Title

N- and C-cycling organisms in the subsurface

Permalink
https://escholarship.org/uc/item/2f1480x2

Journal
Environmental Microbiology, 18(1)

ISSN
1462-2912

Authors
Hug, Laura A
Thomas, Brian C
Sharon, Itai
et al.

Publication Date
2016

DOI
10.1111/1462-2920.12930

Peer reviewed

Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages

Laura A. Hug

Brian C. Thomas

Itai Sharon

Christopher T. Brown

Ritin Sharma

Robert L. Hettich
Michael J. Wilkins

Kenneth H. Williams

Andrea Singh

Jillian F. Banfield
First published: 01 June 2015
https://doi.org/10.1111/1462-2920.12930
Cited by: 34
UC-eLinks

Summary

Nitrogen, sulfur and carbon fluxes in the terrestrial subsurface are determined by the intersecting activities of microbial community members, yet the organisms responsible are largely unknown. Metagenomic methods can identify organisms and functions, but genome recovery is often precluded by data complexity. To address this limitation, we developed subsampling assembly methods to re-construct high-quality draft genomes from complex samples. We applied these methods to evaluate the interlinked roles of the most abundant organisms in biogeochemical cycling in the aquifer sediment. Community proteomics confirmed these activities. The eight most abundant organisms belong to novel lineages, and two represent phyla with no previously sequenced genome. Four organisms are predicted to fix carbon via the Calvin–Benson–Bassham, Wood–Ljungdahl or 3-hydroxyproprionate/4-hydroxybutarate pathways. The profiled organisms are involved in the network of denitrification, dissimilatory nitrate reduction to ammonia, ammonia oxidation and sulfate reduction/oxidation, and require substrates supplied by other community members. An ammonium-oxidizing Thaumarchaeote is the most abundant community member, despite low ammonium concentrations in the groundwater. This organism likely benefits from two other relatively abundant organisms capable of producing ammonium from nitrate, which is abundant in the groundwater. Overall, dominant members of the microbial community are interconnected through exchange of geochemical resources.

Introduction

Microbial metabolism is critical to the creation, maintenance and turnover of carbon and nitrogen sinks in the subsurface. The high proportion of uncultured and little-studied organisms present in subsurface environments (Wrighton *et al*., 2012; Castelle *et al*., 2013), and the continued discovery of new microbial lineages implicated in major geochemical cycles (Green *et al*., 2010; Rasigraf *et al*., 2014) indicates that there is substantial metabolic diversity yet to be discovered within terrestrial sediments. Understanding and modelling geochemical cycling thus requires identification of, and metabolic prediction for, the microbial networks catalysing carbon, nitrogen and sulfur cycling in the subsurface.

The complexity of sediment microbial communities had prevented substantial metagenomic assembly until recently; with advances in sequencing technologies and bioinformatic methods, insight into the structure and function of diverse, low abundance sediment communities is now tractable through assembly and genome curation (Castelle *et al*., 2013; Kantor *et al*., 2013). Metagenomics provides simultaneous taxonomic identification and metabolic profiling for the community, or, with binning and genome curation, for specific organisms from the environment. Draft genomes from sediment-associated organisms have allowed definition of new radiations on the tree of life (Ettwig *et al*., 2009; Wrighton *et al*., 2012; Castelle *et al*., 2013; Rinke *et al*., 2013), and prediction of previously unknown roles in biogeochemical cycles, including carbon fixation in *Chloroflexi* (Hug *et al*., 2013) and an abundance of fermentative processes in the saturated subsurface (Wrighton *et al*., 2012; 2014). Microbial community compositions from sediments separated by tens of metres can share little to no overlap at the species level (Wu *et al*., 2008; Hug *et al*., 2015), meaning current work has only scratched the surface of the true metabolic diversity of these environments.

Assembly and genome curation from metagenomic data are continually improving, including accommodation of uneven depths of coverage (Boisvert *et al*., 2012; Namiki *et al*., 2012; Peng *et al*., 2012) and better resolution of genome bins by combining physical characteristics (e.g., nucleotide composition) with time series coverage information (Dick *et al*., 2009; Albertsen *et al*., 2013; Sharon *et al*., 2013). This is helping to resolve the genomic composition of complex communities. Despite these improvements, genomes from higher abundance organisms tend to fare poorly during assembly (Handley *et al*., 2014; Sharon *et al*., 2015). The under-representation or absence of higher abundance organisms in metagenomic assemblies can lead to inaccurate descriptions of community composition, as well as missed or incomplete metabolic pathways important to community functions.

Here we conducted deep metagenomic sequencing of a sediment core from an alluvial aquifer adjacent to the Colorado River near the town of Rifle, CO, USA. The sediment was derived from a region unaffected by previous acetate amendment experiments within the aquifer (Williams *et al*., 2011), allowing examination of the microbial community under ‘native’ conditions. Subsampling assembly methods were applied to improve re-construction of the more abundant organisms' genomes. Metabolic predictions regarding roles played by these organisms in carbon, nitrogen and sulfur cycles were bolstered by metaproteomic detection of protein expression. The organisms profiled represent previously unstudied lineages, and two are from newly identified phyla. The eight organisms are predicted to play interlinked roles in nitrogen, sulfur and carbon cycling.

Results and discussion

Sediment community metagenomics

A sediment core was drilled in the Rifle, CO, USA, aquifer on 20 July 2011; the location is up-gradient from where previous acetate amendment and other perturbation experiments have been conducted. A monitoring well installed post-drilling allowed detailed geochemical analysis of the groundwater from the site (Table S1). Notably, the nitrate concentration was high in comparison with other nitrogen species [nitrate = 6.4 mg l^{-1} (103 μM), nitrite = 0.144 mg l^{-1} (3.13 μM), ammonium = 0.03 mg l^{-1} (1.67 μM)], as well as in comparison with nitrate concentrations at other monitoring wells in the same time frame (range 0–5.5 mg l^{-1}, average 1.4 mg l^{-1}). The sulfate concentration was 8.5 mM, and the iron (II) concentration low at 0.01 mg l^{-1} (0.18 μM).

Sediment samples were taken from depths of 3, 4, 5 and 6 m, with multiple independent DNA extractions yielding replicate samples for the 5 m depth. In total, seven lanes of Illumina HiSeq sequencing were utilized in this study: one lane for each of the 3, 4 and 6 m depths, and four lanes for the 5 m depth samples. The total sequence generated was 314 Gbp, with an average of 44.9 ± 3.8 Gbp per sample (Table S2). The 5 m sample #4 (5m_4) was selected to facilitate comparisons with a previous sample at 5 m depth from a distant location in the aquifer (Castelle *et al*., 2013; Hug *et al*., 2013; Sharon *et al*., 2015). The 5m_4 sequence data were assembled for community composition and metabolic potential analyses.

Subsampling reads improves assembly of high coverage genomes

Subsampling the 5m_4 sequence data set was conducted to examine the effect of using smaller, less computationally intensive data sets for assembly. We used two approaches. First, we conducted a series of independent experiments using fractions of the reads independently sampled from the 5m_4 data set (i.e., with replacement). Second, we conducted sequential assemblies in which reads that assembled into contigs were removed from the read data set prior to subsequent samplings and assemblies (i.e., without replacement). Initial experiments with replacement were used to determine that 5% of the sequence data (17 M reads) was the minimum amount of sequence that would yield a useful assembly (here, defined as > 10 Mbp assembled, at least one contig > 20 000 bp). Notably, although this assembly generated a small amount of sequence compared with the full assembly (1317 Mbp), the highest coverage scaffolds from this assembly were not represented in the full data set assembly. A similar approach involving the use of a subset of the data to target high abundance genomes was used to reconstruct a draft genome of an *Epsilonproteobacterium* from Rifle sediment (Handley *et al*., 2014). In that study, subsamples of a mere 0.3% of the multiple displacement amplified sequence data resulted in significantly improved initial genome reconstruction.

The experiments conducted without read replacement used an initial 5% subsample, followed by 10%, 20%, 33% and 50% subsamples. Results showed that doubling the amount of data assembled more than doubled the amount of assembled sequence generated (Fig. 1). Further, the community compositions of the different subassemblies varied widely (Fig. 1). The subassemblies were merged and compared with the full assembly that was performed in a single step. Results showed increased number and length of scaffolds with coverage values above 40× in the subassembly based data set (Fig. 2, main). From the subsampled assemblies, a coverage level of 15–25× appears optimal for scaffold construction and extension. The effect of strain variation was examined, but the proportion of single nucleotide polymorphisms (SNPs) present in the subsampled reads versus the total reads were not substantially different. While the subsampled assemblies did not contribute a large proportion of the total assembled sequence length, they enabled reconstruction of higher abundance organisms' genomes that were otherwise not present (Fig. 2, inset). From this, subsampling cannot be considered an acceptable replacement for assembling the complete data set, but instead provides complementary information for better coverage of the total community.

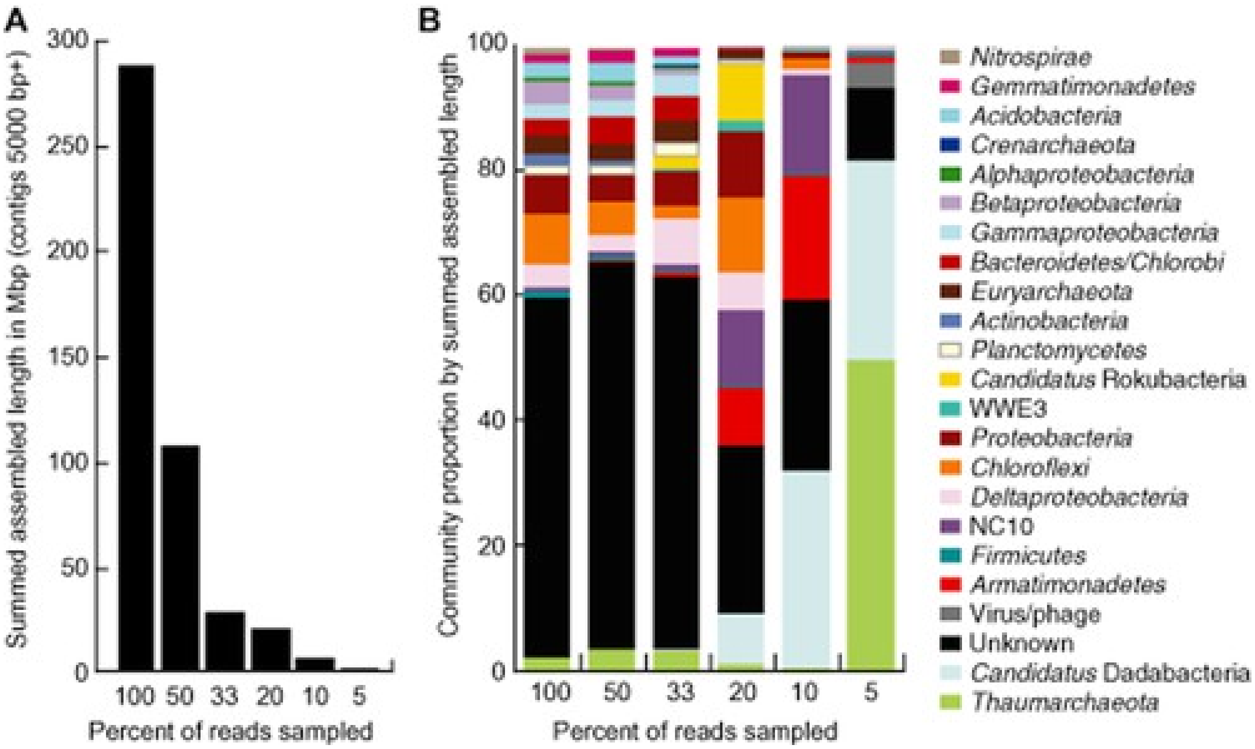

Figure 1

Open in figure viewerPowerPoint

Subsampling assembly methods yield proportionally smaller assemblies, but target different community members.

A. Summed length of assemblies from the 5m_4 sample sequence data under varying degrees of subsampling. Smaller samples yield shorter overall assemblies.

B. Predicted community composition of each assembly from A, coloured by phylum, with proportion defined by summed length of contigs assigned to that phylum. Deeper subsampling dramatically changes the community composition.

Caption

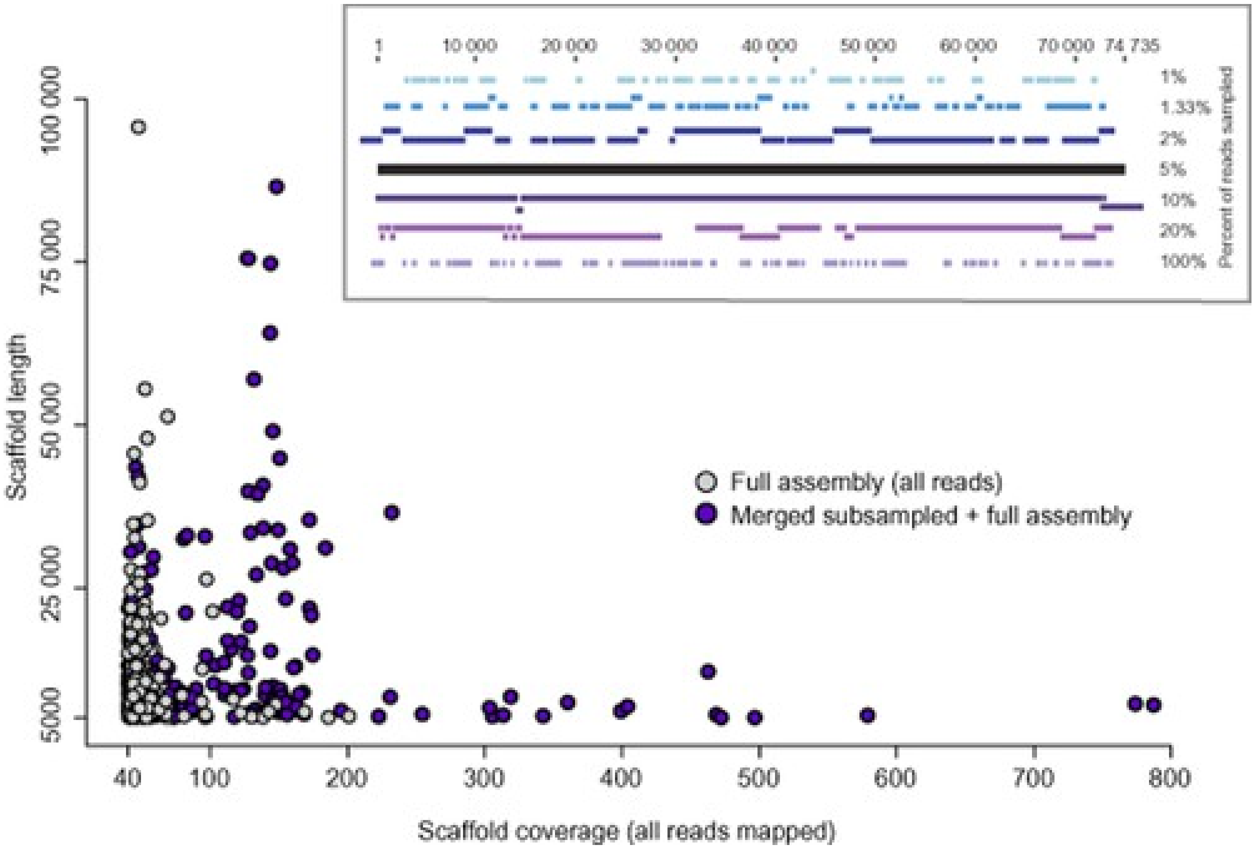

Figure 2

Open in figure viewerPowerPoint

Subsampling assembly methods yield substantially better assembly of abundant community members. Main: coverage versus length plot for scaffolds from the full assembly (grey) and the merged subsampled and full assembly (purple). To highlight higher coverage scaffolds, only those with coverage > 40× and length > 5000 bp were plotted. Shared scaffolds between the two conditions will appear purple based on order of point plotting. Higher coverage scaffolds are over-represented in the subsampled assemblies. Inset: example of the effect of subsampling on reconstruction of the longest scaffold from the 5% sample assembly. The scaffold depicted belongs to the CSP1-1 *Thaumarchaeota* genome, present at ∼ 125× coverage in the full sample. Deeper subsampling resulted in fragmented assembly of this region because of lack of coverage, whereas shallower subsamples and the full assembly also resulted in highly fragmented assembly for this region. For this analysis, the subsampling with replacement of reads method was utilized.

Caption

Draft genomes for abundant organisms representing novel lineages

The sediment microbial community at 5 m depth in the aquifer is complex, characterized by numerous low abundance organisms from many phyla (Fig. 3). In total, 133 individual organisms were assigned taxonomy through phylogenetic analysis of concatenated ribosomal protein alignments, generated from 16 genes located within a syntenic block on most genomes. The use of this set of genes precludes the need for genome binning (Hug *et al*., 2013) (Fig. S1). *Chloroflexi* comprised both the most numerous and most abundant phylum, followed by *Thaumarchaeota* and the class *Deltaproteobacteria* (Fig. 3).

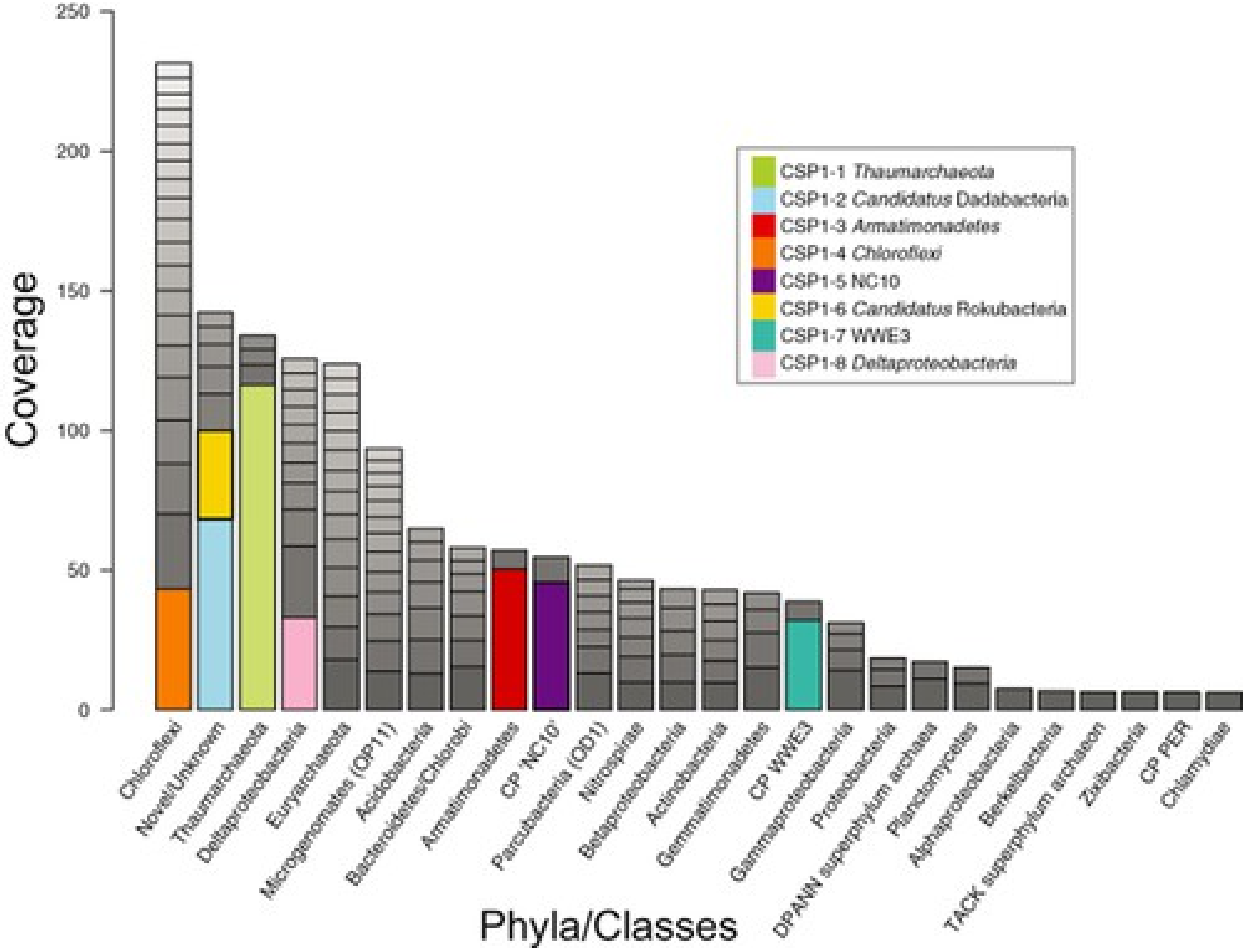

Figure 3

Open in figure viewerPowerPoint

Unamended sediment contains a diverse, even microbial community. The bar plot contains the 133 most abundant organisms from the 5m_4 metagenome assembly, organized by phylum-level lineages with individual organism abundances reflected in the size of the stacked boxes. Read depth (coverage) of the scaffold containing the ribosomal protein syntenic block is plotted as a proxy for organism abundance. Note that the organisms classified as ‘Novel/Unknown’ are not related to one another. The eight genomes selected for curation are highlighted in colour. The most abundant organism, CSP1-1 in light green, is a *Thaumarchaeota* that recruits 0.7% of the total reads to its curated genome.

Caption

All scaffolds from the merged assembly that represent the most abundant organisms (with > 20× coverage) were targeted for hybrid nucleotide composition and time series abundance emergent self-organizing map (ESOM)-based binning (see Experimental Procedures). This yielded nine preliminary genome bins (Fig. S2), one of which (a *Chloroflexi*) was discarded based on estimates that the genome was < 50% complete. Post-curation, the eight genome bins contained between 82% and 100% of the expected single copy marker genes, with an average completeness of 93.7% (Table 1, Table S3). We determined the phylogenetic placement of the eight genomes, taking into consideration the recently suggested definition of 75% 16S rRNA nucleotide identity as a boundary for phylum-level lineages (Yarza *et al*., 2014). However, we made two exceptions in cases where the organism's 16S rRNA sequence shared similar low identity levels with sequences from multiple phyla and had inconsistent placement across different marker gene phylogenies. Importantly, these lineages are clearly defined in phylogenetic analysis of both the 16S rRNA gene and concatenated ribosomal proteins.

Table 1. Overview of the eight curated genomes, including genome size, number of scaffolds and taxonomic placement

Bin name	Taxonomy	Total assembly length (Mb)	#Scaffolds	% Complete
CSP1-1	*Archaea*; *Thaumarchaeota*	1.4	10	95.9
CSP1-2	*Bacteria*; *Candidatus* Dadabacteria	2.5	92	94.2
CSP1-3	*Bacteria*; *Armatimonadetes*	2.5	162	95.7
CSP1-4	*Bacteria*; *Chloroflexi*	2.6	66	97.1
CSP1-5	*Bacteria*; NC10	2.8	45	100

Bin name	Taxonomy	Total assembly length (Mb)	#Scaffolds	% Complete
CSP1-6	*Bacteria*; *Candidatus* Rokubacteria	2.9	166	92.8
CSP1-7	*Bacteria*; WWE3	0.6	7	82.6
CSP1-8	*Bacteria*; *Deltaproteobacteria*	2.2	52	91.3

- Percent completeness was determined using a set of 69 single copy marker genes for bacteria and 49 for the archaeon (see Table S3 for marker gene information).

The most abundant organism, CSP1-1, is a member of the *Thaumarchaeota*, an archaeal phylum identified from a variety of environments and characterized by ammonium oxidization (Stahl and de la Torre, 2012). On a 16S ribosomal RNA gene phylogeny it belongs to a clade related to representatives of the Nitrosopumilales and Cenarchaeales orders (Figs S3 and S4). The 16S rRNA sequence shares 95% identity to the nearest sequenced genome, *Candidatus*Nitrosoarchaeum limnia SFB1, a member of the order *Nitrosopumilales*. The CSP1-1 genome is present at ∼ 125× coverage in the metagenome, collecting 0.7% of all sequence reads.

The remaining seven genomes belong to organisms distributed across the bacterial domain (Figs S1 and S3). None of the organisms shares greater than 89% 16S rRNA identity with sequenced genomes. Environmental clone sequences with 94–99% identity to each 16S rRNA are present in the National Center for Biotechnology Information (NCBI) nr database, suggesting these genomes represent prevalent but genomically uncharacterized lineages.

Two of the bacterial genomes belong to organisms that cannot confidently be placed within a currently defined phylum (CSP1-2 and CSP1-6). The 16S rRNA gene from CSP1-2 shares at most 80–81% identity with genes from sequenced genomes, including representatives from three different phyla (*Proteobacteria*, *Firmicutes* and *Thermodesulfobacteria*). The CSP1-2 16S rRNA gene is assigned to the Deltaproteobacterial lineage Sh765B-TzT-29 by the SILVA database. However, the sequence falls within the deepest branching of three separate Sh765B-TzT-29 clades on a 16S rRNA gene tree encompassing all *Deltaproteobacteria* (Fig. S5), and the clade places as a separate lineage well separated from the *Deltaproteobacteria* and all other named phyla on a global 16S rRNA gene phylogeny (Fig. S3). This reflects the observed polyphyly of the *Deltaproteobacteria* (Lang *et al*., 2013), making the SILVA prediction unreliable in this instance. The concatenated ribosomal protein tree places CSP1-2 as a long branch to a clade containing *Thermodesulfobacteria* and *Deltaproteobacteria*. Given the placement of CSP1-2 as a deep branch to all bacterial lineages on the 16S rRNA gene tree, we conclude that this organism represents a novel phylum rather than an extremely divergent Deltaproteobacterial lineage. For it, we propose the name *Candidatus* Dadabacteria CSP1-2, in recognition of the non-traditional methods used to identify this lineage, in-keeping with the Dada art movement ideals of disruption and novelty. Here, *Candidatus* applies both to the organism as well as the potential new phylum lineage, a method to amalgamate environmentally derived genome sequence taxonomic classification with currently accepted nomenclature standards, as proposed by Hedlund and colleagues (2015).

The CSP1-6 16S rRNA gene shares 82% identity with representatives from the *Deltaproteobacteria*, *Firmicutes* and *Nitrospirae* as well as with *Candidatus* Methylomirabilis oxyfera, the type strain for the NC10 phylum. CSP1-6 places as a deep branch related to, but distinct from, the Nitrospirae and the NC10 in the 16S rRNA gene and ribosomal protein phylogenies, respectively, suggesting it represents a new phylum-level lineage. We propose the name *Candidatus* Rokubacteria for the CSP1-6 lineage, from the Sino-Japanese for the number six.

Three of the remaining five bacterial genomes are affiliated with currently understudied phyla: the *Armatimonadetes* (OP10) and the candidate phyla (CP) NC10 and WWE3. The CSP1-3 genome places as a deep branch associated with, but distinct from, the *Armatimonadetes* in both the 16S rRNA gene and concatenated ribosomal protein phylogenies. It shares only 77.8% 16S rRNA gene identity with *Chthonomonas calidirosea*, the type strain for the *Armatimonadetes*, indicating that it likely represents a novel class. Based on SILVA classification, the genome belongs to the GAL15 group, which appears erroneously classified within the *Thermotogae*: we suggest this group should instead be considered part of the *Armatimonadetes* radiation. CSP1-5 shares 89% 16S rRNA identity with *Candidatus* M. oxyfera, and places as a sibling lineage to this organism on both 16S rRNA gene and ribosomal protein phylogenies (Figs S1 and S3). CSP1-5 places within clade D of the NC10, compared with clade A for *Candidatus* M. oxyfera (Fig. S6, Ettwig *et al*., 2009). CSP1-5 is thus the second sequenced genome for the NC10 phylum and also the organism with the closest sequenced relative of the eight genomes examined here. CSP1-7 places within the basal Candidate Phyla Radiation (CPR) comprising the *Parcubacteria*(OD1), *Microgenomates* (OP11), *Gracilibacteria* (BD1-5 and GN-02), PER, WWE3, SR1 and *Saccharibacteria* (TM7) phyla (Wrighton *et al*., 2012; Kantor *et al*., 2013; Rinke *et al*., 2013; Brown *et al*., 2015). The 16S rRNA gene from CSP1-7 shares at most 77% pairwise identity with sequenced genomes from the WWE3 phylum. Based on a 16S rRNA identity above 75%, as well as consistent placement associated with the WWE3, we define this organism as a representative of a novel class-level lineage within the WWE3.

The most abundant *Chloroflexi* (CSP1-4) shares 84–85% 16S rRNA gene identity with members of the order *Dehalococcoidia* and is placed within the Gitt-GS-136 clade in the SILVA *Chloroflexi*taxonomy. The genome places as a branch distinct from previously described *Chloroflexi*genomes from the Rifle site (Hug *et al*., 2013) on both 16S rRNA gene and ribosomal protein trees.

The final bacterial genome curated belongs to a Deltaproteobacterial lineage observed at Rifle previously (Castelle *et al*., 2013; Sharon *et al*., 2015). It shares 87% 16S rRNA identity with members of the orders *Myxococcales* and *Desulfuromonadales*, and consistently places within the *Deltaproteobacteria*, specifically with the 43F-1404R clade in the SILVA 16S rRNA classification. This is the least abundant organism of the eight, with 0.37% of the total 5m_4 sequencing reads allocated to this genome.

Biogeochemical cycling predictions for abundant sediment-associated organisms

Metabolic profiling of the eight most abundant organisms reveals varied but interconnected responses to the environmental conditions. Each organism, not unexpectedly, encodes a unique suite of pathways relevant to global biogeochemical cycles. Interestingly, the metabolic products from the eight organisms are likely utilized by others, forming an interconnected network of activities that promotes the abundance of the examined organisms (Fig. 4).

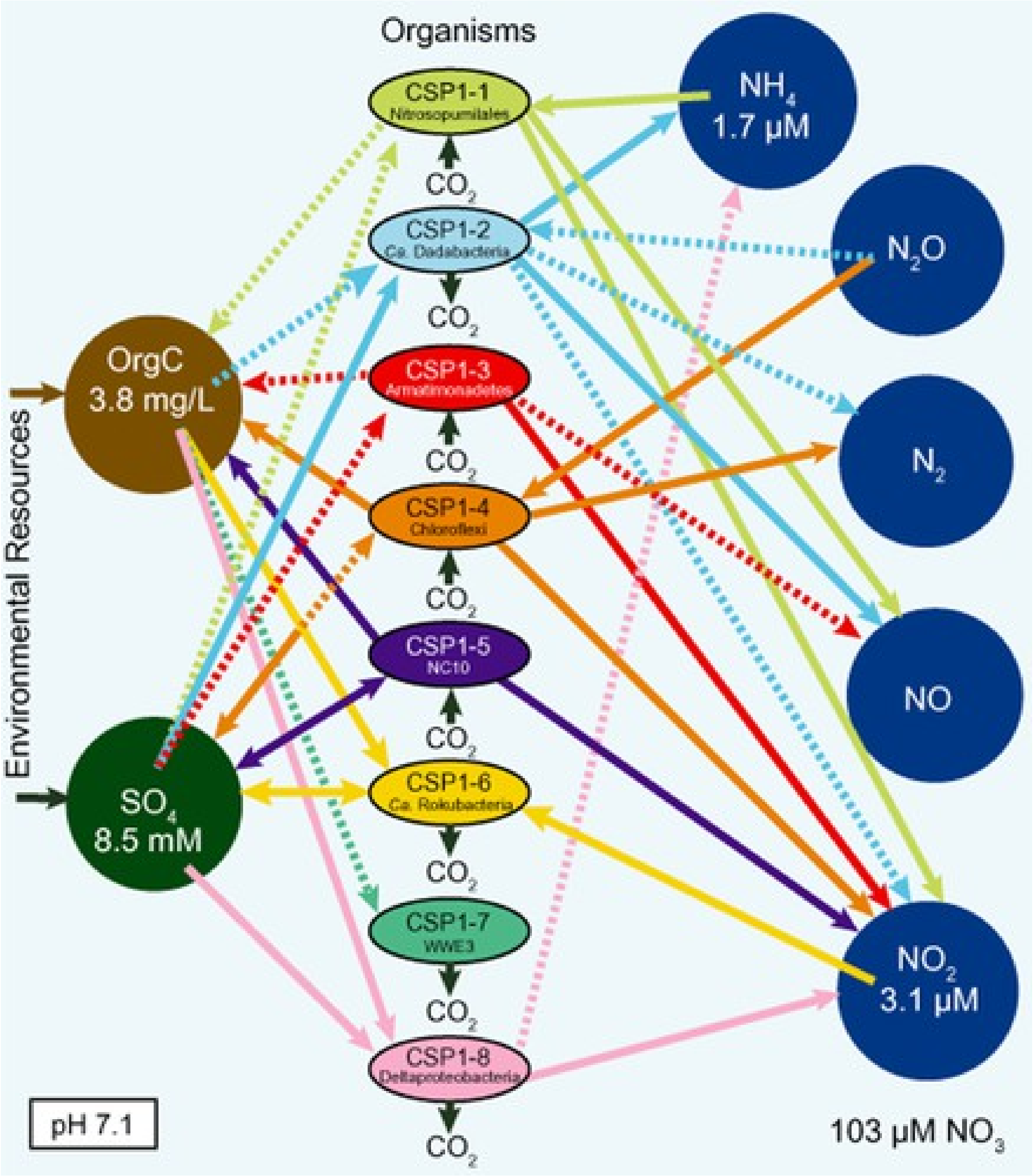

Figure 4

[Open in figure viewerPowerPoint]

Predicted biogeochemical transformations create an interconnected metabolic network of abundant subsurface microbes. Profiled organisms are ovals coloured as in Fig. 3, arrayed vertically in the centre of the diagram. Environmental inputs including organic carbon in the subsurface matrix and local nitrogen and sulfur species are represented as larger nodes on the sides, and are labelled by compound and with local concentrations where available Solid arrows indicate proteins and processes detected from groundwater metaproteomic efforts, dashed arrows correspond to predicted proteins not identified in the proteomic data.

Caption

Carbon

Of the eight profiled organisms, four are predicted to be autotrophic, spanning a variety of carbon fixation processes that convert CO_2 to organic carbon, which is ultimately made available to heterotrophic community members. CSP1-1, the *Thaumarchaeota*, is predicted to be capable of carbon fixation using the archaeal 3-hydroxyproprionate/4-hydroxybutyrate pathway, as described in *Nitrosopumilus maritimus* strain SCM1 (Walker *et al*., 2010). The marine group I *Thaumarchaeota* are thought to be obligate mixotrophs based on metabolic requirements established for two *Thaumarchaeota* isolates (Qin *et al*., 2014), as well as hypotheses from the *N. maritimus* SCM1 genome (Walker *et al*., 2010). The CSP1-1 genome lacks pyruvate kinase and both pyruvate dehydrogenase and pyruvate ferredoxin : oxidoreductase, suggesting glycolysis and other pyruvate-based metabolisms are not operative in this organism. Similarly, the TCA cycle is partial, missing isocitrate dehydrogenase. It is possible this organism is mixotrophic as has been demonstrated for other Thaumarchaeotes but, as for those strains, the genetic mechanism for organic carbon assimilation is not clear.

Two of the bacterial genomes harbour ribulose 1,5-bisphosphate carboxylase (RuBisCO) operons: both large subunits associate with the form I RuBisCOs, specifically the *Chloroflexi*CSP1-4 protein with the group IC RuBisCOs, whereas the NC10 CSP1-5 protein belongs to group ID (Fig. S7). Both operons share a *cbbL-cbbS-cbbX* canonical structure, with the *Chloroflexi* CSP1-4 genome having an additional upstream RuBisCO transcriptional regulator *cbbR*. The *Chloroflexi* CSP1-4 also encodes an aerobic carbon monoxide dehydrogenase complex, which may form CO_2 from CO to feed into the Calvin–Benson–Bassham (CBB) cycle for carbon fixation. In a previous profile of sediment *Chloroflexi* from the Rifle site, no RuBisCOs were identified aside from a small number of RuBisCO-like form IV genes (Hug *et al*., 2013); hence, this is the first report of a form I RuBisCO from a subsurface *Chloroflexi*. The type strain for the NC10 phylum, *Candidatus* M. oxyfera has been shown to encode and transcribe genes for the CBB cycle for carbon assimilation (Rasigraf *et al*., 2014), making this a conserved metabolic feature in the NC10 to date. Beyond the RuBisCOs, both genomes contain predicted genes for the complete CBB cycle, several of which are detected in the metaproteomic data (Table S4).

Finally, the *Armatimonadetes* CSP1-3 encodes a carbon monoxide dehydrogenase/acetyl-CoA synthetase complex that is predicted to generate acetyl-CoA from CO_2 as part of the Wood–Ljungdahl pathway. The CSP1-3 genome encodes a complete Wood–Ljungdahl pathway on a single scaffold whose final gene is a formate dehydrogenase alpha subunit (*fdhA*). The other *fdh* subunits are absent, but are expected to be present in the complete genome.

The remaining four organisms derive carbon for growth from organic sources. The *Candidatus*Dadabacteria CSP1-2's genome encodes glycolysis, a complete TCA cycle, and a canonical electron transport chain for oxidative phosphorylation, including a cytochrome c oxidase. The Rifle bulk groundwater contained 0.784 mg l^{-1} dissolved oxygen (Table S2), meaning aerobic metabolisms may be functional in microenvironments. The Dadabacteria genome also encodes a nitrite reductase forming ammonium (see below), making it likely that nitrate is the terminal electron acceptor for CSP1-2 in anaerobic conditions. The *Candidatus* Rokubacteria CSP1-6 is a predicted acetoclastic heterotroph likely also utilizing beta-oxidation of fatty acids for energy. The CSP1-6 genome encodes an AMP-dependent acetyl-CoA synthetase, predicted to convert acetate to acetyl-CoA, which then feeds into a complete TCA cycle and downstream electron transport chain. A second annotated acetyl-CoA synthetase is a mitochondrial-type medium chain acyl-CoA synthetase (MACS). MACS enzymes function in beta-oxidation in eukaryotes, forming the initial acyl-CoA substrate, but have been shown to activate shorter chain fatty acids like 2-methylbutyrate, acetate and butyrate in microorganisms (Meng *et al*., 2010). A complete beta-oxidation pathway was identified in the genome, including 8 acyl-CoA synthetases (long, medium and short-chain specific) as well as 16 annotated enoyl-CoA hydratases, making fatty acid degradation likely an important aspect of CSP1-6's metabolism. The genome additionally contains a complete pathway for butyrate metabolism, making butyrate a likely by-product released to the environment. The CSP1-6 genome is missing several key glycolysis genes, including phosphoglycerate kinase and pyruvate kinase, suggesting it cannot utilize sugars for energy. Like CSP1-2, the CSP1-6 genome encodes a single-chain pyruvate ferrodoxin oxidoreductase linked to electron transfer to nitrate during denitrification. Finally, the CSP1-6 genome contains an aerobic carbon monoxide dehydrogenase, *coxL-coxS* gene pair found in carboxydotrophs, meaning it may be able to oxidize CO to CO_2.

In keeping with the predictions of universally conserved fermentative and symbiotic lifestyles for members of the CPR (Wrighton *et al*., 2012; Kantor *et al*., 2013; Rinke *et al*., 2013; Brown *et al*., 2015), the WWE3 CSP1-7 genome lacks many core metabolic pathways, including a complete absence of a TCA cycle, electron transport chain, ATP synthase and most amino acid synthesis pathways, as well as an incomplete glycolysis pathway. The majority of the enzymes of the pentose phosphate pathway are present, possibly for generation of reducing equivalents. In the absence of any predicted enzymes for nucleotide or aromatic amino acids synthesis, it is unlikely the pentose phosphate pathway products act as precursor molecules as seen in other organisms. The CSP1-7 genome contains a nickel-dependent coenzyme-F420-reducing hydrogenase, but none of the genes involved in F420 cofactor synthesis, nor does it encode any canonical fermentative pathways. The means of energy generation for this organism are enigmatic, but likely involve fermentative processes encoded in the ∼ 40% of proteins with no predicted function or annotation on the genome. The presence of a WWE3 among the most abundant organisms, and the identification of several other members of the CPR within the community (Fig. 3) are evidence that native conditions in the aquifer support significant fermentative growth and underline the likely importance of this mode of life in subsurface environments.

The final heterotrophic organism is the *Deltaproteobacteria* CSP1-8, whose draft genome encodes near-complete glycolysis and TCA cycles, lacking pyruvate kinase and succinyl-CoA synthetase respectively. The genome also encodes a pyruvate dehydrogenase complex and complete canonical electron transfer chain (Table S4). Given that the genome encodes a respiratory nitrate reductase (NarG), this member of the *Deltaproteobacteria* may be capable of anaerobic respiration as well as aerobic growth. The missing enzymes may be present on the complete genome, or the organism may instead rely on a complete beta-oxidation pathway also present for energy generation. The genome does not contain hydrogenases, genes associated with acetogenesis, the final steps of butyrate fermentation or other hallmarks of a fermentative metabolism.

The production of CO_2 by heterotrophs and its utilization by autotrophs forms an interlinked subsurface microbial community network. New carbon resources are introduced to the community through the degradation of organic matter by heterotrophic and mixotrophic metabolisms and by autotrophic fixation of CO_2 derived from biotic and abiotic sources within the saturated sediment matrix.

Nitrogen

The CSP1 groundwater near the time of sediment sampling contained high nitrate (103 μM) and low ammonium (1.64 μM), a general trend that has been stable at this site for at least 2 years (Hug *et al*., 2015). Despite the conditions at the site, the most abundant organism in this data set, CSP1-1 *Thaumarchaeota*, is not a nitrate reducer, but rather a potential ammonia-oxidizing archaea (AOA). The CSP1-1 genome contains a complete ammonia monooxygenase (*amoABC*) operon, the machinery for AOA's high-affinity ammonia oxidation activity (Martens-Habbena *et al*., 2009; Walker *et al*., 2010). Although the concentration of ammonia in the groundwater at the site is low, ammonia sorbed to sediments as well as ammonia produced by the surrounding microbial community may lead to significantly higher local ammonia concentrations than are detectable in bulk groundwater measurements. Further, two of the other most abundant organisms encode dissimilarity nitrate reduction to ammonia (DNRA) pathways (the *Candidatus* Dadabacteria and *Deltaproteobacteria*, Figs 4 and 5) that may generate ammonia as a substrate for the CSP1-1 organism. The Amo enzyme complex's substrate threshold has been identified as ≤ 10 nM (Martens-Habbena *et al*., 2009), meaning even the low ammonia concentrations in groundwater are sufficient to support this metabolism. The CSP1-1 catalytic AmoA displays congruent phylogenetic placement compared with the 16S rRNA gene (Fig. S4). This genome places between the marine group I/*Nitrosopumilales* and the soil *Thaumarchaeota* clades. Although marine group I organisms are, to date, all confirmed ammonia oxidizers, the soil organisms' Amo can instead function in the biodegradation of organic material as a hydroxylase (Mussmann *et al*., 2011). It is unclear which function the CSP1-1 Amo is conducting in the subsurface. In addition to the Amo operon, the CSP1-1 genome encodes one *nirK* nitrite reductase, compared with two encoded by the type strain *Nitrosopumilus maritimus* strain SCM1 (Walker *et al*., 2010).

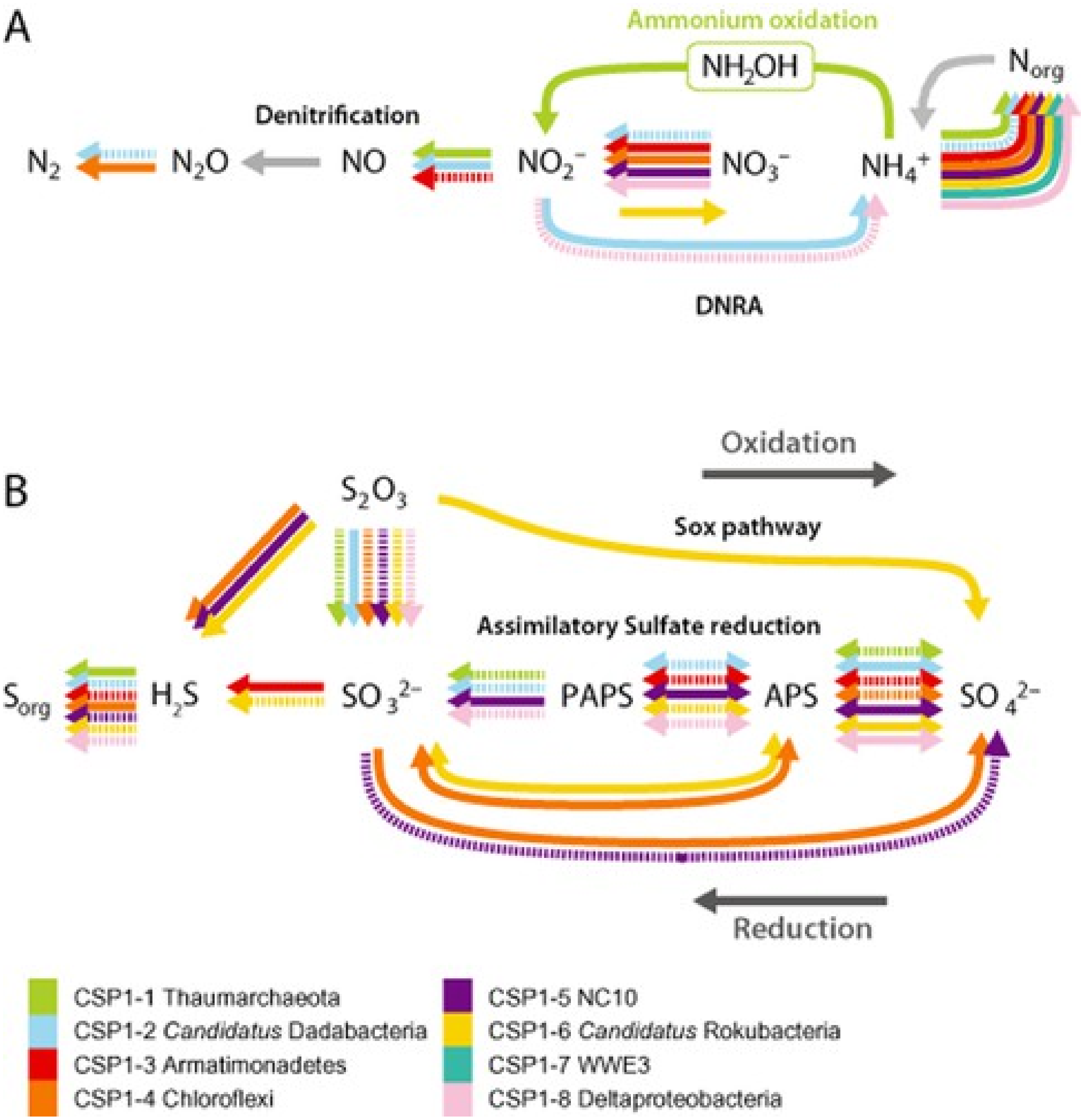

Figure 5

Open in figure viewerPowerPoint

Biogeochemical activities predicted for the eight profiled organisms based on genome content. Arrows are coloured by organism as in Fig. 3. Solid arrows indicate proteins and processes detected from groundwater metaproteomic efforts, dashed arrows correspond to predicted proteins not identified in the proteomic data.

A. Predicted nitrogen cycle activities. DNRA = dissimilatory nitrate reduction to ammonia.

B. Predicted sulfur cycle activities.

Caption

The abundant bacteria from the CSP1_5m_4 community are nearly all predicted to function in nitrogen cycling, with specific compounds apparently acting as metabolic hand-offs between community members (these organisms or others within the broader community). Only the WWE3 CSP1-7 is not predicted to use nitrogen compounds, a finding in-keeping with the expected reduced metabolic lifestyle for organisms within the CPR (Wrighton *et al*., 2012; Kantor *et al*., 2013; Brown *et al*., 2015). Nitrate reduction is a common feature of the abundant organisms from this site: five of the seven bacteria encode nitrate reductases (Fig. 5). CSP1-2, the *Ca*. Dadabacteria, is predicted to convert nitrate to nitrite, and then nitrite either to ammonia, based on the presence of *nirBD*, or to nitric oxide (NO), based on an encoded *nirK*. The *Deltaproteobacteria* CSP1-8 shares the predicted DNRA function, whereas the *Armatimonadetes* CSP1-3 shares the *nirK* predicted nitrite reduction to NO. Although none of the eight organisms profiled here encode a nitric oxide reductase, there are at least three *norBC*(*DEQ*) operons within the assembled but unbinned CSP1_5m_4 metagenome, suggesting this missing link for denitrification is present within the community. Completing denitrification, nitrous oxide is predicted to be converted to nitrogen gas by both the CSP1-1 *Ca.*Dadabacteria, and the CSP1-4 *Chloroflexi*, whose genomes encode *nosZ*. The CSP1-6 genome, the *Ca*. Rokubacteria, is unique in encoding multiple nitrite oxidoreductases, predicted to convert nitrite to nitrate, with two *nxrAB* operons present on the genome. Each of the seven bacterial genomes encodes a unique suite of genes relevant to the nitrogen cycle, forming an interconnected network of possible activities (Figs 4 and 5).

Notably, the CP NC10 CSP1-5 does not appear to share the nitrite-dependent anaerobic methane oxidation activities of the type strain for the phylum. The CSP1-5 genome does not contain any genes associated with methane oxidation (*pmoABC*) or the *nirSJFD* operon for conversion of 2 NO_2 to 2 NO. Additionally absent is a quinol-dependent NO reductase. The genome does contain a *nar* operon as well as a blue copper *nirK* nitrite reductase, suggesting it is actively involved in nitrogen cycling, although not capable of conducting methane oxidation. The genome also encodes the machinery for methylotrophy, including a methanol dehydrogenase, formate dehydrogenase, a bifunctional 5,10-methylene-tetrahydrofolate dehydrogenase/5,10-methylene-tetrahydrofolate cyclohydrolase and a methylene tetrahydromethanopterin dehydrogenase (mtdB). The phylogenetic distance between the CSP1-5 NC10 and the type strain *Candidatus* M. oxyfera clarifies the boundaries of denitrification coupled to methane oxidation as a trait within the NC10 phylum (Fig. S6).

Sulfur

The ambient sulfate concentration at the site was 8.5 mM (816 mg l^{-1}) at the time of sampling, which is classified as high concentration by the United States Geological Survey (USGS) (> 30 mg l^{-1}, Sacks, 1996), and is significantly above the U.S. Environmental Protection Agency (EPA) maximum contaminant level for drinking water of 250 mg l^{-1}. Given the abundant sulfate at the site, it is not surprising that the majority of the abundant organisms are predicted to reduce sulfate (Fig. 5). The network of sulfur compound conversions predicted from the profiled genomes includes complete assimilatory sulfate reduction, sulfur oxidation and thiosulfate reduction and assimilation. Of the eight organisms, only the WWE-3 CSP1-7 is not predicted to be involved in sulfur conversions.

Six of the seven organisms predicted to assimilate sulfur through sulfate reduction encode partial or fragmented pathways, suggesting that complete sulfate reduction occurs through the functions of multiple organisms, with compounds handed off at different points in the pathway. The capacity to incorporate thiosulfate is typically encoded on these genomes, with its reduction to sulfite predicted for six of the seven organisms, and its direct conversion to hydrogen sulfide additionally encoded by three genomes (Fig. 5). The CSP1-1 *Thaumarchaeota*encodes a disjointed sulfate assimilation pathway, with sulfate adenylate transferase (*sat*), phosphoadenosine phosphosulfate reductase (PAPS reductase, *cysH*) and thiosulfate sulfurtransferase present on the genome. Sulfite is predicted to be formed from both PAPS and thiosulfate, but no mechanisms for conversion of adenosine phosphosulfate to PAPS, or for sulfite to hydrogen sulfide were identified on the genome. The CSP1-2 *Ca.* Dadabacteria encodes sulfate reduction through to sulfite, as well as thiosulfate reduction to sulfite, but no further steps for sulfite reduction. The genome does contain a *cysK* for sulfide assimilation to organic sulfur, but no mechanism for conversion of sulfite to sulfide, suggesting that reaction may instead be occurring in the flanking microbial community. The CSP1-3 *Armatimonadetes*has a near-complete assimilatory sulfate reduction pathway, lacking only a PAPS reductase or other mechanism to form sulfite. Unlike the other six organisms, CSP1-3 does not encode any means to recruit thiosulfate into its sulfate assimilation pathway. The CSP1-4 *Chloroflexi* and CSP-5 NC10 genomes encode identical functions within the sulfur cycle: near-complete sulfate reduction pathways lacking only sulfite reductase. Both encode sulfite oxidases, suggesting the main direction of sulfite conversion in these organisms is oxidation to sulfate, coupled to electrons passed to the complete electron transport chains for downstream ATP synthesis. The CSP1-4 and CSP1-5 genomes are predicted to be able to convert thiosulfate to hydrogen sulfide and subsequently to organic sulfur. The CSP-8 *Deltaproteobacteria* rounds out the six organisms with partial sulfate reduction pathways. The genome lacks sulfite reductase as well as the ability to convert thiosulfate to hydrogen sulfide – another example where a different member of the microbial community may be responsible for conversion of sulfite to hydrogen sulfide.

The CSP1-6 *Ca.* Rokubacteria contains a remarkable suite of sulfur genes, including the only complete assimilatory sulfate reduction pathway found in the eight genomes surveyed. The genome additionally encodes the genes for conversion of thiosulfate to both sulfite and hydrogen sulfide, for eventual sulfur assimilation. Remarkably, and uniquely within the abundant organisms, the *Ca.* Rokubacteria also encodes a *sox* operon for oxidation of thiosulfate to sulfate. This reaction is likely energetically unfavourable given the local concentrations of sulfate, but the pathway contributes to this organism's potential repertoire of sulfur conversions. In summary, *Ca.* Rokubacteria has the ability to oxidize thiosulfate to sulfate and the complete sulfate reduction pathway, as well as two additional mechanisms to incorporate thiosulfate into organic matter. As this is the first genome sequence for this lineage, it will remain to be seen whether these functions are typically associated with this candidate phylum. If so, the *Ca.* Rokubacteria may represent important and currently unrecognized players within the sulfur cycle.

Metaproteomics confirms expression of key genes in the subsurface

From a concurrent experiment conducted at the Rifle site, we had access to metaproteomic data from groundwater sampled in June and July of 2013 from a site adjacent to the CSP1 sediment core. The eight organisms discussed here were present in the groundwater at this time, although at markedly lower abundances (average coverage of 2.35× ± 3.36 in groundwater sequence compared with average coverage of 52.5× ± 28.53 in the sediment metagenome) (Hug *et al.*, 2015). Many key proteins for the metabolic pathways described above were identified from 2 dimensional liquid chromatography tandem mass spectrometry (2D-LC-MS/MS) peptide spectra, although the low abundance of the profiled organisms precluded sufficient spectral counts for statistical analyses of expression differences across filter sizes or environmental conditions. Detection of these proteins, including both RuBisCOs, numerous nitrate reductase subunits, the CSP1-1 AmoABC and components of the Sox pathway from the CSP1-6 *Ca.* Rokubacteria (Fig. 5), provides evidence that these functions are active within the sediment-associated community and may be contributing to geochemical cycling as predicted. The full list of described genes from the eight profiled organisms is presented in Table S4, alongside the peptide hits for each protein of interest.

Summary

The microbial community-colonizing aquifer sediment was identified and genomically characterized using metagenomic sequencing. Assembly of the metagenome was conducted with subsampling using 5%, 10%, 20%, 33% and 50% of the reads. In combination with a full 100% assembly, the subsampling approach leads to a more accurate microbial community representation and improved recovery of abundant organisms' genomes. The eight most abundant organisms in the community were selected for genome binning and reconstruction, leading to curated draft genomes for organisms representing novel lineages within the *Thaumarchaeota*, *Armatimonadetes*, *Chloroflexi*, *Deltaproteobacteria* and the CP NC10 and WWE3. Two of the curated genomes were from organisms belonging to previously unknown phyla for which we propose the names *Candidatus* Dadabacteria and *Candidatus* Rokubacteria. Metabolic predictions from the genome sequences implicate the eight profiled organisms in the carbon, nitrogen and sulfur cycles. Key predicted activities include carbon fixation through the CBB cycle for the *Chloroflexi* and NC10 organisms, ammonia oxidation by the *Thaumarchaeota* with ammonia partially supplied through DNRA by the *Ca.* Dadabacteria and *Deltaproteobacteria*, and sulfur oxidation by the *Ca.* Rokubacteria. Many of the proteins involved in the predicted activities were also identified in metaproteomic data, confirming expression of these key genes in the subsurface environment. The results indicate biogeochemical cycling in the subsurface may involve a complex network of reactions and metabolic hand-off points, orchestrated by a wide diversity of organisms, including relevant lineages not previously implicated in these processes.

Experimental procedures

Sequence origin

A sediment core was recovered during installation of groundwater monitoring well FP-101 at the U.S. Department of Energy (DOE) Rifle research site in Rifle, Colorado (USA) on July 20th, 2011, from a 6–7 m thick aquifer adjacent to the Colorado River (Latitude 39.52927920, Longitude −107.77162320, altitude 1618.31 m above sea level; Williams *et al.*, 2011). Sediment samples from 3, 4, 5 and 6 m depths were stored within gas-impermeable sample bags at −80°C, and kept frozen during transport and prior to DNA extraction. For the 5 m sample, the frozen sediment was split into four subsamples prior to DNA extraction. Each sample comprised ∼ 100 g of sediment (Table S2), from which 10 independent DNA extractions of ∼ 10 g of sediment were conducted using the PowerMax® Soil DNA Isolation Kits (MoBio Laboratories, Carlsbad, CA, USA) with the following modifications to the manufacturer's instructions. Sediment was vortexed at maximum speed for an additional 3 min in the sodium dodecyl sulfate (SDS) reagent, and then incubated for 30 s at 60°C in place of extended bead beating. DNA extractions were concentrated using a sodium acetate/ethanol/glycogen precipitation and replicate extractions pooled to generate sufficient DNA for sequencing.

Sequencing and assembly

Illumina HiSeq paired-end sequencing was conducted by the DOE Joint Genome Institute, with one sample of DNA per lane for a total of seven lanes of sequencing: one from each of the 3, 4 and 6 m depths, and four from the 5 m depth. Reads were 150 bp long, with an insert size of ∼ 500 bp. All reads were pre-processed using SICKLE (https://github.com/najoshi/sickle) with default settings. Only paired-end reads were used in the assemblies.

The 5m_4 sample reads were assembled using IDBA_UD (Peng *et al*., 2012) under default parameters. In parallel, the sequence data were iteratively subsampled and assembled. Initially, the data set was sampled with replacement of reads, with assemblies of 1%, 1.5%, 2%, 5%, 10% and 20% of reads conducted. Following this, the data set was sampled without replacement, with assemblies of approximately 5%, 10%, 20%, 33.3% and 50% of the original reads, in that order (see Table S5 for details). For subsampling without replacement, after each subsample, all 5m_4 reads were mapped to scaffolds using BOWTIE2 (Langmead *et al*., 2009), and reads mapping to scaffolds with ≥ 10× coverage were excluded from the subsequent subsamples. To generate the final data set, the full and the without-replacement subsampled assemblies were combined using MINIMUS2 (Sommer *et al*., 2007) to condense replicate regions. Unless otherwise stated, the subsampled assemblies discussed refer to those conducted without replacement.

Binning and genome reconstruction

The most abundant organisms were binned into putative genome data sets using ESOMs (Dick *et al*., 2009). Raw reads from 13 existing metagenomic samples were mapped against the final assembly, including the seven samples from this study, the 4, 5 and 6 m samples from an earlier sediment sample (Castelle *et al*., 2013; Hug *et al*., 2013; Sharon *et al*., 2015), one sediment column metagenome (AAC1, Kantor *et al*., 2013) and two planktonic community groundwater filtrate metagenomes from 0.2 μm and 0.1 μm sequential filters (GW2011 A1 and A2; Hug *et al*., 2015; Luef *et al*., 2015). The resultant coverage profile across data sets for each scaffold was combined with mono-, di- and tri-nucleotide frequencies as inputs to the ESOM, a hybrid time series abundance and nucleotide composition methodology (Sharon *et al*., 2013). The ESOM was generated using only those scaffolds with a coverage of 20× or greater in the merged metagenome assembly to target the more abundant organisms.

Bins for the eight most abundant organisms were determined based on ESOM clusters (Fig. S1), phylogenetic affiliation of the genes and scaffolds, and consistent %GC content and coverage. Reads from the complete 5m_4 sequence data set were mapped to each of the eight genome bins, and the matching reads extracted for single-genome re-assemblies using Velvet (Zerbino and Birney, 2008). Each genome was automatically curated by identifying reads mapped to the ends of scaffolds, followed by scaffold extension and connection through site-specific reassembly. Further manual curation using read mapping to extend scaffold ends and to close internal gaps enabled significant improvements in the genome properties (Table S6). A final curation step involved identification of areas of zero read coverage from stringent read mapping, followed by automated re-assembly to correct SNPs and small in/dels in the assemblies. Following curation, the draft genomes were re-annotated with the same workflow as for the full data set.

Gene calling and annotation

A functional prediction was conducted on all identified open reading frames as described previously, with similarity searches conducted using UCLUST (Edgar, 2010), and omitting the INTERPROSCAN analysis (Hug *et al*., 2013). The annotation included the taxonomic affiliation of the best-match protein, generating a phylogenetic fingerprint for each scaffold. More detailed analyses were conducted for proteins of interest. Metabolic profiles were re-constructed from gene annotations through a combination of the KEGG Automated Annotation System (Kanehisa *et al*., 2012, http://www.genome.jp/tools/kaas) and ggKbase list functionality (http://ggkbase.berkeley.edu). For proteins of interest, BLASTP and associated domain hidden Markov models (HMMs) (Altschul *et al*., 1990), protein alignments, tree analyses and SwissProt modelling of predicted enzyme structures (swissmodel.expasy.org) were additionally conducted. Absent genes may be a result of missing sequence information in the curated genome, or a true absence of this encoded function on the genome; it is not possible to distinguish between these options with metagenomic methods.

Phylogenetic profiling

All trees were generated with reference sets from NCBI, Integrated Microbial Genomes (IMG) and SILVA databases, with nearest-neighbour reference sequences identified and included. The taxonomic placements of the eight organisms with curated genomes were determined using phylogenies for both the 16S rRNA gene and a syntenic block of ribosomal proteins (Hug *et al*., 2013). For 16S rRNA trees, alignments were generated using the SILVA SINA alignment algorithm (Pruesse *et al*., 2012) with positions containing > 97% gaps stripped from the alignment. All protein alignments were generated using MUSCLE (Edgar, 2004), and manually curated to remove single-taxon insertions and to unif y within-gene lengths. Phylogenetic trees were conducted using RAXML-HPC version 7.2.8 (Stamatakis, 2006) under the GTR-GAMMA model for nucleotide alignments and the PROTGAMMALG model for protein alignments. All trees were conducted with 100 bootstrap re-samplings. Where condensed trees are displayed for visual clarity the complete trees in newick format are included in Text S1.

Metaproteomics

Total proteins were determined from groundwater samples pumped from a monitoring well in June and July 2013. For each sampling, approximately 36 000 l of groundwater was pumped from well FP-101 through serially connected polyethersulfone membrane filter cartridges with pore sizes of 1.2, 0.2 and 0.1 μm (Graver Technologies, Glasgow, DE, USA). The 1.2, 0.2 and 0.1 μm filters were back-flushed with ∼ 3000 ml of distilled, de-ionized water containing 0.5% Tween 80, 0.01% sodium pyrophosphate and 0.001% Antifoam Y-30 emulsion (all reagents, Sigma-Aldrich, St. Louis, MO, USA), the back-flushed solution centrifuged at 10 000 r.p.m., supernatant removed and the biomass kept frozen at −80°C for transport and storage.

Protein extractions were conducted on the six groundwater samples using one third of the collected biomass. Protein extraction, trypsin digestion and analysis were conducted as described previously (Glass *et al*., 2014). In brief, proteins were extracted through lysis and sonication, purified through precipitation and then digested overnight with trypsin. Duplicate samples per filter were desalted and analysed on an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific). The raw MS/MS data were searched using MyriMatch V2.1 (Tabb *et al*., 2007) against the predicted protein database from the eight curated genomes. Resulting peptide hits were filtered and assembled using IDPicker V. 3.0.564 (Ma *et al*., 2009). Proteins with at least one unique and one non-unique peptide match were reported. Resulting peptides were filtered and assembled using IDPicker keeping False discovery rate (FDR) below 2% at peptide level. False discovery rates at the protein level varied between 0% and 2.7%.

Sequence availability

Genome sequences, predicted genes and annotations for the eight organisms are available through http://genegrabber.berkeley.edu/CSP-1_EM_2015/organisms. The eight genomes were deposited in DDBJ/EMBL/GenBank under Bioproject PRJNA262935, under the Whole Genome Shotgun accessions LDXK00000000-LDXR00000000. The versions described in this paper are versions LDX[K-R]01000000. The ribosomal protein scaffolds used for taxonomic identification of the total microbial community have been deposited in DDBJ/EMBL/GenBank under Biosample SAMN03092877, accession numbers KT006933 - KT007086.

Metagenomic sequence read data sets used for read mapping and binning are available through the JGI portal system (http://genome.jgi.doe.gov) under the following Project IDs: this study, 3m depth: 1008424, 4 m depth: 1008430, 5m_1: 1008442, 5m_2: 1008445, 5m_3: 1008448, 5m_4: 1008451, and 6 m depth: 1008457. Previous studies: RBG_4m: 1016235, RBG_5m: 1016238, RBG_6m: 1016241, GW_2011_A1: 1006501 and GW_2011_A2: 1006504. The AAC1 sediment column sample's reads are deposited in the NCBI SRA database under accession SRX329136.

Acknowledgements

The authors acknowledge the DOE Joint Genome Institute sequencing facility for generating the metagenome sequence via the Community Science Program. This material is based upon work supported as part of the Lawrence Berkeley National Laboratory's Sustainable Systems Scientific Focus Area funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under contract DE-AC02-05CH11231. LAH was partially supported by an NSERC Post-Doctoral Fellowship. The authors state they have no conflicts of interest to declare.