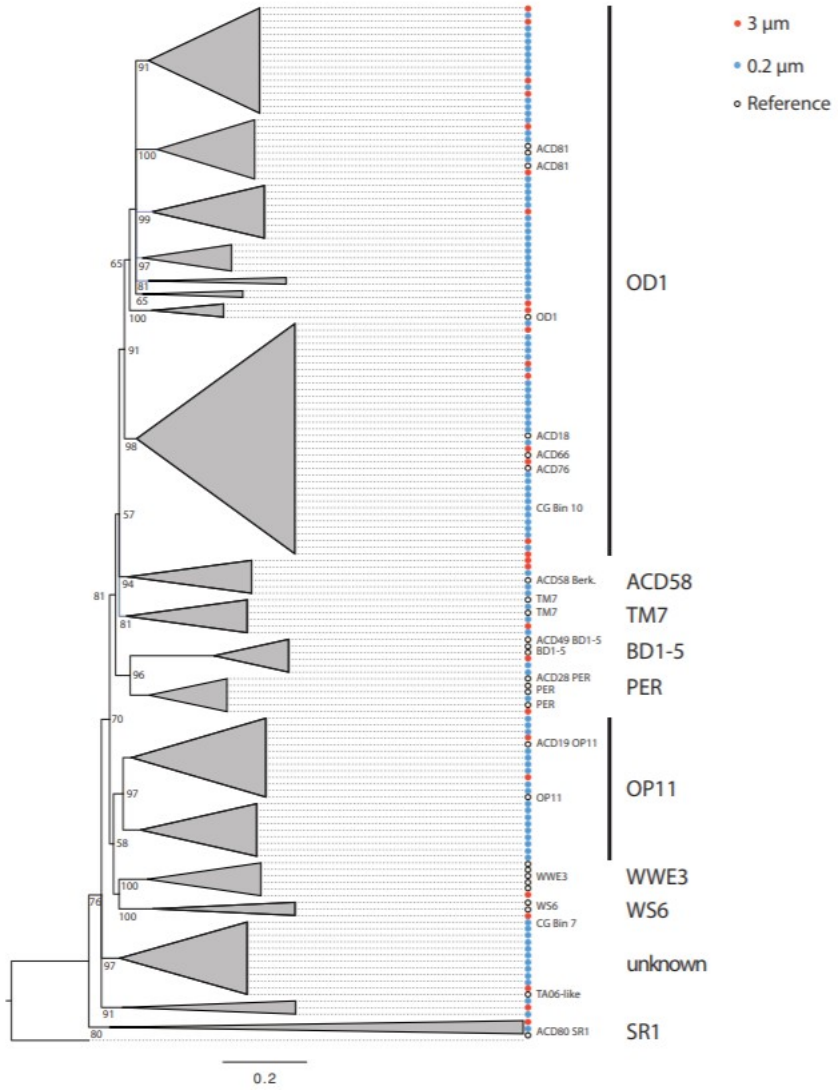

Figure 2. Candidate phyla DNA gyrase A protein sequence tree. DNA gyrase A protein sequences predicted to belong to CP were collected from CG2_0.2A, CG2_0.2B and CG2_3.0 assemblies, and a small number of reference sequences (including those labeled 'ACD' from Wrighton *et al*., 2012) were included for each phylum for improved alignment. The sequences labelled 'CG Bin' correspond to the numbered genome bins described in the text from CG_0.2A. Sequences were aligned using MUSCLE, and a maximum-likelihood RAxML tree was generated (RAxML-HPC-PTHREADS under the PROTGAMMAWAG model of evolution, with 100 bootstraps) and edited with FigTree and Adobe Illustrator. The size of collapsed nodes indicates the number of OTUs in each group, with each dotted line leading to a small, coloured circle indicating a single OTU. The colour of each small circle indicates the size fraction from which the sequence was assembled (0.2 µm = blue; 3 µm = red; reference sequences = white with black outline).

Interestingly, the community compositions of the 0.2–3 µm and > 3 µm size fractions are distinct. In particular, the abundances of CP (many of which are predicted to have small cell size, e.g. Kantor *et al*., 2013) and *Zetaproteobacteria* are higher in the smaller size fraction, whereas the abundances of SM1 Euryarchaea and *Alphaproteobacteria* are higher in the larger size fraction. Also, predicted neutrophilic, iron-oxidizing bacteria (FeOB), including *Gallionellales Betaproteobacteria* and *Mariprofundus*

Zetaproteobacteria, are the most abundant populations in the smaller size fraction, whereas *Hydrogenophilales* sulfur-oxidizing bacteria are most abundant in the larger size fraction (still, both predicted iron- and sulfur-oxidizing bacteria were found at high abundance in both size fractions). Although Archaea are relatively abundant in both size fractions, many archaeal groups exhibit different abundances between filters.

The abundance of putative neutrophilic, FeOB in CG is consistent with circumneutral pH (Table 1) and high iron concentrations (Table S4). At low oxygen concentrations (< 50 μM) in systems with abundant Fe(II), neutrophilic Fe(II) oxidation can provide energy for metabolism (at higher oxygen concentrations, abiotic iron oxidation outcompetes biological iron oxidation) (Singer and Stumm, 1970; Roden *et al*., 2004; Emerson *et al*., 2010). The diversity and physiology of neutrophilic, FeOB have been recently reviewed (Emerson *et al*., 2010), and this group includes *Gallionellales* and *Zetaproteobacteria*, as found in the CG system. Specifically, *Gallionella* spp. are *Betaproteobacteria* that have been shown to grow both chemoautotrophically and mixotrophically (i.e. with complex carbon as a carbon source) (Hallbeck and Pedersen, 1991), and both *Sideroxydans* spp. (also in the *Gallionellales* order and closely related to *Gallionella*) and *Mariprofundus ferrooxydans* PV-1, the genome-sequenced type strain of the *Zetaproteobacteria*, are reported to be obligate iron-oxidizing chemolithoautotrophs (Emerson and Moyer, 1997; Emerson *et al*., 2007; Singer *et al*., 2011). *Gallionella* and *Sideroxydans* spp. tend to be among the most abundant FeOB in freshwater systems, whereas the *Zetaproteobacteria* tend to be found at higher salinity, primarily in marine systems but also in brackish waters (e.g. Emerson *et al*., 2010; McBeth *et al*., 2013; Dang *et al*., 2011).
Though *Zetaproteobacteria* and *Gallionellales* have been shown to co-occur in at least one other system (McBeth *et al*., 2013), our results suggest that both ‘marine’ and ‘freshwater’ FeOB can be at high abundance in the same ecosystem (Fig. 1 and Table S2). However, it is possible that both are present in separate microniches in the geyser system and only become mixed as a result of mechanical processes.

RuBisCO genes

Ribulose-1,5 bisphosphate carboxylase-oxygenase (RuBisCO) is essential for carbon fixation via the Carbon-Benson-Bassham Cycle (Bassham *et al*., 1950), and RuBisCO type I functions best at medium-to-low CO_2 (0.1–1%) with oxygen present, whereas type II has a low discrimination against oxygen as an alternative substrate and is predicted to be most common at high CO_2 (>1.5%) and low oxygen concentrations (Badger and Bek, 2008; Singer *et al*., 2011). Many previously genomically sequenced neutrophilic FeOB, including *M. ferrooxydans* PV-1, have both type I and type II RuBisCO genes, suggested to be an adaptation to fluctuating CO_2 concentrations (Emerson *et al*., 2007; Singer *et al*., 2011). The detection of only type II but not type I RuBisCO in the reconstructed genomes from CG2_0.2A (described

below) may be indicative of adaptation to the high CO_2 concentrations in the geyser system.

Notably, several genome fragments from CG2_0.2B and CG2_3.0 encode type III RuBisCO protein complexes typically associated with archaea. These gene products appear to play a role in the AMP salvage pathway for CO_2 fixation (Sato *et al.*, 2007; Tabita *et al.*, 2007; 2008). Further, there is evidence for type I RuBisCO-encoding genes and for a newly reported set of hybrid type II/III RuBisCO-encoding genes that have also been inferred to confer CO_2 fixation capacity (Wrighton *et al.*, 2012). Overall, there is a strong genomic signal for the capacity for CO_2 fixation in the aquifer from which CG is sourced (Fig. 3).

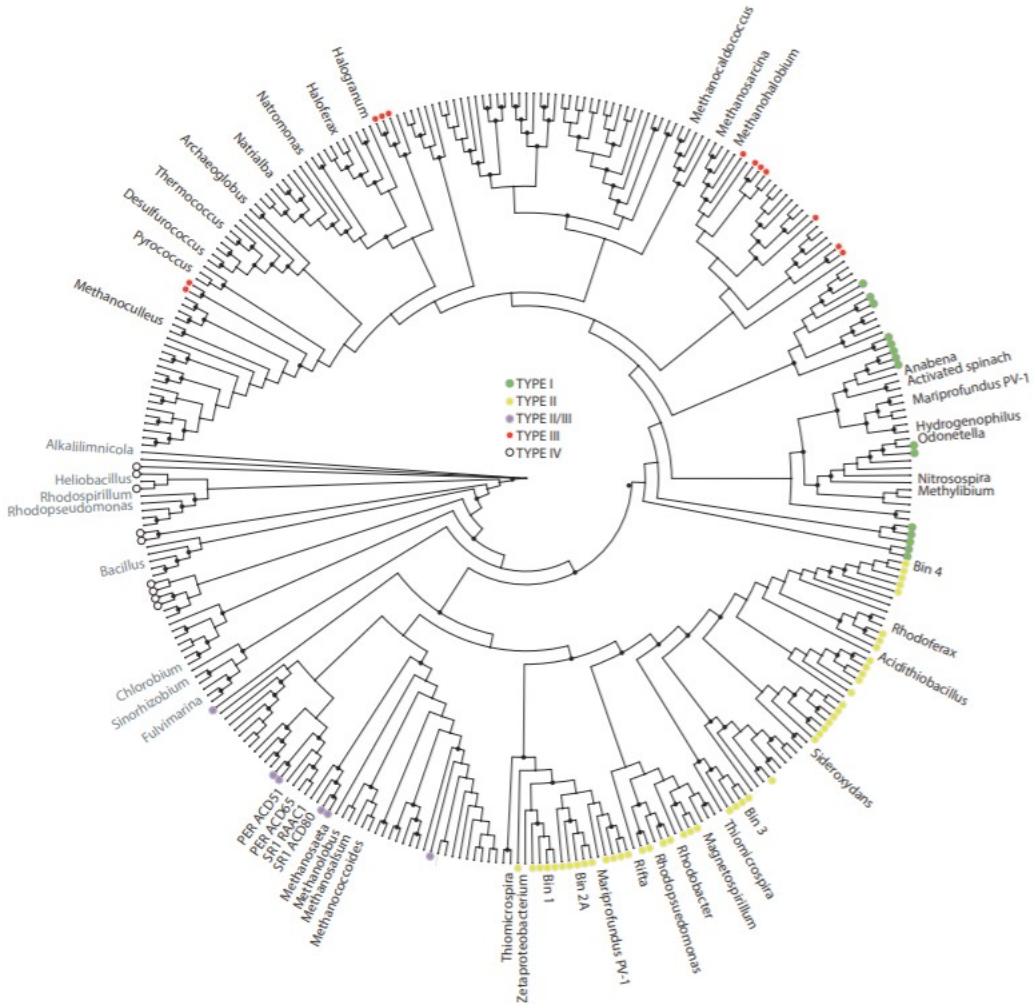

Figure 3. RuBisCO protein sequence tree. RuBisCO predicted protein sequences were collected from CG2_0.2B and CG2_3.0 assemblies, along with four RuBisCO protein sequences predicted from CG2_0.2A assemblies (labelled with bin numbers). Sequences > 196 amino acids in length were aligned with database sequences using MUSCLE, and the alignment was trimmed at both ends. Columns with > 97% gaps were removed and the resulting alignment used to construct a maximum-likelihood phylogenetic tree with RAxML (RAxML-HPC-PTHREADS under the PROTGAMMAWAG model of evolution, with 100 bootstraps) and edited with FigTree and Adobe Illustrator. Circles at branch tips (colour-coded by RuBisCO type, see legend) represent CG sequences, and tips without circles indicate reference sequences. Small black circles indicate bootstrap values > 80 (out of 100).

Genomic reconstructions

Genome reconstruction-based analyses focused on the CG2_0.2A dataset. We *de novo* assembled 9 million 150 bp paired-end Illumina sequencing reads (Table 1) and achieved genomic reconstruction for bacteria, archaea, viruses and plasmids. Of the reads that passed quality control filtering, 52% were assembled into contigs > 1 kb and 28% were assembled into contigs > 5 kb. The largest contig was 185 604 bp, and there were 20 contigs > 100 000 bp. We visualized tetranucleotide frequency patterns of all contigs > 5 kb on an emergent self-organizing map (ESOM; Fig. 4) and used this map in combination with phylogenetic information, measures of genome completeness and other information to bin genomes for genomic analyses (see *Experimental procedures* for details). Nearly all of the contigs > 5 kb were binned into genomes (97% of the reads from contigs > 5 kb were binned, or 27% of the total reads in the dataset). We reconstructed three near-complete (> 90% single-copy genes present) and seven partial (> 50% complete) genomes, and we binned contigs from ~ 10 additional genomes from bacteria and archaea, including previously genomically unsampled or little-sampled lineages. We also partially reconstructed five putative bacteriophage genomes and two putative plasmids (degree of completeness unknown).

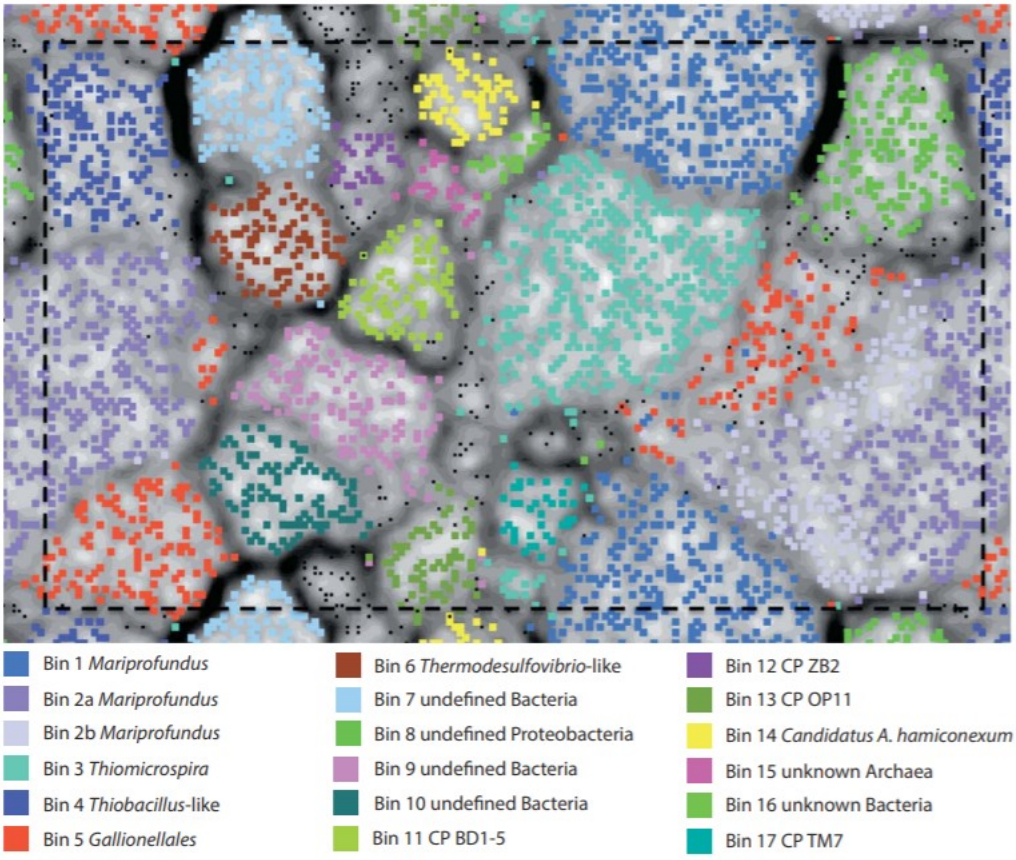

Figure 4. ESOM map of tetranucleotide frequency bins. CG2_0.2A genome bins were generated from tetranucleotide frequencies in this ESOM. Black-to-dark gray regions are topographically raised,

delineating differences in tetranucleotide frequency patterns between contigs in valleys (white-to-light gray). Darker delineations indicate more pronounced differences in tetranucleotide frequency patterns. Dots represent 5 kb contig regions, colour coded by bin number, as in the legend. Small black dots represent contigs that were either not binned, removed from a bin or that came from small bins not significantly described in the text, including putative plasmids and viruses. Bins with 'undefined' in the name could not be assigned to a taxonomic group for reasons outlined in the *Experimental procedures* section; bins with 'unknown' in the name were too incomplete for robust phylogenetic inference. The black dotted line outlines one instance of the map and highlights its periodicity.

In the sections that follow, we detail the phylogenetic placement and metabolic potential of bacteria and archaea with genomes > 50% complete recovered from CG2_0.2A. Information about binned genomes < 50% complete can be found in Appendix S1. Although we focus on the CG2_0.2A dataset for genomic analyses, additional information from CG2_0.2B and CG2_3.0 has been incorporated, as indicated.

CG *Zetaproteobacteria*

Zetaproteobacteria are predicted to be among the most abundant microorganisms in CG, particularly in the 0.2–3.0 μm filter size fraction, based on RpS3 and EMIRGE data from the CG2_0.2B and CG2_3.0 metagenomes. A total of seven zetaproteobacterial OTUs were identified in those metagenomes (Table S2), and three zetaproteobacterial genomes were partially genomically reconstructed from CG_0.2A metagenomic data (*Mariprofundus* CG Bins 1, 2a and 2b). Our metabolic analyses focus on two of these genomes, which we were able to significantly recover (Bins 1 and 2a, close to 100% and 80% complete, respectively, based on single-copy gene inventory; Table 2; Table S5). These *Zetaproteobacteria* are expected to show some functional similarity to the type strain, *M. ferrooxydans* PV-1 (Singer *et al*., 2011), based on similar protein complements and phylogenetic analysis of EMIRGE-reconstructed 16S rRNA genes (see below). The majority of the proteins predicted from genes in the PV-1 genome are also predicted from the CG *Mariprofundus* Bin1 and Bin 2a genomic data. Similarities to the sequenced PV-1 strain include the presence of genes encoding proteins involved in: motility (flagella and chemotaxis), glycolysis and the tricarboxylic acid (TCA) cycle, along with genes encoding numerous transporters, particularly metal transporters. Based on phylogeny (according to 16S rRNA gene, *rpS3* gene and *gyrA* gene trees) and gene inventory, the CG populations have a predicted iron-oxidizing metabolism with oxygen as the terminal electron acceptor.

Table 2. CG2_0.2A genome bin summary statistics.

Final bin	Original ESOM bin[a]	Phylogenetic affiliation	Original single-copy genes (out of 30)	Final single-copy genes (out of 30)	Split or combined bin?	Number of contigs (final bin)	Genome size (total bp in final bin)	Average % GC	Average coverage CG_0.2A	Per cent relative abundance (gyrA)[c]
1	1	*Mariprofundus*	30	30	No	35	2313086	46.4	37	10
2a	2	*Mariprofundus*	31	25	Split	61	2344113	58.7	22	5.9
2b	2	*Mariprofundus*	N/A	6	Split	123	1010809	55.3	7	2
3	3	*Thiomicrospira crunogena*	30	30	No	131	2257568	43.5	31	7.1
4	4	*Thiobacillus*-like *Hydrogenophilales*	17	17	No	124	1015462	61	9	2.9
5	5	*Gallionellales*[b]	40	40	No	201	1493567	54.4	27	1.8
6	6	*Thermodesulfovibrio*-like	12	12	No	83	614406	44.2	1	No gyrA
7	7	Unknown bacteria	8	8	No	93	913688	42.8	8	2.1
8	8	Unknown *Proteobacteria*	6	6	No	120	1033127	61.9	10	3
9	9	Unknown bacteria	17	17	No	59	753732	46.6	23	6
10	10	Unknown bacteria	18	18	No	67	687190	48.6	8	1.7
11	11	Candidate phylum BD1-5	21	21	No	67	513393	38.8	7	No gyrA
12[d]	12	Candidate phylum ZB2	0	15	Combined	81	508790	41.9	7	1.7
12	25	Candidate phylum ZB2	5	N/A	Combined	N/A	N/A	N/A	N/A	N/A
12	26	Candidate phylum ZB2	2	N/A	Combined	N/A	N/A	N/A	N/A	N/A
12	27	Candidate Phylum ZB2	8	N/A	Combined	N/A	N/A	N/A	N/A	N/A
13	13	Candidate phylum OP11	8	8	No	53	391996	36.1	4	1.5
14	14	*Candidatus* Altiarchaeum hamiconexum	14	14	No	63	462410	32.7	13	No gyrA
15	15	Archaea	6	6	No	29	181298	34.5	4	1
16	16	Bacteria	1	1	No	33	213208	37.4	5	No gyrA
17	17	Candidate phylum TM7	9	9	No	32	237016	40.6	5	No gyrA

a. Assembled contigs > 5 kb were used for the ESOM genome binning analysis.
b. Bin 5 contains multiple, unresolved *Gallionella*-like and *Sideroxydans*-like genomes.
c. Per cent relative abundance was calculated based on the average coverage of the contig containing the DNA gyrA gene (if present), and this coverage depth was extrapolated across the full length of the genome (we assume a genome size of 2 Mbp if the size is not otherwise known) and divided by the total number of reads in CG2_0.2A, times 100.
d. Bin 12 is a combination of four neighbouring ESOM bins on the original ESOM map (not shown), and this row shows the genome summary results for the final, combined Bin 12.
Binning information and genome summary statistics for each genome bin in the CG2_0.2A metagenome. The 'final bin' column indicates the bin ID, as described in the text and colour-coded in Fig. 4. The 'original ESOM bin' column indicates the original ESOM bin (map not shown, but similar to Fig. 4) considered for combining with or splitting from other bins to generate single genomes, based on other data presented in this table (single-copy genes, %GC, and coverage).

Notable differences between the CG and PV-1 *Mariprofundus* genomes include the presence of only type II RuBisCO genes in the genomes of the geyser *Zetaproteobacteria* (Fig. 3), consistent with zetaproteobacterial genomic data from single-cell sequencing (Field *et al*., 2014). In addition, a complete nitrogen fixation pathway was recovered from the *Mariprofundus* CG Bin 2a genome. The CG genomes also contain catalase-peroxidase and glutathione peroxidase genes, which function in the removal of toxic oxidants and are found in nearly all aerobic bacteria, although they were notably absent from the PV-1 genome (Singer *et al*., 2011). The detection of only type II RuBisCO genes in the CG *Mariprofundus* genomes is consistent with CO_2 saturation in the geyser system.

Based on the PV-1 genome (Singer *et al*., 2011), it has been suggested that *Mariprofundus* species may be able to grow mixotrophically (i.e. they may be able to assimilate complex carbon as a carbon but not an energy source). Genes encoding inventory for sugar uptake and metabolism would support the possibility of mixotrophy in the CG genomes, although these genes could also indicate the potential for organotrophy. Consistent with the sequenced isolate, the capacity for energy storage and for the breakdown of both polyphosphate and glycogen are indicated in the CG genomes.

To characterize the phylogenetic relationships of the reconstructed *Mariprofundus* genomes to *Zetaproteobacteria* from environmental 16S rRNA gene surveys, we identified two *Mariprofundus* 16S rRNA gene sequences via EMIRGE (Miller *et al*., 2011) in the CG2_0.2A dataset, each with 96% identity to the *M. ferrooxydans* PV-1 reference sequence and 95% identity to each other at the nucleic acid level (Table S1). In a phylogenetic tree of known *Zetaproteobacteria*, modified from McAllister and colleagues (2011) and including sequences from Dang and colleagues (2011), one of the CG 16S rRNA gene sequences clusters with ‘OTU 9’, which is populated by marine subsurface clones, and the other clusters with sequences from several OTU groups from diverse, iron-rich marine environments (Fig. S1). Phylogenetic placement of the CG *Mariprofundus* sp. within the *Zetaproteobacteria* class but separate from the previously sequenced *M. ferrooxydans* PV-1 isolate indicates that we have generated information on the metabolic potential of *Zetaproteobacteria* distinct from the type strain, and the abundance of *Zetaproteobacteria* in the geyser system extends the environmental distribution of this group to the terrestrial subsurface.

Thiomicrospira crunogena CG Bin 3

Relative abundance estimates from the CG2_0.2B and CG2_3.0 datasets suggest that *Thiomicrospira crunogena* is among the more abundant bacteria in CG, particularly in the 0.2–3 μm size fraction, in which it is predicted to be the fourth most abundant population. The *T. crunogena* CG Bin 3 genome is estimated to be nearly 100% complete (Table 2). In general,

the *T. crunogena* CG Bin 3 genome is highly similar to the previously sequenced strain, *T. crunogena* XCL-2 (Scott *et al*., 2006), isolated from a deep-sea hydrothermal vent, and is inferred to be functionally similar. This is consistent with 99% nucleic acid sequence identity to the isolate in the 16S rRNA gene sequence from the CG2_0.2A EMIRGE analysis (Table S1). Similar to the sequenced isolate, the genome is predicted to encode a full suite of Sox proteins for sulfur oxidation, a complete glycolytic pathway, and is motile, with flagella, pili and chemotaxis genes. As expected, the genome has evidence for hydrogen metabolism in the form of a number of hydrogenase genes, including a Ni,Fe-hydrogenase operon, and it has genomic inventory encoding a cbb3-type cytochrome C oxidase, indicative of microaerobic respiration (Kulajta *et al*., 2006) and NO defence (Stein *et al*., 2007). One notable difference to XCL-2 and similarity to other CG predicted chemoautotrophs is the presence of genes encoding only type II RuBisCO proteins in the genome of the geyser *Thiomicrospira* organisms, indicating adaptation to the CO_2-saturated geyser system (Fig. 3).

Thiobacillus-like CG Bin 4

Notably, a predicted RpS3 sequence that groups with the *Hydrogenophilales* was the most abundant in the CG2_3.0 metagenome, although this sequence could not be definitively linked to CG Bin 4, as no *rpS3* gene sequence was identified in the Bin 4 draft genome (Table S2). In general, relative abundance information from CG2_0.2B and CG2_3.0 place *Thiobacillus*-like bacteria among the most abundant populations in both size fractions (most abundant in the > 3 µm size fraction and 11th most abundant in the 0.2–3 µm size fraction, out of 311 OTUs total). The majority of best BLAST hits for proteins encoded by the *Thiobacillus*-like CG Bin 4 genome are to *Betaproteobacteria*, particularly *Thiobacillus denitrificans*, a member of the *Hydrogenophilales* order.

Consistent with the sequenced *T. denitrificans* genome (Beller *et al*., 2006b), CG Bin 4 contains genes predicted to code for proteins involved in sulfur oxidation, including a number of *dsr* genes in a large, sulfur-associated operon. Although *dsr* genes could potentially indicate dissimilatory sulfate reduction, many of the predicted protein sequences from this genome have best BLAST hits to *T. denitrificans*, which has been shown to use these genes for thiosulfate oxidation (Beller *et al*., 2006a). The CG genome also contains other genes predicted to encode proteins involved in sulfur oxidation (e.g. flavocytochrome C). The presence of cytochrome C oxidase and a periplasmic nitrate reductase complex suggests a facultatively aerobic metabolism with the ability to grow anaerobically on nitrate, similar to the sequenced *T. denitrificans* genome (Beller *et al*., 2006b). Also similar to the sequenced genome, there are a number of predicted hydrogenase and hydrogenase accessory protein genes, particularly genes encoding Ni-dependent hydrogenases. Interestingly, the CG Bin 4 genome contains genes for both arsenate oxidase and reductase, and has a number of arsenate and arsenite accessory genes. The sequenced isolate was shown to have high

metal resistance and is the only organism known to oxidize uranium anaerobically, so arsenate metabolism is generally consistent with metal resistance and redox chemistry in this group of bacteria. As with the previously described CG chemoautotrophs, *Thiobacillus*-like CG Bin 4 encodes only type II RuBisCO and accessory proteins, indicating an ability to fix CO_2 chemoautotrophically under high CO_2 conditions (Fig. 3).

Gallionellales CG Bin 5 genomes

Gallionellales bacteria are predicted to be among the top five most abundant populations in both the 0.2–3 µm and > 3 µm size fractions, based on data from CG2_0.2B and CG2_3.0 (Fig. 1). Based on contig coverage depth (range 3–42×), GC content (range 46.7–64.7%), single-copy gene inventory (Table S5) and best BLAST hits for each contig, we determined that CG Bin 5 contains multiple genomes, predominantly from bacteria belonging to the *Gallionellales* order of *Betaproteobacteria*. As such, all interpretations from this section should be understood to come from multiple, unresolved genomes. The presence of multiple *Gallionellales* genomes is consistent with the EMIRGE-generated 16S rRNA gene sequences from all three metagenomes and with the RpS3 abundance data from CG2_0.2B and CG2_3.0, all of which indicate multiple, abundant populations of *Gallionella*-like and *Sideroxydans*-like bacteria (Fig. 1; Tables S1 and S2). A large number of bacteriophage, clustered regularly interspaced short palindromic repeats regions (Jansen *et al*., 2002; Barrangou *et al*., 2007), transposases and recombinases were binned with the CG Bin 5 genomes, suggesting a high degree of phage pressure and/or extensive horizontal gene transfer in this group.

A number of Bin 5 contigs with high coverage had best BLAST hits to bacterial genomes from multiple phyla, possibly representing new genomic regions within *Gallionellales* genomes (i.e. regions not affiliated with previously identified *Gallionellales* and perhaps obtained through horizontal gene transfer) or representing an unidentified population that co-binned with this group. At high coverage in CG *Gallionellales* Bin 5 were a number of cytochrome genes, consistent with respiration, although a cytochrome-based terminal oxidase (e.*g*, cbb3) was not identified. At lower coverage, Bin 5 contains a large operon (encoding the NAP complex) predicted to code for a number of hydrogenases, including Ni,Fe-hydrogenase subunits and accessory proteins, indicative of hydrogen metabolism and possibly the ability to use molecular hydrogen as an energy source.

One contig in CG Bin 5 contains a respiratory nitrate reductase operon encoding NarGH, and on the same small contig is a cytochrome C gene with a best BLAST hit (from its predicted protein sequence) to *Sideroxydans lithotrophicus* ES-1 (Emerson and Moyer, 1997), perhaps indicative of a *Sideroxydans*-like population capable of anaerobic respiration. Similarly, a small contig with a Sox operon and two best BLAST hits to *S. lithotrophicus* ES-1 may suggest that a *Sideroxydans*-like population can

acquire energy through sulfur oxidation in the geyser system. Although we cannot say with certainty that the identified nitrate reduction and sulfur oxidation operons are accurately placed within the assembled *Gallionellales* genomes, our data suggest that these metabolic capacities may be present in this clade. While there are no known *Gallionellales* isolates that can reduce nitrate, predicted nitrate reduction in the CG *Gallionellales* is consistent with an enrichment culture-based study of coupled Fe(II) oxidation and nitrate reduction, in which *Sideroxydans*-like *Betaproteobacteria* were found to be the most abundant organisms (Blöthe and Roden, 2009). In contrast, there is presently no information in the literature that suggests the potential for sulfur oxidation in the *Gallionellales*. The CG *Gallionellales* may also fix CO_2, given the presence of *Gallionellales* type II RuBisCO genes in the CG2_0.2B and CG2_3.0 metagenomes (Fig. 3).

Undefined bacteria CG Bins 9 and 10

CG Bins 9 and 10 are characterized as undefined bacteria because of best BLAST hits to many different phyla throughout the annotation (as opposed to best BLAST hits predominantly to a single phylum, characteristic of most other CG bacterial genomes) and the lack of definitive information from protein trees constructed from alignments of single-copy genes (i.e. affiliation with different phyla in different protein trees). However, it is possible that these bacteria are members of known CP that have minimal or no prior genomic sampling. For example, the protein predicted from the DNA *gyrA* gene for CG Bin 10 affiliates with CP OD1 (Fig. 2), but proteins predicted from other single-copy genes from Bin 10 affiliate with a variety of different phyla.

A wealth of genes for sugar uptake and metabolism in both genomes (CG Bins 9 and 10), e.g. citrate transporters, sugar ABC transporters, and genes for glycolysis and pyruvate metabolism, is consistent with fermentation, which is the metabolism predicted for a number of recently genomically sampled CP (Wrighton *et al.*, 2012). The CG Bin 9 genome also has two genes, malate dehydrogenase and succinate dehydrogenase, involved in central metabolism. Both genomes contain a number of nucleic acid and protein metabolic genes, and the CG Bin 10 genome indicates the capacity for carbon storage as glycogen. CG Bins 9 and 10 have predicted twitching motility, and CG Bin 10 has a nickel-dependent hydrogenase gene, consistent with hydrogen metabolism reported for other genomically sampled CP.

CP BD1-5 and ZB2 CG Bins 11 and 12

Based on consistent best BLAST hits for single-copy genes (and their respective predicted proteins) to sequences from specific CP in ggkBase (Wrighton *et al.*, 2012), along with phylogenetic affiliation with specific CP from a single-genome sequencing study (Rinke *et al.*, 2013), we predict that CG genomes from Bins 11 and 12 belong to bacterial CP BD1-5 and ZB2

respectively. While the majority of genes from these genomes are of unknown function or affiliation, and neither of the genomes is complete (the BD1-5 and ZB2 genomes are approximately 70% and 50% complete, respectively, based on single-copy gene inventory, Tables 2 and S5), the predicted functions are consistent with a fermentative metabolism, as predicted for other CP (Wrighton *et al*., 2012). For instance, terminal oxidases were not identified (although it is possible that they are present in genomic regions that were not binned), and a large number of genes from each genome are involved in sugar metabolism, including a variety of glycosyltransferase genes. Both genomes have genes encoding inventory required for pilus synthesis and twitching motility. Consistent with previously genomically sampled CP, CG Bin 12 (ZB2) has evidence for a Gram-positive cell wall, including a gene predicted to code for a Gram-positive anchor protein.

Consistent with previous metagenomic reconstruction of BD1-5 genomes, the CG Bin 11 BD1-5 is predicted to use an alternative genetic code (Wrighton *et al*., 2012), in which the UGA stop codon is translated as glycine, as predicted for related CP SR1 bacteria (Campbell *et al*., 2013; Kantor *et al*., 2013). This coding was recently confirmed by proteomic analysis (Hanke *et al*., 2014). Notably, the incidence of the UGA codon in the CG BD1-5 genotype is low (0.74%), relative to the incidence in other similarly alternatively coded BD1-5 and SR1 genomes (0.81–4.72%) (Wrighton *et al*., 2012; Kantor *et al*., 2013) (Table 3). Still, the genomes with this coding use the UGA codon far more frequently than it is employed in some other genomes that use the standard bacterial code (0.07–0.09%).

Table 3. 'Stop' codon frequency in BD1-5 and other genomes.

Phylum	Genome	TAA	TAG	TAA/TAG	TGA
BD1-5	ACD_3	0.25	0.05	4.65	4.72
BD1-5	ACD_2	0.25	0.06	4.13	4.56
BD1-5	ACD_4	0.27	0.05	5.41	4.50
BD1-5	ACD_Cluster49	0.26	0.05	4.97	3.88
SR1	RAAC1	0.24	0.05	4.45	2.09
BD1-5	ACD_78	0.25	0.08	3.18	0.81
BD1-5	CG_Bin11	0.24	0.06	4.37	0.74
Proteobacteria	ACD_46	0.19	0.05	3.61	0.09
PER	ACD_28	0.16	0.08	2.00	0.09
PER	ACD_51	0.18	0.08	2.20	0.06
Proteobacteria	ACD_69	0.16	0.07	2.14	0.07

The frequency of codons that normally code for stopping translation is presented for BD1-5 and SR1 bacteria, which are predicted to use UGA to code for glycine. For reference, we also show these codon frequencies in four other bacteria that are predicted to use UGA as a stop codon. All genomes were assembled from metagenomes, either from this study (CG_Bin11) or from a shallow aquifer in Rifle, CO (all other genomes; Wrighton *et al*., 2012; Kantor *et al*., 2013). Codons are shown in DNA sequence form (e.g. the UGA stop codon is shown as TGA).

Candidatus Altiarchaeum hamiconexum (SM1 Euryarchaeota) CG Bin 14

The CG Bin 14 draft genome has the majority of its best BLAST hits to methanogens, but the primary means of energy generation for this SM1 Euryarchaeal population is not obvious from the genome at its current level of completion (~ 50%; Table 2). With a genome that is ~ 50% complete, we cannot make definitive metabolic predictions, but some inferences can be made from the genes that are present. The CG Bin 14 *Candidatus* Altiarchaeum hamiconexum draft genome contains a V-type ATPase gene, which is found predominantly in vacuoles and other eukaryotic structures but also in methanogenic and other archaea (e.g. Thermoplasmatales) and some bacteria (Yokoyama *et al*., 2000). The CG Bin 14 genome has indications of anaerobic glucose metabolism, including a pyruvate-formate lyase-activating enzyme gene, along with a methionine adenosyltransferase gene to make its required cofactor, S-adenosylmethionine. The genome also contains a glycogen synthase gene, indicating carbon storage as glycogen, and it has an alpha-amylase gene, coding for the breakdown of complex carbon and perhaps suggesting the potential for heterotrophy or chemoorganotrophy. It contains the gene for a ferripyochelin-binding protein, which binds siderophores and iron, suggesting the ability to uptake iron from the environment. Further information about the biology and metabolic potential of *Candidatus* Altiarchaeum hamiconexum, including genomic information from populations from CG and from a cold, sulfidic spring near Regensburg, Germany, can be found in (Probst *et al*., 2014).

Community ecological predictions

We now use the genomic data from the CG2_0.2A assemblies (described above), in combination with our geochemical measurements and previously published geological data, to make predictions about microbial community ecology in the subsurface aquifer from which CG is sourced. This synthesis is based largely on genomes recovered only from the 0.2–3 µm size fraction from a single water sample, including partial genomes of 13 of the 20 microorganisms predicted to be the most abundant in this size fraction and five (including the top two) of the 20 most abundant in the > 3 µm size fraction (Fig. 1). The two genomically reconstructed *Mariprofundus*-like *Zetaproteobacteria* (CG Bins 1 and 2a), *T. crunogena* (CG Bin 3), and *Thiobacillus*-like *Hydrogenophilales* (CG Bin 4) appear capable of carbon fixation, and as such, they are likely to be primary producers in the geyser system. *Thiomicrospira crunogena* (CG Bin 3), *Thiobacillus*-like *Hydrogenophilales* (CG Bin 4) and an unidentified member of the *Gallionellales* in CG Bin 5 show genomic evidence for sulfur oxidation. Genomic evidence for dissimilatory sulfate reduction appears in the *Thermodesulfovibrio*-like CG Bin 6 and *Proteobacteria* CG Bin 8 genomes (see Appendix S1). *Mariprofundus* (CG Bin 2a) appears to represent one of the primary sources of nitrogenase-dependent nitrogen fixation, and the CG Bin 8 *Proteobacteria* can likely also fix nitrogen. *Thiobacillus*-

like *Hydrogenophilales* (CG Bin 4), *Mariprofundus* (CG Bin 2b) and *Gallionella* (CG Bin 5) all appear capable of nitrite and/or nitrate reduction. All of the *Mariprofundus Zetaproteobacteria* are predicted to oxidize iron, as are most of the *Gallionellales* in CG Bin 5. As a consequence of this metabolism, many previously characterized neutrophilic FeOB have been shown to produce extracellular iron oxide stalks or sheaths, or amorphous iron oxyhydroxides (Emerson and Ghiorse, 1993; Emerson and Moyer, 1997; Chan *et al*., 2011; Comolli *et al*., 2011; Krepski *et al*., 2012). As suggested previously (Wrighton *et al*., 2012), members of bacterial CP and unidentified bacterial lineages in CG may be mostly fermentative, feeding on the metabolic products of other microorganisms in the system.

Overall, the community structure and genetic potential encountered in CG are consistent with the geochemistry (Table S4) and local geology. Similarities to hydrothermal vent communities and genomes are consistent with gradients of CO_2 and O_2 in the geyser system and the absence of light for primary production. Similarities to neutrophilic FeOB are consistent with circumneutral pH and high iron content (Tables 1). Differences in community composition and genomic content, relative to other systems, can be attributed to high CO_2 (e.g. the CG RuBisCO complement) and intermediate salinity (allowing for the coexistence of ‘marine’ and ‘freshwater’ neutrophilic FeOB) in the geyser system.

We are not aware of a study that has identified the geological source for the iron and sulfur that are presumably being metabolically oxidized by the microbial community within the CG water reservoir, but both are available in formations below the geyser (Hintze, 1988, chart 67). For example, the Jurassic and Triassic sandstones of this part of the Colorado Plateau show the presence of abundant ferric iron-rich minerals, which could potentially dissolve under anoxic conditions to provide the ferrous iron required for microbial Fe(II) oxidation. Sulfur could come from the dissolution of gypsum in the Jurassic Carmel Formation and/or the Pennsylvanian Paradox Formation.

Conclusions

We demonstrate that a subsurface, CO_2-saturated aquifer can support a phylogenetically diverse microbial community with the metabolic potential to harvest energy through the oxidation of inorganic compounds and to acquire carbon from the CO_2-saturated solution. The populations in the geyser system that are presumed to be the most abundant, based on metagenomic assembly data, appear capable of (and perhaps restricted to) chemolithoautotrophy, indicating that carbon assimilation solely from CO_2 is likely to be a major component of the carbon cycle in this system. This implies that abundant complex carbon sources are not likely to be a requirement for subsurface ecosystems associated with geologic carbon sequestration and that microbial communities in the deep subsurface may respond to elevated CO_2 from the external environment (i.e. anthropogenic

CO_2 additions as a part of geologic carbon sequestration efforts) with increased growth. Whether this actually occurs, and if so, on what timescale, is unknown, but interestingly, communities similarly populated by putative iron- and sulfur-oxidizing bacteria have been recovered from the CarbFix subsurface geologic carbon sequestration pilot site in Iceland (B. Menez, pers. comm.), highlighting the relevance of this work to geologic carbon sequestration. Overall, the diversity of Bacteria and Archaea affiliated with CP and previously unidentified lineages in the CG system should serve as impetus for future studies at CG and other subsurface sites with a range of geochemical conditions to increase insight into the subsurface biosphere, which may be responsible for a significant amount of the carbon cycling on Earth.

Experimental procedures

Field site and sample collection

CG is located at 38° 56.3′ N, 110° 8.1′ W, on the east bank of the Green River, 6 km south of the town of Green River, Utah, USA. Using a large (~ 2 ft diameter), sterile, plastic funnel placed in a sterile container, we collected 65 l of geyser water as it erupted in November 2009. We collected two samples (Table 1), one for solution chemistry from an eruption on 6 November 2009 (sample CG1), and a second for solution chemistry and metagenomics on 8 November 2009 (sample CG2). Water samples for solution chemistry were passed through 0.2 µm filters and placed in 30 ml acid-washed Nalgene bottles (three bottles were collected for each water sample: one untreated, one acidified with HNO_3 and one acidified with HCl; concentrations for each cation in each sample were calculated from all three preparations, Table S4). Bottles were stored at 4°C until processing via inductively coupled plasma mass spectrometry and ion chromatography (Table S4). Water samples for metagenomics (65 l) were filtered sequentially through 3.0 and 0.2 µm polyethersulfone filters (Pall Corporation, NY, USA) via a peristaltic pump, and filters were immediately frozen on dry ice in the field and then stored at −80°C in the laboratory until processing.

It should be noted that, at the aquifer depth from which the geyser is sourced (~ 200–500 m; Wilkinson *et al.*, 2009), the CO_2 is not predicted to be supercritical, as would be expected at a geologic CO_2 sequestration site. In addition, we acknowledge the potential for contamination from near-surface water and/or surfaces along the geyser conduit walls. However, given the narrow conduit diameter (a drill hole) and the duration and extent of eruptions, this is likely to be a minimal contribution to the overall DNA in our samples.

DNA extraction and sequencing

Metagenomic DNA was extracted from the sample CG2 0.2 µm filter for metagenomes CG2_0.2A and CG2_0.2B and from the sample CG2 3.0 µm filter for CG2_3.0. A sterile razor blade was used to cut each filter into small

pieces, which were added to the PowerMax Soil kit (MoBio, Carlsbad, CA, USA) and processed according to the manufacturer's instructions, followed by ethanol DNA precipitation. DNA from both CG2 filters was sent to Eureka Genomics (Hercules, CA, USA) for the construction of two TruSeq Illumina paired-end libraries (one per filter), each with 300 bp inserts. The first round of sequencing (CG2_0.2A) was 150 bp paired-end sequencing from the CG2 0.2 µm filter library on Eureka's Illumina GAIIx machine (30% of one lane). The same (already prepared) 0.2 and 3.0 µm filter libraries were then sent to the University of California Berkeley DNA Sequencing Facility, which generated 100 bp paired-end sequencing reads on the Illumina HiSeq platform (one HiSeq2000 flow cell per library, CG2_0.2B and CG2_3.0; CG2_0.2B is a second round of sequencing on the same library as CG2_0.2A). Reads from CG2_0.2A, CG2_0.2B and CG2_3.0 have been deposited to GenBank (BioProject ID PRJNA229517; accession numbers: SRP045164, SRS672458, SRX667892, SRR1534387, SRP045164, SRS672319, SRX667739, SRR1534154), and assembled contigs from all three metagenomes are available via ggkBase (Wrighton *et al*., 2012), including annotation, GC% and coverage data for each contig, along with bin information for CG2_0.2A. URLs for ggKbase data are as follows: http://ggkbase.berkeley.edu/CG1_02A/organisms (CG2_0.2A); http://ggkbase.berkeley.edu/CG2_30_FULL/organisms (CG2_3.0); http://ggkbase.berkeley.edu/CG1_02_FULL/organisms (CG2_0.2B).

Metagenomic assembly, annotation and binning

CG2_0.2A, CG2_0.2B and CG2_3.0 reads were trimmed for quality, resulting in the removal of ends of reads that had an Illumina quality indicator of ‘B’, defined in the Illumina manual as strings of Phred quality scores below 15 at either end of a sequence. The IDBA_UD assembly algorithm, version 1.0 (Peng *et al*., 2012) (default parameters) was applied to each metagenome, using all read pairs in which both reads were > 60 bp. Genes were predicted with PRODIGAL (Hyatt *et al*., 2010) on all contigs > 1 kb, using the meta option. Proteins were predicted from these gene sequences, and protein sequences derived from contigs > 5 kb were phylogenetically and functionally annotated, based on best BLAST hits (bit score ≥ 60) to the UniRef database, UniProt Release 2012_02 (Suzek *et al*., 2007). The tetranucleotide signatures of all CG2_0.2A contigs > 5 kb were submitted to an ESOM analysis (Dick *et al*., 2009). Tetranucleotide frequencies were calculated on all 5 kb contig segments, using standard approaches described in Dick and colleagues (2009), and an ESOM topographic map was generated using the Databionics application, ESOM-1.1 software, with default parameters and the ‘batch’ algorithm (Ultsch and Moerchen, 2005) to cluster and visualize shared tetranucleotide frequency patterns (Fig. 4). For further binning (confirmation of the existing ESOM bins as representative of single genomes, or the separation/combination of ESOM bins to generate single genomes where possible), the first step for each bin was to assess the distribution of each of the 30 single-copy genes (see details below and Table S5). If more than one

copy of any of the 30 single-copy genes was detected in a given bin, that bin became a candidate for separation into multiple bins. If co-located bins (bins in neighbouring locations on the ESOM map) had complementary distributions of single-copy genes (i.e. genes not detected in one bin were present in the other and vice versa), those bins became candidates for combination. At that point, all bins (those that were predicted to represent single genomes, those that were candidates for separation into multiple genomes and those that were candidates to be combined) were assessed for average coverage depth (from IDBA_UD output), taxonomy of best BLAST hits (the taxonomy of the top BLAST hit for the predicted protein for each gene) and average GC content on each contig. Consistency across these parameters, particularly with regard to coverage depth, was used to confirm or combine bins, and differences, particularly in coverage depth and GC content, were used to separate genomes appearing in the same ESOM bin. In general, this binning approach follows that of several prior studies (Dick *et al*., 2009; Wrighton *et al*., 2012; Castelle *et al*., 2013; Kantor *et al*., 2013). Bacteria and archaea identified through this binning procedure in CG2_0.2A are named according to their numbered genome bin (e.g. CG Bin 1, CG Bin 2a, CG Bin 2b, etc.) and should not be confused with names of unknown Bacteria and Archaea from CG2_0.2B and CG2_3.0 (e.g. CG1, CG2, etc. in Fig. 1).

Abundance information for CG2_0.2B and CG2_3.0 metagenomes was calculated by a mapping approach, using BOWTIE 2 (Langmead and Salzberg, 2012) with default settings to count the number of reads mapped to each contig. We also accounted for read length and contig length, as follows: coverage = the number of reads that mapped to the contig × the 100 bp read length/the length of the contig in bp.

Estimates of genome completeness and phylogenetic analyses

A list of 30 conserved, single-copy genes was identified, based on the literature (Wu and Eisen, 2008; Wu *et al*., 2009), specifically, *gyrA*, *gyrB*, *recA*, *radA* and ribosomal protein genes (rp) L1, L2, L3, L4, L5, L6, L10, L11, L13, L14, L15, L16, L18, S2, S3, S4, S7, S8, S9, S10, S11, S12, S13, S15, S17 and S19. The percentage of these genes present in each CG2_0.2A ESOM bin was used to assess the completeness of each genome, as described above (Table S5).

A multi-tiered approach was used to assess the phylogeny of genomes binned from CG2_0.2A. For binned genomes with an assigned taxonomic affiliation, all approaches converged on the same phylogeny and taxonomic affiliation, unless otherwise stated in the text. Specifically, we used protein trees predicted from single-copy genes where possible, particularly from ribosomal protein S3 (RpS3) and DNA GyrA. For each tree, predicted protein sequences from previously sequenced genomes were used as references, including representatives of all branches of the *Proteobacteria*, all candidate bacterial phyla for which protein sequences for a given gene were available,

and all bacterial and archaeal groups predicted to be in the geyser system, based on annotation and 16S rRNA gene analyses. For 10 bins (Bins 1, 2a, 3, 4, 5, 6, 8, 11, 14 and 17) that had at least five of 10 selected ribosomal proteins (L5, L6, L15, L16, L18, L22, L24, S3, S8, S19, chosen because they were present in most bins), a concatenated protein tree was generated from these predicted protein sequences to confirm phylogeny (data not shown). As a final source of information, the taxonomic affiliation of best BLAST hits for each predicted protein in the annotation for each genome bin was used to corroborate phylogeny and/or to assist in predicting the taxonomic affiliation of binned genomes without a clear affiliation based on the methods just described. Genome bins predicted to belong to unknown or little genomically sampled lineages were required to: (i) affiliate with different taxonomic groups (different at the phylum level in the case of genome bins predicted to be undefined at the phylum level) in different single-copy gene predicted protein trees; and (ii) have best BLAST hits to many different phyla throughout the annotation.

We used the EMIRGE algorithm (Miller *et al*., 2011) to reconstruct near full-length 16S rRNA gene sequences from CG2_0.2A, CG2_0.2B and CG2_3.0 sequencing reads. All paired reads for which both pairs were > 60 bp in length after quality trimming were included in the analysis. Because of differences in sequencing throughput, we performed 20 iterations for CG2_0.2B and CG2_3.0 and 40 iterations for CG2_0.2A. The resulting sequences were phylogenetically characterized, based on the following, in priority order: (i) SILVA (SINA online) nucleotide alignments and the associated SINA sequence classifications (Pruesse *et al*., 2007), with minimum identity to the query sequence set to 0.95 and 10 neighbors per query sequence; and (ii) for sequences that were ‘unclassified’ through SINA, we used best BLAST hits to NCBI's nr and 16S rRNA gene databases for identification (online BLASTn search, conducted on 12 April 2013). The relative abundance of each OTU was calculated statistically via the EMIRGE algorithm, based on ‘prior probabilities’ of read coverage depth (see Miller *et al*., 2011). Because some of the EMIRGE sequences from CG2_0.2A were used for phylogenetic analyses, chimera detection was performed via DECIPHER (Wright *et al*., 2012) on the 12 EMIRGE sequences from CG2_0.2A. One putative chimera was detected, but upon manual evaluation of the BLAST alignment for the putative chimera with its best BLAST hit in NCBI's nr database (94% similar to *S. lithotrophicus* ES-1 across 98% of the sequence, as described above), we determined that the sequence was not chimeric (see Table S1).

Acknowledgements

We thank Maxwell Rudolph (University of California Berkeley) for field assistance, Laura Hug (University of California Berkeley) for phylogenomics assistance, Sean McAllister (University of Delaware) for providing zetaproteobacterial 16S rRNA gene sequences and Joern Larsen and his group (Lawrence Berkeley National Laboratory) for generating the solution

chemistry data. Funding for this work was provided by Department of Energy (DOE) Award DE-AC02-05CH11231.

References

Badger, M.R., and Bek, E.J. (2008) Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO_2 acquisition by the CBB cycle. *J Exp Botany* 59: 1525– 1541.

Baer, J.L., and Rigby, J.K. (1978) Geology of the Crystal Geyser and environmental implications of its effluent. *Utah Geology* 5: 125– 130.

Barnes, F.A. (1996) *Canyon County Explorer*. Moab, UT, USA: Canyon Country Publications.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., *et al*. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315: 1709– 1712.

Bassham, J.A., Benson, A.A., and Calvin, M. (1950) The path of carbon in photosynthesis. *J Biolog Chem* 185: 781– 787.

Beller, H.R., Letain, T.E., Chakicherla, A., Kane, S.R., Legler, T.C., and Coleman, M.A. (2006a) Whole-genome transcriptional analysis of chemolithoautotrophic thiosulfate oxidation by *Thiobacillus denitrificans* under aerobic versus denitrifying conditions. *J Bacteriol* 188: 7005– 7015.

Beller, H.R., Chain, P.S.G., Letain, T.E., Chakicherla, A., Larimer, F.W., Richardson, P.M., *et al*. (2006b) The genome sequence of the obligately chemolithoautotrophic, facultatively anaerobic bacterium *Thiobacillus denitrificans*. *J Bacteriol* 188: 1473– 1488.

Bickle, M., and Kampman, N. (2013) Lessons in carbon storage from geological analogues. *Geology* 41: 525– 526.

Blöthe, M., and Roden, E.E. (2009) Composition and activity of an autotrophic Fe(II)-oxidizing, nitrate-reducing enrichment culture. *Appl Environ Microbiol* 75: 6937– 6940.

Boone, D.R., Liu, Y., Zhao, Z.J., Balkwill, D.L., Drake, G.R., Stevens, T.O., and Aldrich, H.C. (1995) *Bacillus infernus* sp. nov., an Fe(III)- and Mn(IV)-reducing anaerobe from the deep terrestrial subsurface. *Intl J System Bacteriol* 45: 441– 448.

Brown, M., and Balkwill, D. (2009) Antibiotic resistance in bacteria isolated from the deep terrestrial subsurface. *Microb Ecol* 57: 484– 493.

Burnside, N.M., Shipton, Z.K., Dockrill, B., and Ellam, R.M. (2013) Man-made versus natural CO_2 leakage: a 400 k.y. history of an analogue for engineered geological storage of CO2. *Geology* 41: 471– 474.

Campbell, J.H., O'Donoghue, P., Campbell, A.G., Schwientek, P., Sczyrba, A., Woyke, T., *et al*. (2013) UGA is an additional glycine codon in uncultured

SR1 bacteria from the human microbiota. *Proc Natl Acad Sci* 110: 5540– 5545.

Castelle, C.J., Hug, L., Wrighton, K.C., Thomas, B.C., Williams, K.H., Wu, D., *et al*. (2013) Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Comm* 4: 2120.

Chan, C.S., Fakra, S.C., Emerson, D., Fleming, E.J., and Edwards, K.J. (2011) Lithotrophic iron-oxidizing bacteria produce organic stalks to control mineral growth: implications for biosignature formation. *ISME J* 5: 717– 727.

Chandler, D.P., Brockman, F.J., Bailey, T.J., and Fredrickson, J.K. (1998) Phylogenetic diversity of archaea and bacteria in a deep subsurface paleosol. *Microb Ecol* 36: 37– 50.

Colwell, F.S., and D'Hondt, S. (2013) Nature and extent of the deep biosphere. *Rev Mineralogy and Geochem* 75: 547– 574.

Comolli, L.R., Luef, B., and Chan, C.S. (2011) High-resolution 2D and 3D cryo-TEM reveals structural adaptations of two stalk-forming bacteria to an Fe-oxidizing lifestyle. *Environ Microbiol* 13: 2915– 2929.

Dang, H., Chen, R., Wang, L., Shao, S., Dai, L., Ye, Y., *et al*. (2011) Molecular characterization of putative biocorroding microbiota with a novel niche detection of *Epsilon*- and *Zetaproteobacteria* in Pacific Ocean coastal seawaters. *Environ Microbiol* 13: 3059– 3074.

Dick, G., Andersson, A., Baker, B., Simmons, S., Thomas, B., Yelton, A.P., and Banfield, J. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10: R85.

Doelling, H.H. (2002) Interim geologic map of the San Rafael Desert, Emery and Grand Counties, Utah. Utah Geological Survey, Open-File Report, pp. 404.

Edwards, K.J., Becker, K., and Colwell, F. (2012) The deep, dark energy biosphere: intraterrestrial life on Earth. *Ann Rev Earth and Planet Sci* 40: 551– 568.

Emerson, D., and Ghiorse, W.C. (1993) Ultrastructure and chemical composition of the sheath of *Leptothrix discophora* SP-6. *J Bacteriol* 175: 7808– 7818.

Emerson, D., and Moyer, C. (1997) Isolation and characterization of novel iron-oxidizing bacteria that grow at circumneutral pH. *Appl Environ Microbiol* 63: 4784– 4792.

Emerson, D., Rentz, J.A., Lilburn, T.G., Davis, R.E., Aldrich, H., Chan, C., and Moyer, C.L. (2007) A novel lineage of *Proteobacteria* involved in formation of marine Fe-oxidizing microbial mat communities. *PLoS ONE* 2: e667.

Emerson, D., Fleming, E.J., and McBeth, J.M. (2010) Iron-oxidizing bacteria: an environmental and genomic perspective. *Ann Rev Microbiol* 64: 561– 583.

Feng, L., Wang, W., Cheng, J., Ren, Y., Zhao, G., Gao, C., *et al*. (2007) Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *Proc Natl Acad Sci* 104: 5602– 5607.

Field, E.K., Sczyrba, A., Lyman, A.E., Harris, C.C., Woyke, T., Stepanauskas, R., *et al*. (2014) Genomic insights into uncultivated marine *Zetaproteobacteria* at Loihi Seamount. doi:10.1038/ismej.2014.183.

Gouveia, F.J., and Friedmann, S.J. (2006) Timing and prediction of CO_2 eruptions from Crystal Geyser, UT. Lawrence Livermore National Laboratory, Open-File Report. [WWW document]. URL https://e-reports-ext.llnl.gov/pdf/334382.pdf.

Hallbeck, L., and Pedersen, K. (1991) Autotrophic and mixotrophic growth of *Gallionella ferruginea*. *J Gen Microbiol* 137: 2657– 2661.

Han, W.S., Lu, M., McPherson, B.J., Keating, E.H., Moore, J., Park, E., *et al*. (2013) Characteristics of CO_2-driven cold-water geyser, Crystal Geyser in Utah: experimental observation and mechanism analyses. *Geofluids* 13: 283– 297.

Hanke, A., Hamann, E., Sharma, R., Geelhoed, J.S., Hargesheimer, T., Kraft, B., *et al*. (2014) Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between *Alpha*- and *Gammaproteobacteria* in a tidal sediment microbial community naturally selected in a laboratory chemostat. *Frontiers Microbiol* 5: 231.

Hintze, L.F. (1988) Geologic history of Utah. In *Department of Geology*. Provo, UT, USA: Bringham Young University, p. 203.

Hyatt, D., Chen, G.L., LoCascio, P., Land, M., Larimer, F., and Hauser, L. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.

Jansen, R., Embden, J.D., Gaastra, W., and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Molec Microbiol* 43: 1565– 1575.

Jiao, N., Herndl, G.J., Hansell, D.A., Benner, R., Kattner, G., Wilhelm, S.W., *et al*. (2010) Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nat Rev Microbiol* 8: 593– 599.

Kampman, N., Maskell, A., Bickle, M.J., Evans, J.P., Schaller, M., Purser, G., *et al*. (2013) Scientific drilling and downhole fluid sampling of a natural CO_2 reservoir, Green River, Utah. *Sci Drill* 16: 33– 43.

Kampman, N., Bickle, M.J., Maskell, A., Chapman, H.J., Evans, J.P., Purser, G., *et al*. (2014) Drilling and sampling a natural CO_2 reservoir: implications for

fluid flow and CO_2-fluid-rock reactions during CO_2 migration through the overburden. *Chem Geol* 369: 51– 82.

Kantor, R.S., Wrighton, K.C., Handley, K.M., Sharon, I., Hug, L.A., Castelle, C.J., *et al*. (2013) Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* 4: e708– e713.

Kovacik, W.P., Takai, K., Mormile, M.R., McKinley, J.P., Brockman, F.J., Fredrickson, J.K., and Holben, W.E. (2006) Molecular analysis of deep subsurface Cretaceous rock indicates abundant Fe(III)- and S-reducing bacteria in a sulfate-rich environment. *Environ Microbiol* 8: 141– 155.

Krepski, S.T., Hanson, T.E., and Chan, C.S. (2012) Isolation and characterization of a novel biomineral stalk-forming iron-oxidizing bacterium from a circumneutral groundwater seep. *Environ Microbiol* 14: 1671– 1680.

Kulajta, C., Thumfart, J.O., Haid, S., Daldal, F., and Koch, H.G. (2006) Multi-step assembly pathway of the cbb3-type cytochrome c oxidase complex. *J Molec Biol* 355: 989– 1004.

Langmead, B., and Salzberg, S.L. (2012) Fast-gapped read alignment with Bowtie 2. *Nat Methods* 9: 357– 359.

McAllister, S.M., Davis, R.E., McBeth, J.M., Tebo, B.M., Emerson, D., and Moyer, C.L. (2011) Biodiversity and emerging biogeography of the neutrophilic iron-oxidizing *Zetaproteobacteria*. *Appl Environ Microbiol* 77: 5445– 5457.

McBeth, J.M., Fleming, E.J., and Emerson, D. (2013) The transition from freshwater to marine iron-oxidizing bacterial lineages along a salinity gradient on the Sheepscot River, Maine, USA. *Environ Microbiol* 5: 453– 463.

Miller, C., Baker, B., Thomas, B., Singer, S., and Banfield, J. (2011) EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* 12: R44.

Mitchell, A.C., Dideriksen, K., Spangler, L.H., Cunningham, A.B., and Gerlach, R. (2010) Microbially enhanced carbon capture and storage by mineral-trapping and solubility-trapping. *Environ Sci & Tech* 13: 5270– 5276.

Morozova, D., Wandrey, M., Alawi, M., Zimmer, M., Vieth, A., Zettlitzer, M., and Wuerdemann, H. (2010) Monitoring of the microbial community composition in saline aquifers during CO_2 storage by fluorescence in situ hybridisation. *Int J Greenhouse Gas Control* 4: 981– 989.

Morozova, D., Zettlitzer, M., Let, D., and Wuerdemann, H. (2011) Monitoring of the microbial community composition in deep subsurface saline aquifers during CO_2 storage in Ketzin, Germany. *Energy Procedia* 4: 4362– 4370.

Mu, A., Boreham, C., Leong, H.X., Haese, R., and Moreau, J.W. (2014) Changes in the deep subsurface microbial biosphere resulting from a field-scale CO_2 geosequestration experiment. *Frontiers Microbiol* 5: 209.

Naftz, D.L., Peterman, Z.E., and Spangler, L.E. (1997) Using δ87 Sr values to identify sources of salinity to a freshwater aquifer, Greater Aneth Oil Field, Utah, USA. *Chem Geol* 141: 195– 209.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012) IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420– 1428.

Probst, A.J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J.B., Rattei, T., *et al*. (2014) Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat Commun* 5: 5497.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Ac Res* 35: 7188– 7196.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499: 431– 437.

Roden, E.E., Sobolev, D., Glazer, B., and Luther, G.W. (2004) Potential for microscale bacterial Fe redox cycling at the aerobic-anaerobic interface. *Geomicrobiol J* 21: 379– 391.

Sahl, J.W., Schmidt, R., Swanner, E.D., Mandernack, K.W., Templeton, A.S., Kieft, T.L., *et al*. (2008) Subsurface microbial diversity in deep-granitic-fracture water in Colorado. *Appl Environ Microbiol* 74: 143– 152.

Sato, T., Atomi, H., and Imanaka, T. (2007) Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science* 315: 1003– 1006.

Schrag, D.P. (2007) Preparing to capture carbon. *Science* 315: 812– 813.

Scott, K.M., Sievert, S.M., Abril, F.N., Ball, L.A., Barrett, C.J., Blake, R.A., *et al*. (2006) The genome of deep-sea vent chemolithoautotroph *Thiomicrospira crunogena* XCL-2. *PLoS Biol* 4: e383.

Shipton, Z.K., Evans, J.P., Kirschner, D., Kolesar, P.T., Williams, A.P., and Heath, J. (2004) Analysis of CO_2 leakage through low-permeability faults from natural reservoirs in the Colorado Plateau, east-central Utah. *Geol Soc, London, Spec Publ* 233: 43– 58.

Singer, E., Emerson, D., Webb, E.A., Barco, R.A., Kuenen, J.G., Nelson, W.C., *et al*. (2011) *Mariprofundus ferrooxydans* PV-1 the first genome of a marine Fe(II) oxidizing Zetaproteobacterium. *PLoS ONE* 6: e25386.

Singer, P.C., and Stumm, W. (1970) Acidic mine drainage: the rate-determining step. *Science* 167: 1121– 1123.

Stein, L.Y., Arp, D.J., Berube, P.M., Chain, P.S.G., Hauser, L., Jetten, M.S.M., *et al*. (2007) Whole-genome analysis of the ammonia-oxidizing

bacterium, *Nitrosomonas eutropha* C91: implications for niche adaptation. *Environ Microbiol* 9: 2993– 3007.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282– 1288.

Szewzyk, U., Szewzyk, R., and Stenstrom, T.A. (1994) Thermophilic, anaerobic bacteria isolated from a deep borehole in granite in Sweden. *Proc Natl Acad Sci* 91: 1810– 1813.

Tabita, F.R., Hanson, T.E., Li, H., Satagopan, S., Singh, J., and Chan, S. (2007) Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol Molec Biol Rev* 71: 576– 599.

Tabita, F.R., Hanson, T.E., Satagopan, S., Witte, B.H., and Kreel, N.E. (2008) Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. *Phil Trans Royal Soc B: Biol Sci* 363: 2629– 2640.

Trias, R., Gerard, E., le Campion, P., Aradottir, E.S., Gunnarsson, I., Gislason, S.R., *et al*. (2014) Anthropogenic CO_2-H_2S-H_2 injections in deep basalts stimulate biofilm development and microbially-induced mineralizations that favour reservoir clogging. Abstract, 9th International Symposium on Subsurface Microbiology, October 5–10, Pacific Grove, California.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., *et al*. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: 37– 43.

Ultsch, A., and Moerchen, F. (2005) ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report. Marburg, Germany: Department of Mathematics and Computer Science, University of Marburg, Germany.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66– 74.

Waltham, T. (2001) Crystal Geyser–Utah's cold one. *Geology Today* 17: 22– 24.

Wandrey, M., Pellizari, L., Zettlitzer, M., and Wuerdemann, H. (2011a) Microbial community and inorganic fluid analysis during CO_2 storage within the frame of CO_2SINK – long-term experiments under in situ conditions. *Energy Procedia* 4: 3651– 3657.

Wandrey, M., Fischer, S., Zemke, K., Liebscher, A., Scherf, A.K., Vieth-Hillebrand, A., *et al*. (2011b) Monitoring petrophysical, mineralogical, geochemical and microbiological effects of CO2 exposure – results of long-term experiments under in situ conditions. *Energy Procedia* 4: 3644– 3650.

West, J.M., McKinley, I.G., Palumbo-Roe, B., and Rochelle, C.A. (2011) Potential impact of CO_2 storage on subsurface microbial ecosystems and implications for groundwater quality. *Energy Procedia* 4: 3163– 3170.

Wilkinson, M., Gilfillan, S.M.V., Haszeldine, R.S., and Ballentine, C.J. (2009) Plumbing the depths: testing natural tracers of subsurface CO2 origin and migration, Utah, USA. In *Carbon Dioxide Sequestration in Geological Media – State of the Science*. M. Grobe, J.C. Pashin, and R.L. Dodge (eds). Washington, D.C., USA: The American Association of Petroleum Geologists, pp. 619– 634.

Wright, E.S., Yilmaz, L.S., and Noguera, D.R. (2012) DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl Environ Microbiol* 78: 717– 725.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., *et al*. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337: 1661– 1665.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., *et al*. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462: 1056– 1060.

Wu, M., and Eisen, J. (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9: R151.

Yokoyama, K., Ohkuma, S., Taguchi, H., Yasunaga, T., Wakabayashi, T., and Yoshida, M. (2000) V-type H^+-ATPase/synthase from a thermophilic eubacterium, *Thermus thermophilus*. *J Biol Chem* 275: 13955– 13961.