The ISME Journal (2010) 4, 660–672

www.nature.com/ismej

ORIGINAL ARTICLE

Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community

Christopher L Hemme[1,2], Ye Deng[1], Terry J Gentry[3,2], Matthew W Fields[4], Liyou Wu[1,2], Soumitra Barua[1,2], Kerrie Barry[5], Susannah G Tringe[5], David B Watson[2], Zhili He[1], Terry C Hazen[6], James M Tiedje[7], Edward M Rubin[5] and Jizhong Zhou[1,2]

[1]*Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, OK, USA;* [2]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA;* [3]*Department of Soil Sciences, Texas A&M University, College Station, TX, USA;* [4]*Department of Microbiology, Montana State University, Bozeman, MT, USA;* [5]*US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA;* [6]*Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and* [7]*Center for Microbial Ecology, Department of Soil and Crop Sciences, Michigan State University, East Lansing, MI, USA*

Understanding adaptation of biological communities to environmental change is a central issue in ecology and evolution. Metagenomic analysis of a stressed groundwater microbial community reveals that prolonged exposure to high concentrations of heavy metals, nitric acid and organic solvents (~50 years) has resulted in a massive decrease in species and allelic diversity as well as a significant loss of metabolic diversity. Although the surviving microbial community possesses all metabolic pathways necessary for survival and growth in such an extreme environment, its structure is very simple, primarily composed of clonal denitrifying γ- and β-proteobacterial populations. The resulting community is overabundant in key genes conferring resistance to specific stresses including nitrate, heavy metals and acetone. Evolutionary analysis indicates that lateral gene transfer could have a key function in rapid response and adaptation to environmental contamination. The results presented in this study have important implications in understanding, assessing and predicting the impacts of human-induced activities on microbial communities ranging from human health to agriculture to environmental management, and their responses to environmental changes.
The ISME Journal (2010) **4,** 660–672; doi:10.1038/ismej.2009.154; published online 25 February 2010
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: metagenomics; microbial ecology; bioremediation

Introduction

Microorganisms are the most abundant and diverse group of life on the planet and have an integral function in biogeochemical cycling of compounds crucial to ecosystem functioning (Whitman *et al.*, 1998). Comprehensive characterization of microbial communities in natural systems remains a challenge due to their extremely high diversity and the as-yet uncultivated status of the vast majority of environmental microorganisms. Metagenomics and associated technologies has revolutionized the study of microbial diversity, adaptation and evolution (Riesenfeld *et al.*, 2004; Handelsman *et al.*, 2007; He *et al.*, 2007). Studies of microbial communities from several environments, including acid-mine drainage (Tyson *et al.*, 2004), marine water and sediments (DeLong *et al.*, 2006; Yooseph *et al.*, 2007), human gut (Gill *et al.*, 2006; Turnbaugh *et al.*, 2007) and soils (Smets and Barkay, 2005; Voget *et al.*, 2006), have yielded novel insights on gene discovery, metabolism, community structure, function and evolution. Metagenomic analysis offers an unprecedented opportunity to comprehensively examine ecosystem response to environmental change, but integrated surveys of microbial communities to date have not been reported that examine the responses and adaptation of microbial communities to environmental contaminants.

Although high-throughput sequencing of microbial communities is now possible, the complexity and magnitude of most communities complicate data interpretation. Low-complexity microbial communities from extreme environments, such as acidic geothermal hot springs and contaminated sites, are ideal for high-resolution, in-depth metagenomics

Correspondence: J Zhou, Institute for Environmental Genomics, University of Oklahoma, 101 David L Boren Blvd., Norman, OK 73019, USA.
E-mail: jzhou@rccc.ou.edu
Received 17 September 2009; revised 14 December 2009; accepted 14 December 2009; published online 25 February 2010

studies (Allen and Banfield, 2005). In this study, a microbial community from highly uranium-contaminated groundwater was sequenced using a random shotgun sequencing-based strategy with the aim of addressing the following questions: (1) How does anthropogenic environmental change such as contamination affect groundwater microbial community diversity and structure? (ii) How does a microbial community adapt to severe environmental changes such as heavy metal contamination? (3) Can molecular mechanisms responsible for such environmental changes be predicted from the metagenomic sequence? Results reveal novel insights into microbial community diversity, structure and function in a contaminated ecosystem and predicted processes by which microbial communities adapt to extreme levels of contamination.

Materials and methods

To obtain sufficient biomass for sequencing, we extensively purged the FW106 well (several well volumes of water removed), and pumped ~1700 l of groundwater from the matrix surrounding the screened area (~10 m depth) using peristaltic pumps and passed it through sintered metal (TJ Phelps, unpublished data) (589 l) or 0.2 μm Supor (Pall Corporation, Port Washington, NY, USA) filters (1126 l) to collect the biomass. High molecular weight community DNA was extracted using grinding, freezing-thawing SDS-based methods (Zhou *et al.*, 1996) and the purified DNA was treated with RNase (Zhou *et al.*, 1996). The metagenome was sequenced by JGI using random shotgun methods (Tyson *et al.*, 2004). Double-ended sequencing reactions were performed using PE BigDye terminator chemistry (PE Applied Biosystems, Carlsbad, CA, USA) and resolved using PRISM 3730 capillary DNA sequencer (PE Applied Biosystems). Approximately 53 Mb of high-quality Q20 read sequences were obtained from ~78 Mb raw sequence (three clone libraries: 20.04 Mb small insert (3 kb) pUC library, 23.13 Mb medium insert (8 kb) pMCL, 9.27 Mb large insert (40 kb) pCCiFos). The sequencing reads (66220) were assembled into 421 contigs w/>20 reads (2770 contigs total, ~8.3 Mb assembled DNA) using Phrap (Seattle, WA USA) as previously described (Tyson *et al.*, 2004), and the contigs were further assembled by paired-end analysis into 224 scaffolds ranging in size from 1.8 kb to 2.4 Mb. To account for polymorphisms expected to occur in community DNA, we allowed alignment discrepancies beyond those expected for random sequencing errors if they were consistent with end-pairing constraints. A second assembly of the FW106 metagenome was conducted using Lucy (vector and quality trimming) (Chou and Holmes, 2001) and the Paracel Genome Assembler (pga) (Paracel, Pasadena, CA, USA). Two independent annotations were performed on the Phrap assembly using the JGI-ORNL single genome and JGI-Integrated Microbial Resource (IMG) annotation pipelines, and the pga assembly was annotated using the IMG pipeline. The pga assembly and associated IMG annotation are available at the IMG/m database (Markowitz *et al.*, 2008) and the FW106 read library has been deposited in GenBank (accession number ADIG00000000). Assembled scaffolds resulting from the pga assembly were assigned to phylogenetic bins using PhyloPythia (McHardy *et al.*, 2007) and predicted genes were phylogenetically assigned by BlastP (Altschul *et al.*, 1997) homology using the online IMG Phylogenetic Profiling and Phylogenetic Distribution tools. Gene prediction, functional assignment and metabolic reconstruction were performed automatically using internal JGI protocols. Single-nucleotide polymorphisms (SNPs) were detected from the Phrap assembled metagenomic reads using *ad hoc* computational methods using BioPerl (Stajich *et al.*, 2002). A nucleotide change was classified as an SNP in a manner similar to that described by Wu *et al.* (2006): (1) sequence quality score was >20 in both the contig and read sequences, (2) there was >4-fold coverage in the affected assembly column, (3) there were differences among the other reads at that position and (4) the change was flanked by three invariable nucleotides on each side (Wu *et al.*, 2006). Oligonucleotide primers for cloning experiments were designed based on the assembled FW106 metagenomic sequence, and standard population genetics parameters were determined using DnaSP (Barcelona, Spain) (Rozas and Rozas, 1999). Laterally transferred genes were detected using a combination of composition-based and phylogenetic methods as described in the main text and Supplementary Materials and Methods. Pairwise analyses of major scaffold genes were conducted using PAML (London, UK) (Yang, 1997). Phylogenetic analyses were conducted using MEGA 4.0 (Tempe, AR, USA) (Tamura *et al.*, 2007) for functional gene analysis and with ARB (Munich, Germany; Ludwig *et al.*, 2004) for 16S analysis. Metagenomic sequences are deposited in the JGI-IMG database (Markowitz *et al.*, 2008). Details for all methods are provided in Supplementary Materials and Methods.

Results

Overview of the metagenomic sequencing

Phylogenetic diversity of the sampling site. Groundwater from well FW106 at Oak Ridge Environmental Remediation Sciences Program Field Research Center (FRC) is highly acidic (pH 3.7), and contaminated with extremely high levels of uranium (among the highest in the United States), nitrate, technetium and various organic contaminants (Supplementary Table S1). Microscopic analysis indicated that the microbial community was dominated by organisms representing only 2–3 different cell morphologies

(Supplementary Figure S1). Similarly, SSU rRNA gene-based phylogenetic analysis reveals very low phylogenetic diversity with a total of 13 operational taxonomic units (OTUs) from 619 sequences at the 98% sequence identity cutoff, with ~87% of these sequences corresponding to the BFXI557 γ-proteobacterial clone (Supplementary Figure S2). The community is composed primarily of γ- and β-Proteobacteria and dominated by *Rhodanobacter*-like γ-proteobacterial and *Burkholderia*-like β-proteobacterial species (Supplementary Figure S2).

Metagenomic sequencing. A total of ~70 Mb sequence was obtained from three small, medium and large insert clone libraries and were assembled using Phrap (~8.4 Mb, 2770 contigs) and pga (~9.5 Mb, 6079 contigs) (Supplementary Table S2). Contigs from the pga assembly were assigned to taxonomic bins using PhyloPythia (McHardy *et al.*, 2007) (Table 1; Figure 1a). The most populated bin corresponds to the dominant γ-proteobacterial group identified from the OTU analysis and this bin is designated FW106 γI. Protein recruitment plots show the most similarity to Burkholderiaceae and Xanthomonadaceae lineages (Supplementary Figure S3); however, the lack of closely related reference genomes complicates phylogenetic assignment of this metagenome. Although a complete FW106 γI genome could not be assembled, the relatively high degree of coverage permits extensive assembly of consensus contigs and scaffolds for this phylotype, with the largest scaffold ~2.4 Mb. Comparison of the two assemblies and multiple PCR experiments using primers designed from the assembled sequences suggest that the assemblies are accurate (results not shown). A total of 12 335 putative protein-coding genes were identified from the IMG annotation of the pga assembly and functional assignments were made for ~70% of the predicted genes, with ~64% assigned to COG categories and ~12% assigned to KEGG pathways (Supplementary Table S2; Figure 1a). A total of 3646 (~29%) of the predicted genes had no assigned functions. Protein-coding genes were assigned to phylogenetic taxa by BlastP (NCBI, Bethesda, MD, USA; Altschul *et al.*, 1997) homology using the IMG phylogenetic profiling tool (Figure 1b). Although 16S rRNA analysis and field experiments show dominance of the community by γ-proteobacterial species, β-Proteobacteria constituted the largest reservoir of assigned functional genes (18%) followed by γ- (12%) and α-Proteobacteria (3%) (Figure 1b). This is primarily due to the large degree of assembly of the γ-proteobacterial contigs compared to assembly of β-proteobacterial contigs. The dominant lineages in FW106 based on protein assignment are Burkholderiaceae, Xanthomonadaceae and Comamonadaceae, consistent with previous analyses (Figure 1b).

Abundance of geochemical resistance genes. Abundance profiles of FW106 genes assigned to COG functional categories compared to all sequenced bacteria show an overabundance of genes involved in DNA recombination and repair, defense mechanisms, cell motility, intracellular trafficking, energy production and conversion, lipid metabolism and transport, and secondary metabolite biosynthesis and transport (Figure 2). Overabundance of defense and repair mechanisms for dealing with stress-induced damage and contaminant-specific mechanisms for dealing with heavy metals, low pH, nitrate/nitrite and organic solvents are expected to occur in the acidic heavy metal-contaminated environment of FW106. A more detailed analysis of the abundance of COG functional groups shows a strong overabundance of resistance genes likely propelled by specific contaminants such as nitrate and heavy

Table 1 Binning of metagenomic contigs by PhyloPythia

Domain	*Phylum*	*Class*	*No. of contigs*	*Total contigs (%)*	*Sequence (bp)*	*Total sequence (%)*
Archaea	Crenarchaeota	Thermoprotei	1	0.02	721	0.01
Archaea	Crenarchaeota	Unassigned	1	0.02	845	0.01
Archaea	Euryarchaeota	Unassigned	16	0.26	13 720	0.14
Archaea	Unassigned	Unassigned	34	0.56	27 790	0.29
Bacteria	Actinobacteria	Actinobacteria	9	0.15	19 005	0.20
Bacteria	Actinobacteria	Unassigned	52	0.86	61 095	0.64
Bacteria	Bacteroidetes	Bacteroidetes	1	0.02	949	0.01
Bacteria	Deinococcus-Thermus	Unassigned	2	0.03	1663	0.02
Bacteria	Firmicutes	Bacilli	3	0.05	11 581	0.12
Bacteria	Firmicutes	Unassigned	9	0.15	8441	0.09
Bacteria	Proteobacteria	α-Proteobacteria	9	0.15	15 552	0.16
Bacteria	Proteobacteria	β-Proteobacteria	1659	27.29	2 490 010	26.06
Bacteria	Proteobacteria	ε-Proteobacteria	2	0.03	1669	0.02
Bacteria	Proteobacteria	γ-Proteobacteria	84	1.38	3 629 419	38.00
Bacteria	Proteobacteria	Unassigned	450	7.40	552 980	5.79
Bacteria	Unassigned	Unassigned	3471	57.10	2 636 705	27.60
Eukaryota	Arthropoda	Unassigned	4	0.07	3518	0.04
Eukaryota	Chordata	Unassigned	1	0.02	714	0.01
Unassigned	Unassigned	Unassigned	271	4.46	78 167	0.82

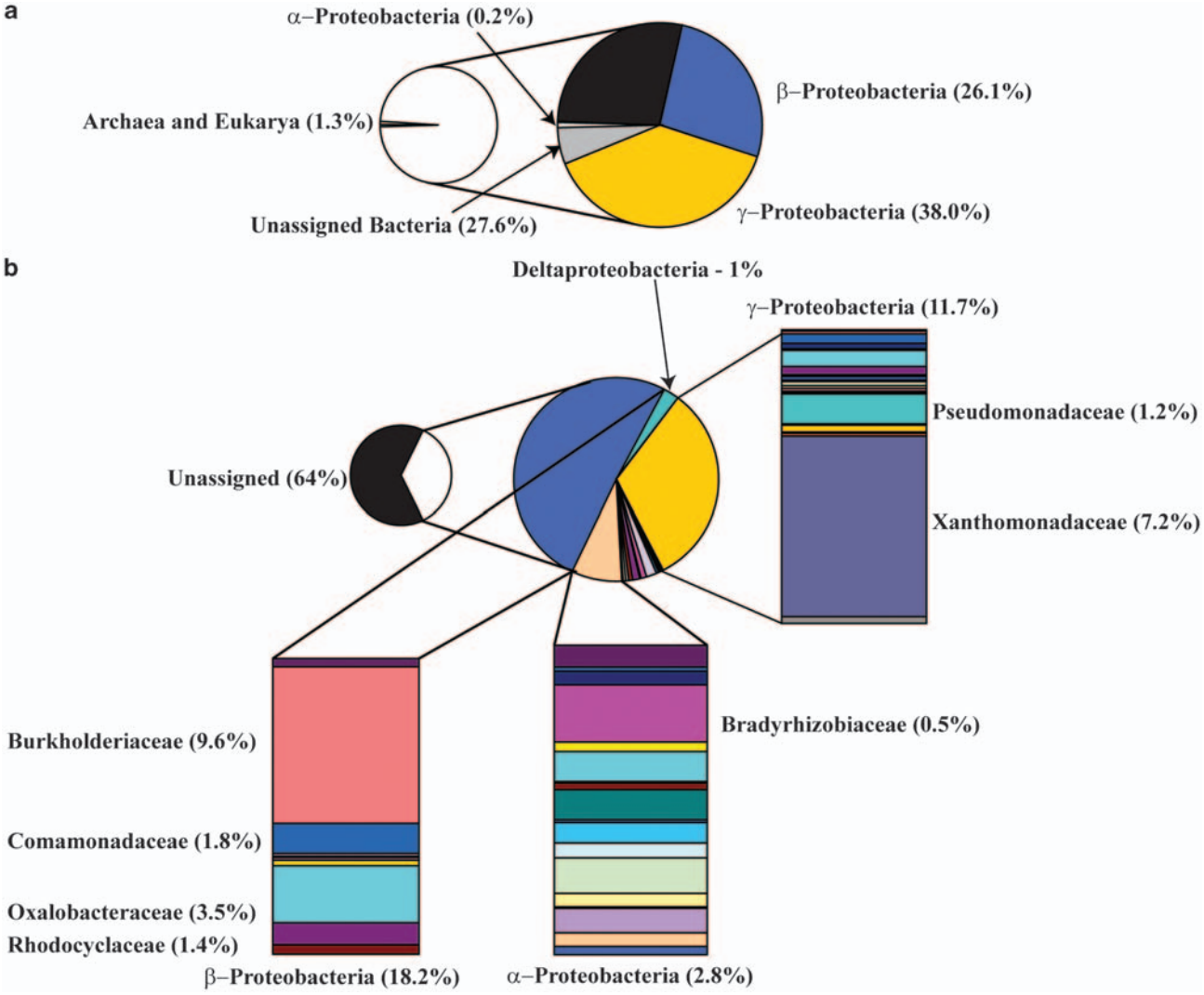

Figure 1 Phylogenetic profiling of FW106 metagenome. (**a**) Binning of FW106 contigs by PhyloPythia (see also Table 1). (**b**) Binning of FW106 genes based on IMG phylogenetic profiling tools. Percentage values represent the number of genes assigned to a particular taxon compared to all genes in the metagenome.

metals. These resistance genes include toxin transport genes such as NarK nitrate/nitrite antiporters and $Cd^{2+}/Zn^{2+}/Co^{2+}$ efflux components (CzcABC, CzcD) (Figure 3). Accumulation of genes involved in resistance and stress response mechanisms thus appears to be a basic survival strategy used by the community in response to the specific contaminants in FW106.

Metabolic reconstruction of FW106 γI

To better understand how the FW106 microbial community responds to stress on a genomic scale and to gain a comprehensive view of the metabolic capabilities of the community, we performed metabolic reconstruction for the dominant FW106 γI phylotype. Sequence coverage of the metagenome was sufficient to produce a comprehensive metabolic reconstruction of the consensus FW106 γI species (Figure 4) and a partial reconstruction of FW106 βI (data not shown). Although these reconstructions are incomplete and likely represent composite cell networks, the information obtained is sufficient to address specific questions regarding the metabolic potential of the community and to correlate this data to the FW106 contamination profile (Supplementary Table S1).

Reconstruction of central carbon pathways and identification of carbon transport systems suggest the community subsists primarily on simple mono- and disaccharides, including cellulosic degradation products (for example, cellobiose) that may permeate into the groundwater from adjacent soil. Limited metabolism of complex carbohydrates by FW106 γI is implied by the presence of genes encoding an exoxylanase and xylose interconversion enzymes (Figure 4). Complete glycolytic, TCA, pentose phosphate, Entner-Doudroff and methylglyoxal

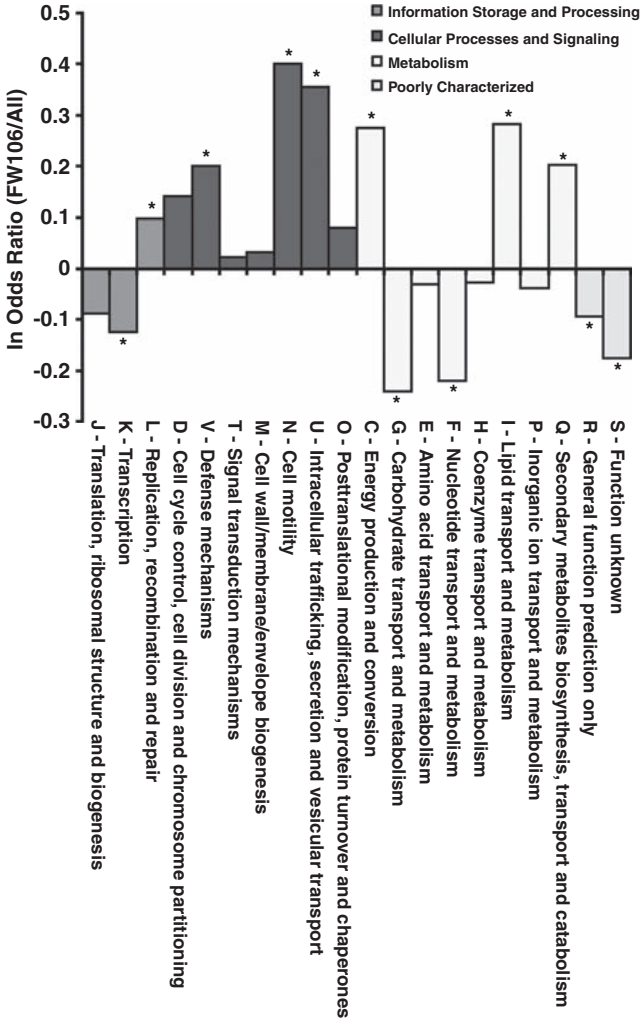

Figure 2 Odds ratios of FW106 genes compared to all sequenced bacteria for genes assigned by COG functional categories. Asterisks indicate significant deviation from the null hypothesis (ln odds ratio = 0) at the 95% confidence level by one-tailed Fisher exact test (Rosner, 2005).

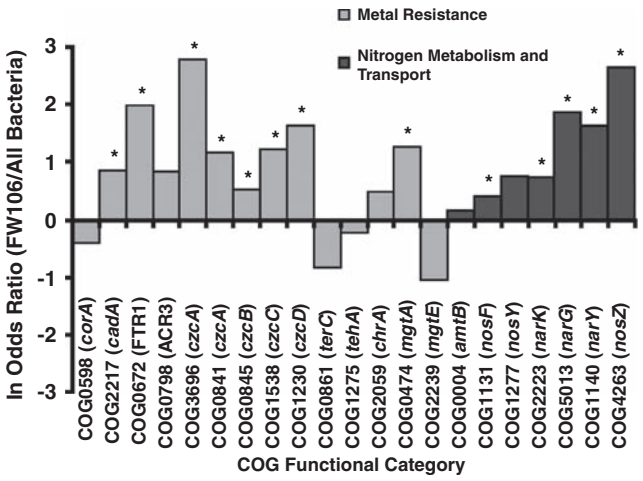

Figure 3 Odds ratios of FW106 genes compared to all sequenced bacteria for specific COG functional groups containing selected geochemical resistance genes. Asterisks indicate significant deviation from equality (ln odds ratio = 0) at the 95% confidence level by one-tailed Fisher exact test.

pathways are identified, as well as partial or complete organic acid metabolism pathways (acetate, lactate, butyrate, propionate and formate). Pathways are also identified for degradation of specific organic contaminants (for example, acetone, 1, 2-dichloroethene, methanol and formaldehyde). Pyruvate dehydrogenase complex components are present but not fermentative pyruvate conversion enzymes (for example, pyruvate formate-lyase or pyruvate/ferredoxin oxidoreductase). It is not known if the community carries out fermentation to a significant degree versus respiration, though *Clostridia* and other fermentative species may be present in the community at extremely low abundance.

Respiration is of particular interest because one of the major contaminants of the FRC, nitrate, is also an exceptional anaerobic terminal electron acceptor. FW106 γI (and possibly FW106 βI) uses a complete denitrification pathway for the conversion of nitrate and nitrite to N_2 (Figure 4, brown pathways). The abundant supply of terminal electron acceptor, the apparent lack of fermentation activity in the community and the low dissolved oxygen content of the site ($0.26\,mg\,l^{-1}$) suggest an obligate respiratory community deriving energy primarily from denitrification. FW106 γI encodes genes for *nasA* (assimilatory nitrate reductase) and *amt* (ammonium uptake transporters), as well as genes for two ammonium assimilation pathways (glutamate dehydrogenase and glutamine synthetase/glutamate synthase) and associated regulatory mechanisms (*ntrBC*, *glnBD*).

No evidence for the presence of sulfate-reducing bacteria or dissimilatory sulfate reduction was observed in the FW106 metagenome. FW106 γI does, however, encode a complete assimilatory sulfate reduction pathway. Reduction of sulfite to sulfide appears to be possible in FW106 βI, but a complete dissimilatory sulfate reduction pathway is not identified in this species; instead, sulfur assimilation in βI may involve uptake and interconversion of sulfur-containing amino acids such as taurine.

Metabolic adaptation to stress

A comprehensive list of genes relevant to survival under the unique geochemical conditions of FW106 is provided in Supplementary Table S3. Adaptations observed for specific stressors are described as follows.

Nitrate stress. Extremely high levels of nitrate impose severe stress on the community through the generation of toxic nitrite, and appropriate genetic determinants are needed for survival and

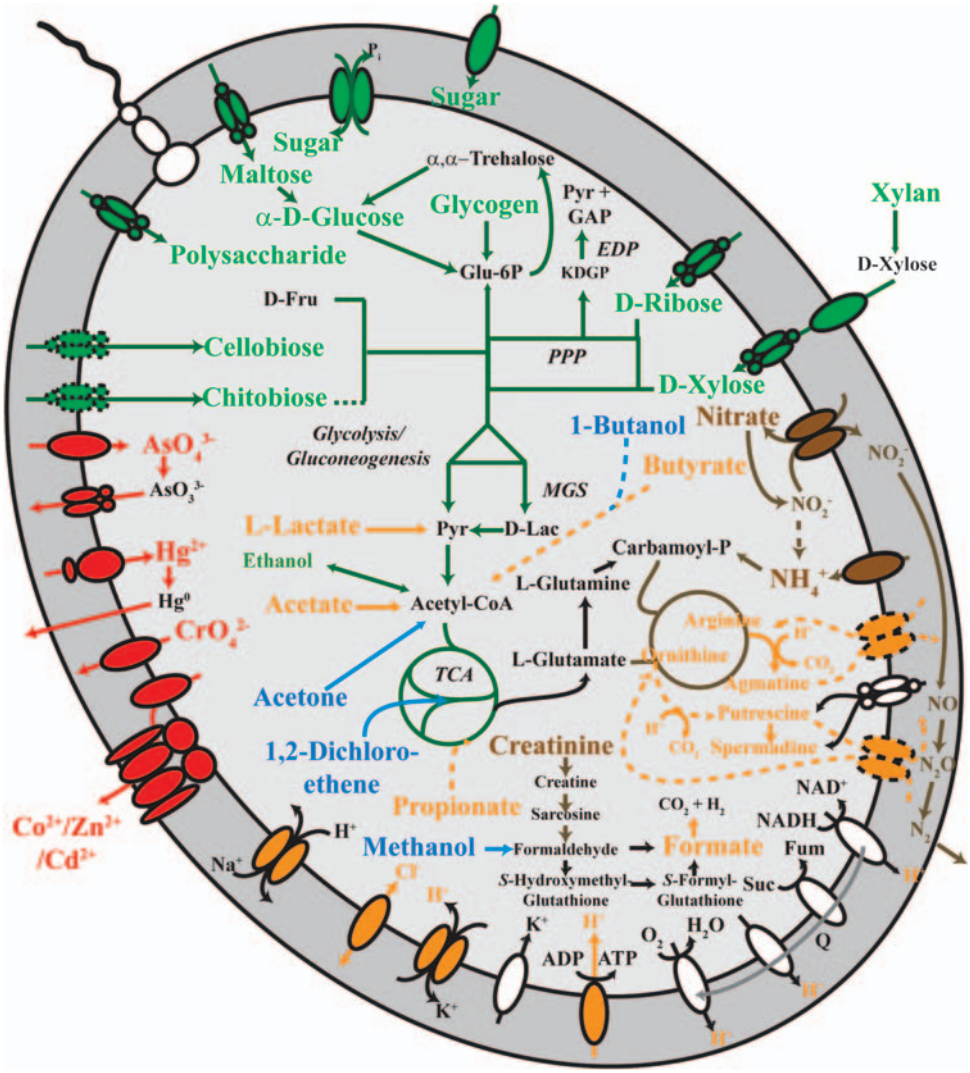

Figure 4 Reconstructed community metabolisms of the putative FW106 γI species. Partial, ambiguous or missing pathways/complexes are indicated by dashed lines. Pathways, compounds and transporters are colored as follows: Carbon metabolism (green), organic solvent detoxification (blue), heavy metal detoxification (red), denitrification and nitrogen metabolism (brown) and acid resistance (orange).

growth. Abundance profiling reveals an overabundance of *narK* nitrate/nitrite antiporters (COG2223, 10 genes), which transports nitrate from the periplasm to the cytoplasm where it is reduced to nitrite by NarG (DeMoss and Hsu, 1991). Nitrite is then transported to the periplasm, again by NarK, where it is ultimately converted to N_2 by denitrification (Figure 4, brown pathways).

pH stress. Metabolic reconstruction suggests several possibilities for acid resistance response. Under acidic conditions, protonated organic acids freely permeate the cell membrane and dissociate within the cytoplasm, resulting in decreased intracellular pH and disruption of the chemiosmotic gradient (Bearson *et al*., 1997). Maintenance of the chemiosmotic gradient under acidic conditions can be achieved by modulation of the intracellular pH by metabolism of organic acids, consumption of protons by amino-acid decarboxylation and/or by transport of protons and other small ions between the cytoplasm and periplasm (Bearson *et al*., 1997). Several such systems are implied by the FW106 γI metabolic network, including proton and small ion transport and organic acid metabolism pathways (Figure 4, orange pathways). Additional general stress response systems implicated in acid resistance (*rpoS*, *gshB*) were also identified. Although it is difficult to elucidate the acid stress response using genomic data alone, the metagenome does suggest possible mechanisms by which the FW106 community responds to acid stress.

Organic solvent stress. Degradation of organic contaminants typically requires specialized multistep pathways specific to a given class of compounds (Horvath, 1972). The FW106 metagenome reveals several putative degradation pathways

for dealing with specific organic contaminants present in the FW106 environment. In particular, FW106 γ1 uses pathways for the degradation of 1,2-dichloroethene and acetone, major contaminants of the site (Figure 4, cyan pathways). 1,2-Dichloroethene is a degradation product of tetrachloroethene, analysis of which has shown to be a major factor in controlling community structure in the FRC environment (Fields *et al.*, 2006). Additional pathways for the metabolism of methanol and detoxification of formaldehyde were also identified. Butanol may be degraded through the butyrate pathway, though not all of the necessary genes (for example, butanol dehydrogenase) are identified (Figure 4). In contrast, no complete pathways are identified for degradation of other major organic contaminants of the site, including aromatic compounds. The lack of specific degradation pathways in the FW106 community may be compensated for by more general stress response system such as the highly abundant AcrA/CzcA-like RFD multidrug efflux proteins.

Heavy metal stress. In contrast to organic contaminants, the metabolic mechanisms for resistance to heavy metal ions are relatively simple, typically involving (1) conversion of the ion to a less toxic form followed by efflux (for example, Hg^{2+}), (2) export of the metal ion to the periplasm followed by reduction to a lower oxidation state and decreased solubility of the ion (for example, U^{6+}, Cr^{6+}) and (3) export of the ion from the cell entirely (for example, Co^{2+}, Cd^{2+}, Zn^{2+}) (Silver and Phung, 1996). Many of the genes imparting these activities are known to be plasmid-borne and may easily be transferred between species (Silver and Phung, 1996). The FW106 community contains a variety of heavy metal resistance systems, including CadA-like heavy metal translocating ATPases (17 genes, COG0598/COG2217), ChrAB chromate efflux (4 genes, COG2059/COG4275), CzcABC $Co^{2+}/Zn^{2+}/Cd^{2+}$ efflux (62 genes, COG3696/COG1538/COG0845), CzcD-like $Co^{2+}/Zn^{2+}/Cd^{2+}$ efflux (14 genes, COG1230), *mer* operon mercuric resistance/regulation (35 genes, COG0789/COG1249/COG2608), TerC tellurium resistance (2 genes, COG0861) and CopRS-type heavy metal responsive two-component systems (8 genes, COG0745/COG6042) (Figure 4, red pathways; Supplementary Table S3). Heavy metals are predicted to represent a major stress on the community, and the abundance and diversity of predicted metal efflux proteins suggests that adaptation to metal stress may be of particular importance to community survival and has been a major factor in shaping the FW106 microbial community composition and structure.

Evolutionary processes affecting stress adaptation

Positive selection, gene duplication and lateral gene transfer (LGT) are three main evolutionary processes that propel evolution, but debate remains regarding the relative importance of these processes in microbial genome and community evolution (Ge *et al.*, 2005; Smets and Barkay, 2005). The relative importance of positive selection, gene duplication and LGT in microbial community evolution is examined in detail using computational and experimental metagenomic data.

Positive selection. The high concentrations of multiple contaminants at FW106 are expected to exert strong selective pressures on the community. Multiple statistical analyses have been developed to identity the effects of selection on related sequences. The ratio of nonsynonymous (d*N*) to synonymous (d*S*) nucleotide substitutions is a common albeit conservative method for identifying positive selection ($dN/dS > 1$) (Yang, 2003). Metagenome-wide pairwise dN/dS analyses of FW106 genes compared to closely related reference genes from GenBank were conducted using the Nei-Gojobori (Nei and Kumar, 2000) and maximum likelihood (Yang, 1997) methods. Analysis shows no definitive evidence of positive selection at the genetic level and that most genes are instead under strong negative selection (results not shown).

A total of 6161 SNPs are identified from the assembled FW106 read libraries, corresponding to ~1.2 SNP per kb. Of these SNPs, 2701 occur within coding sequences (835 synonymous, 1866 nonsynonymous). The overwhelming majority of the SNPs occur at low frequencies, almost always occurring only once in the assembled reads, suggesting clonal populations. This pattern of rare polymorphisms is consistent with models of recurrent selective sweeps, background selection and/or a recent population expansion (Nei and Kumar, 2000), followed by gradual accumulation of nearly neutral mutations. To further differentiate between these models, we directly amplified five representative genes of interest from FW106 metagenomic DNA. The resulting amplicons were used to construct a clone library for sequencing and population genetics analysis (Table 2). None of the five loci contain SNPs in the assembled metagenome but do exhibit a range of diversity when sequenced directly (3–186 segregating sites). Whether this discrepancy reflects additional diversity lost in metagenomic sequencing or inclusion of closely related orthologs and/or paralogs is unknown. To test for evidence of positive selection, we determined several population genetics statistics commonly used to detect the effects of selection and drift based on polymorphism data (Tajima's D, which detects deviations from neutrality by comparing nucleotide diversity to the number of segregating sites (Tajima, 1989); Fu and Li's D and F, which are variations of Tajima's test extended to include out-group sequences (Fu, 1997); Fu and Li's F_S, a neutrality test based on haplotype diversity (Fu, 1997); ZZ, a measure of linkage

Table 2 Population genetics analysis of sequenced FW106 genes

IMG GOID[a]	*2005744412*	*2005746176*	*2005744727*	*2005744725*	*2005742341*
Gene name	MFS	*adh*	*czcD*	*czcD*	*czcD*
Outlier gi #[b]	111017022	110832861	124514842	124267542	124265193
Sample size	30	66	77	52	20
No. of haplotypes	3	27	34	33	17
S[c]	3	38	186	32	16
# Syn Subs[d]	1	7	128	15	7
# Non Subs[d]	2	29	53	17	9
π_S[e]	0.00026	0.00168	0.05985	0.01539	0.01663
π_a[f]	0.00020	0.00236	0.00819	0.00336	0.00262
π_a/π_S[g]	0.788	1.403	0.132	0.216	0.156
K_a/K_S[h]	0.261	0.087	0.125	0.233	0.230
k[i]	0.200	1.747	16.827	3.588	3.511
π[j]	0.00022	0.00218	0.02155	0.00657	0.00643
θ_w[k]	0.00082	0.01076	0.04846	0.01297	0.00826
ZZ[l]	0.1665	0.1488	0.1748	0.2111	−0.0242
Tajima's D	−1.73 (NS)	−2.62***	−2.00*	−1.64 (NS)	−0.82 (NS)
Fu and Li's D*	−2.69*	−5.53**	−2.50*	−4.06*	−1.70 (NS)
Fu and Li's F*	−2.79*	−5.29**	−2.76*	−3.80*	−1.67 (NS)
Fu's F_s	−1.627*	−28.495*	−1.805 (NS)	−30.936*	−13.709*
Fu and Li's D[m]	−2.36 (NS)	−4.38*	−1.81 (NS)	−3.13**	−0.60 (NS)
Fu and Li's F[m]	−2.47*	−4.43**	−2.26 (NS)	−3.03**	−0.66 (NS)
Fay and Wu's H[m]	0.13 (NS)	−10.41***	−100.07*	−7.17 (NS)	−5.49*

Sequences were aligned and analyzed as described in Supplemental Materials and Methods.
Significance of each statistic is indicated as: *95% confidence level; **98% confidence level; ***99% confidence level; NS, not significant.
[a]IMG Gene Object Identifier number for FW106 loci corresponding to the clone group.
[b]gi # of GenBank best hit of FW106 reference gene based on TBLASTN (used as outlier).
[c]No. of segregating sites.
[d]No. of synonymous and nonsynonymous substitutions.
[e]Synonymous nucleotide diversity.
[f]Nonsynonymous nucleotide diversity.
[g]Intraspecific diversity.
[h]Interspecific divergence.
[i]Average number of nucleotide differences.
[j]Nucleotide diversity (per site).
[k]θ Per site, calculated from *S*.
[l]Test for level of linkage disequilibrium between polymorphic sites in relation to distance (Rozas *et al.*, 2001).
[m]Analyses using outlier sequences.

disequilibrium (Rozas *et al.*, 2001)). Negative values for Tajima's D and Fu and Li's D and F were obtained for all five loci, suggesting that negative selection has acted on these loci (Table 2). ZZ values suggest a low rate of recombination among loci. In this situation, the purging of deleterious loci by negative selection would result in the loss of diversity in linked loci (background selection), which could explain the observed metagenome-wide loss of diversity. However, negative values for Tajima's D can also result from demographic effects of a recent population expansion, which would affect allelic diversity across the entire genome through the process of random genetic drift (Fu, 1997). Fu's F_S statistic, which is sensitive to demographic effects, is significantly negative for four of the five loci, suggesting a recent population expansion. It thus appears that a combination of strong negative selection and a recent population expansion have reduced allelic diversity across the entire metagenome resulting in clonal populations, and positive selection appears to have little role in the microbial community evolution at the genetic level.

Lateral gene transfer. LGT has been suggested to be the primary evolutionary process by which short-term adaptation occurs in stressed soil communities (Rensing *et al.*, 2002). Previous studies suggest that LGT of geochemically relevant genes actively occurs between FRC populations (Martinez *et al.*, 2006). The FW106 metagenome permits a community-scale survey of such processes within an ecological context. Analysis of the major scaffolds (scaffold >100 kb, 3901 genes in total) revealed 277 (~7%) putative alien genes using SIGI-HMM, a scoring-based method for identifying genomic islands using hidden Markov models (Waack *et al.*, 2006; Langille *et al.*, 2008) (Supplementary Table S4). A manual survey of mobile elements (for example, transposons, insertion elements, integrases) suggests a rate of ~12 transpositions per Mb in the FW106 community. This is within the observed range of *Xanthomonas* species, the closest sequenced relatives of FW106 γI (Supplementary Table S5). These results suggest that the frequency of fixation of laterally transferred genes in FW106 γI is not significantly greater than in reference strains despite the stresses imposed on the cells. COG categories R

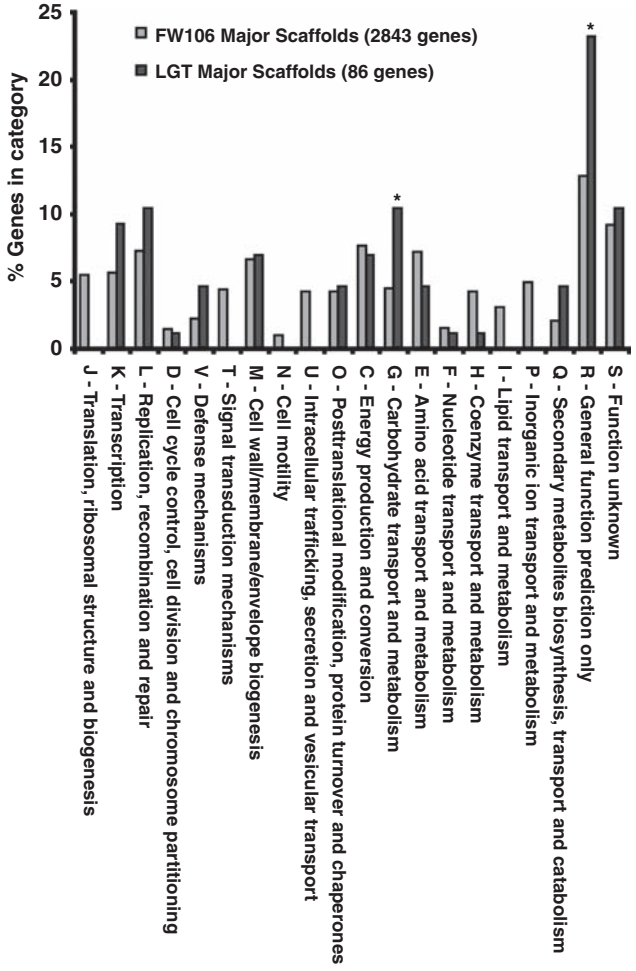

Figure 5 Percentage of laterally transferred genes of the major scaffolds (>100 kb) based on SIGI-HMM prediction. 277 total genes were detected, 86 of which are assigned to COG categories (3901 total major scaffold genes, 2843 assigned to COG categories). COG categories significantly enriched in the LGT dataset compared to major contig genes ($P>0.05$, binomial test) are indicated with an asterisk.

(general function prediction only) and G (carbohydrate transport and metabolism) are significantly overrepresented in the laterally transferred gene data set compared to all major scaffold genes (Figure 5).

Of particular interest are recently acquired genes fixed in the population as a result of contamination, as these genes are more likely to be relevant to survival under stressed conditions. Recent laterally transferred genes are expected to undergo little to no amelioration and thus are more likely to show distinct characteristics (for example, %GC, codon bias) compared to the genomic background (Lawrence and Ochman, 1997). Several methods are used to identify recently acquired genomic islands in the major scaffolds (>100 kb) and a comprehensive list of putative transferred genes is provided in Supplementary Table S5. Where statistical methods result in ambiguity, phylogenetic methods are used as well. Representative LGT events of geochemical interest are described below.

(a) *Acetone carboxylase*. The best example of a geochemically relevant LGT event observed in the community is the acquisition of at least one acetone carboxylation operon by FW106 γI (Figure 6). The predominant acetone metabolism pathway in bacteria, represented by *Xanthobacter autotrophicus*, involves the multistep conversion of acetone to acetyl-CoA, allowing the cell to subsist on acetone as the sole carbon source (Sluis and Ensign, 1997). LGT of the *Xanthobacter*-like acetone carboxylase Operon A is strongly implied by multiple lines of evidence. Discriminant analysis, SIGI-HMM and visual inspection show significant deviations in sequence characteristics (for example, %GC) in the operon from the genomic background (Supplementary Figure S4; Supplementary Table S6). Phylogenetic analysis of the concatenated acetone carboxylase subunits suggests that the genes of Operon A are likely functional orthologs of the characterized *Xanthobacter* genes and further suggests a β-proteobacterial origin (Figure 6; Supplementary Figure S5). Finally, both operons are associated with transposons and other mobile elements (Figure 6). Multiple lines of evidence thus suggest lateral acquisition of acetone carboxylase activity by the dominant γ-proteobacterial species.

(b) *Mercuric resistance (mer) operons*. Mercury is a major contaminant at the FRC (de Lipthay *et al.*, 2008) and mercuric resistance genes in general are known to be frequent targets of lateral transfer (Silver and Phung, 1996). Supplementary Figure S6 shows the distribution and arrangement of *mer* operon genes compared to the typical arrangement and gene complement of the *mer* operon (reference *mer* operon). Eight partial or complete *mer* operons as well as additional *mer* operon genes were identified in the FW106 metagenome (Supplementary Figure S6) and many of these gene clusters are associated with mobile elements and other metal resistance genes. The association of many of these operons with mobile element genes, the abundance of *mer* operon genes in the metagenome and shuffling of the *mer* operon genes within the metagenome suggest active lateral transfer of mercuric resistance within the population in response to mercury contamination.

(c) *czcD divalent cation transporter*. One of the most abundant genes in the FW106 metagenome encodes the CzcD efflux complex that transports divalent cations from the cytosol to the periplasm and ultimately to the cell exterior (in concert with CzcABC). The high abundance of these genes suggests they have a critical function in heavy metal resistance by the FW106 community. Phylogenetic analysis of FW106 *czcD* genes further suggests that some of these genes may have originated from α-Proteobacteria and Actinobacteria species (Supplementary Figure S7), which are known to be

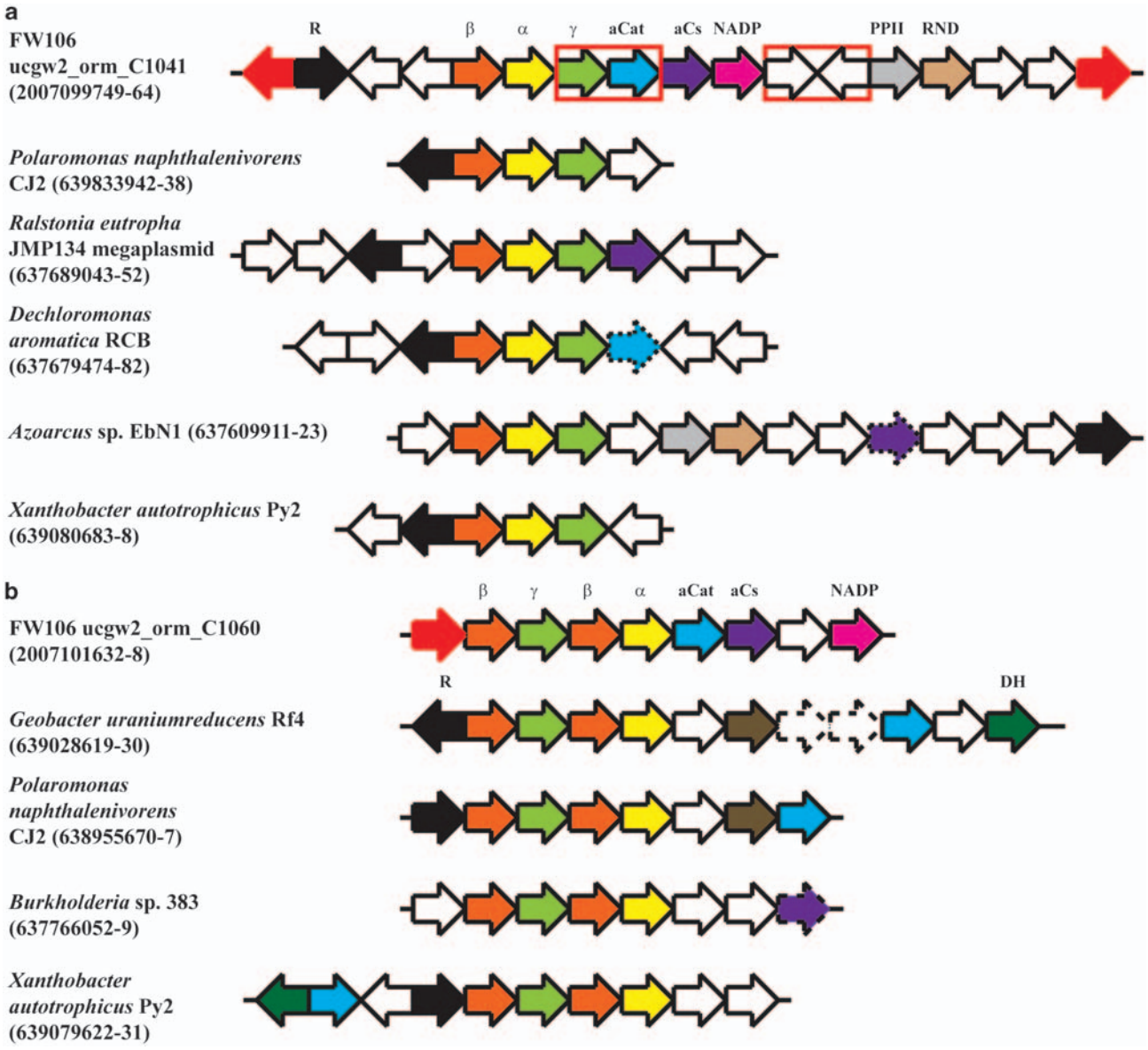

Figure 6 Structure of two putative acetone carboxylation operons of FW106. The classical operon structure is shown in (**a**) and an operon with a modified structure containing two β subunits is shown in (**b**). Open reading frames (ORFs) colored red represent mobile elements. ORFs colored white represent non-homologous genes and all other colored ORFs indicate orthologous groups. Genes with dotted outline represent putative non-orthologous functional analogs. Red boxes indicate putative alien genes as determined by SIGI-HMM. IMG gene object identifiers are listed after the species name and conserved genes are labelled as follows: aCat, acetyl-CoA acetyltransferase; aCs, acyl-CoA synthetase; DH, alcohol dehydrogenase; PPII, Uncharacterized protein related to plant photosystem II stability/assembly factor; R, Fis-type helix–turn–helix activator of acetoin/glycerol metabolism; RND, predicted exporters of the RND superfamily; α/β/γ, subunits of acetone carboxylase.

present in abundance in pristine FRC groundwater communities (Fields *et al.*, 2005).

Discussion

The study of microbial ecology and evolution has been revolutionized by culture-independent metagenomic analysis (Handelsman *et al.*, 2007). In this study, metagenomic approaches were used to analyze the diversity, structure and evolution of a groundwater microbial community in an extreme low-pH environment contaminated with high levels of uranium, nitric acid, technetium and organic solvents. This represents the first metagenomic analysis focusing on the responses and adaptation of groundwater microbial communities to human-induced environmental change. Because groundwater is a key limiting resource and its restoration after pollution is of major importance, the results from this study are of great interest to scientists from broad fields such as geochemists, biologists, ecologists, hydrologists, regulatory officials and policy makers.

Both SSU rRNA gene-based cloning and random shotgun sequencing approaches reveal a very simple FW106 community with less than 13 OTUs and dominated by denitrifying γ- and β-proteobacterial species. Previous studies have observed approximately ~160 (97% cutoff) OTUs pristine groundwater from the FRC background site (FW300, 2 km away (Fields *et al.*, 2005)). These results suggest that anthropogenic chemical contamination has had a dramatic negative impact on microbial community diversity, with an order of magnitude reduction in OTU abundance.

The introduction of contaminants has not only dramatically reduced microbial community diversity at the site but has also had a significant effect on community metabolic diversity. Previous studies based on functional gene markers (*nirS, nirK, dsrAB, amoA, pmoA*) have revealed very high microbial functional diversity at the FRC (Yan *et al.*, 2003; Palumbo *et al.*, 2004; Fields *et al.*, 2005; Hwang *et al.*, 2009), suggesting that the key biogeochemical functional processes such as denitrification, sulfate reduction, nitrification and methane oxidation exist in the subsurface environment. Also, several types of known metal- and sulfate-reducing bacteria (for example, *Geobacter*, *Anaeromyxobacter*, *Desulfovibrio*, *Desulfitobacterium*) have been observed in various FRC sites (Petrie *et al.*, 2003; Brodie *et al.*, 2006; Hwang *et al.*, 2009). However, metabolic reconstruction based on metagenomic sequencing suggests that the FW106 community has retained denitrification activity, but not dissimilatory sulfate reduction, metal reduction, nitrification and methane oxidation activities. However, due to potential undersampling and/or low abundance of these functional groups, direct functional activity analyses are needed to verify this finding.

Analysis suggests that specific contaminants at the site impose strong selective pressures that act to shape the structure of the community. Nitrate likely acts both as the primary terminal electron acceptor for the community and as the primary source of biological nitrogen. Furthermore, the high nitrate concentrations favor denitrifying species while suppressing the activity and abundance of sulfate- and Fe-reducing bacteria at this site despite the fact that such bacteria are known to be active at the FRC. These observations, coupled with the loss of most complex carbohydrate metabolic activities, have resulted in a heterotrophic community that produces energy primarily through denitrification and/or oxygen respiration. The FW106 community has also accumulated genes for degradation of specific organic contaminants including acetone and chlorinated hydrocarbons and may possibly be able to subsist on some of these compounds as carbon sources (for example, acetone). Finally, the complement and abundance of numerous heavy metal resistance systems (that is, *czc* divalent cation transporters, mercuric resistance genes, cytochromes implicated in dissimilatory metal resistance) is consistent with the hypothesis that heavy metal stress is a key environmental factor shaping the structure and function of the FW106 community. However, experimental analysis will be necessary to test this hypothesis.

Adaptation of biological communities to environmental stress is a critical issue in ecology. Metagenomic analyses indicate that the microbial community is well adapted to the geochemical conditions at this site as evidenced by the overabundance of key genes conferring resistance to specific contaminants. Nitrate, heavy metals (for example, divalent cations, mercury) and organic solvents (for example, chlorinated hydrocarbons, acetone) in particular have key functions in shaping the genome and community structure of FW106. Although the majority of microbial populations may have become extinct after the introduction of contaminants, certain community members with key metabolic activities related to denitrification and metal resistance survived to form the foundation of the new community. The results have important implications in understanding, assessing and predicting the impacts of anthropogenic activities on microbial communities ranging from human health to agriculture to environmental management, and their responses to environmental changes.

Sequence analysis revealed no definitive evidence for positive selection in the metagenome, though the extremely low allelic diversity and accumulation of geochemical resistance genes indirectly suggest recurrent selective sweeps. Complicating efforts to detect positive selection events is the possible role of niche differentiation in the FRC communities. The FRC site is a complex three-dimensional geochemical network where local geological conditions can have a significant effect on local geochemistry over short distances, possibly resulting in the formation of ecological traps (Dwernychuk, 1972; Phillips *et al.*, 2008). As such, mutations in a particular genetic background may only confer adaptive phenotypes in a very specific niche or micro-niche (Sokurenko *et al.*, 2004). Thus, further work, including high-resolution temporal and spatial metagenomic sequencing, is necessary to verify the adaptive processes at work in stressed groundwater ecosystems. In addition, many genes important to geochemical resistance appear to have been laterally transferred within the community, and thus LGT could be important in the adaptation of the microbial community to contaminant stress. However, it is not clear whether these events occurred before or after the introduction of the contaminants, the latter of which would indicate adaptation in response to contaminant-induced stress. The working hypothesis is thus that most of the observed LGT events were fixed in the population in response to contamination and hence likely occurred after the introduction of contaminants to the site. Alternatively, the LGT events could have occurred before introduction of contamination and only existed at

low frequency until selected for after the introduction of contaminants. However, the currently available FRC metagenome sequence data are insufficient to resolve this issue. Whole-genome sequencing of culturable isolates from other contaminated and noncontaminated sites as well as targeted metagenome sequencing designed to establish temporal and spatial relationships between FRC sites will be needed to test this hypothesis.

Acknowledgements

We thank Dr Fares Najar and Dr Bruce Roe for providing sequencing services, and Dr Tommy Phelps and Dr Christopher W Schadt for assisting groundwater sampling. This research was supported by The United States Department of Energy under the Environmental Remediation Sciences Program (ERSP), and Genomics: GTL program through the Virtual Institute of Microbial Stress and Survival (VIMSS; http://vimss.lbl.gov), Office of Biological and Environmental Research, Office of Science, and by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under Contract No. DE-AC02-06NA25396. Oak Ridge National Laboratory is managed by University of Tennessee UT-Battelle LLC for the Department of Energy under Contract No. DE-AC05-00OR22725. All authors contributed intellectual input and assistance to this study and paper preparation. The original concept and experimental strategy were developed by JZ and MWF. Sampling collections and DNA preparation were performed by TG and LW. DW performed chemical analysis of the groundwater sample. KB and SGT oversaw metagenomic sequencing and assembly. CH performed all sequence and evolutionary analysis. YD assisted in computational analysis of metagenome sequences. SB performed PCR experiments for population genetics analysis and LGT confirmation. JZ and CH performed data synthesis, and took the lead in writing the paper.

References

Allen EE, Banfield JF. (2005). Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* **3**: 489–498.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**: 3389–3402.

Bearson S, Bearson B, Foster JW. (1997). Acid stress responses in enterobacteria. *FEMS Microbiol Lett* **147**: 173–180.

Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL *et al.* (2006). Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Appl Environ Microbiol* **72**: 6288–6298.

Chou H-H, Holmes MH. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics* **17**: 1093–1104.

de Lipthay JR, Rasmussen LD, Oregaard G, Simonsen K, Bahl MI, Kroer N *et al.* (2008). Acclimation of subsurface microbial communities to mercury. *FEMS Microbiol Ecol* **65**: 145–155.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

DeMoss JA, Hsu PY. (1991). NarK enhances nitrate uptake and nitrite excretion in *Escherichia coli*. *J Bacteriol* **173**: 3303–3310.

Dwernychuk L. (1972). Ducks nesting in association with gulls—an ecological trap. *Can J Zool* **50**: 559.

Fields MW, Yan T, Rhee SK, Carroll SL, Jardine PM, Watson DB *et al.* (2005). Impacts on microbial communities and cultivable isolates from groundwater contaminated with high levels of nitric acid-uranium waste. *FEMS Microbiol Ecol* **53**: 417–428.

Fields MW, Bagwell CE, Carroll SL, Yan T, Liu X, Watson DB *et al.* (2006). Phylogenetic and functional biomakers as indicators of bacterial community responses to mixed-waste contamination. *Environ Sci Technol* **40**: 2601–2607.

Fu YX. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

Ge F, Wang L-S, Kim J. (2005). The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol* **3**: e316.

Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS *et al.* (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355–1359.

Handelsman J, Tiedje JM, Alvarez-Cohen L, Ashburner M, Cann IKO, DeLong EF *et al.* (2007). *Committee on Metagenomics: Challenges and Functional Applications*. National Academy of Sciences: Washington, DC.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC *et al.* (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* **1**: 67–77.

Horvath RS. (1972). Microbial co-metabolism and the degradation of organic compounds in nature. *Bacteriol Rev* **36**: 146–155.

Hwang C, Wu W, Gentry TJ, Carley J, Corbin GA, Carroll SL *et al.* (2009). Bacterial community succession during *in situ* uranium bioremediation: spatial similarities along controlled flow paths. *ISME J* **3**: 47–64.

Langille M, Hsiao W, Brinkman F. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinform* **9**: 329.

Lawrence JG, Ochman H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**: 383–397.

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

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D *et al.* (2008). IMG/M: a data management and analysis system for metagenomes. *Nucl Acids Res* **36**(Database issue): D534–D538. gkm869.

Martinez RJ, Wang Y, Raimondo MA, Coombs JM, Barkay T, Sobecky PA. (2006). Horizontal gene transfer of PIB-type ATPases among bacteria isolated from

radionuclide-and metal-contaminated subsurface soils. *Appl Environ Microbiol* **72**: 3111–3118.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**: 63–72.

Nei M, Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press: New York, NY.

Palumbo AV, Schryver JC, Fields MW, Bagwell CE, Zhou JZ, Yan T *et al.* (2004). Coupling of functional gene diversity and geochemical data from environmental samples. *Appl Environ Microbiol* **70**: 6525–6534.

Petrie L, North NN, Dollhopf SL, Balkwill DL, Kostka JE. (2003). Enumeration and characterization of Iron(III)-reducing microbial communities from acidic subsurface sediments contaminated with Uranium(VI). *Appl Environ Microbiol* **69**: 7467–7479.

Phillips DH, Watson DB, Kelly SD, Ravel B, Kemner KM. (2008). Deposition of uranium precipitates in dolomitic gravel fill. *Environ Sci Technol* **42**: 7104–7110.

Rensing C, Newby DT, Pepper IL. (2002). The role of selective pressure and selfish DNA in horizontal gene transfer and soil microbial community adaptation. *Soil Biol Biochem* **34**: 285–296.

Riesenfeld CS, Schloss PD, Handelsman J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38**: 525–552.

Rozas J, Rozas R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174–175.

Rozas J, Gullaud M, Blandin G, Aguade M. (2001). DNA variation at the rp49 gene region of drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics* **158**: 1147–1155.

Silver S, Phung LT. (1996). Bacterial heavy metal resistance: new surprises. *Annu Rev Microbiol* **50**: 753–789.

Sluis MK, Ensign SA. (1997). Purification and characterization of acetone carboxylase from *Xanthobacter* strain Py2. *Proc Natl Acad Sci USA* **94**: 8456–8461.

Smets BF, Barkay T. (2005). Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. *Nat Rev Microbiol* **3**: 675–678.

Sokurenko EV, Feldgarden M, Trintchina E, Weissman SJ, Avagyan S, Chattopadhyay S *et al.* (2004). Selection footprint in the fimh adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. *Mol Biol Evol* **21**: 1373–1383.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C *et al.* (2002). The bioperl toolkit: perl modules for the life sciences. *Genome Res* **12**: 1611–1618.

Tajima F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Tamura K, Dudley J, Nei M, Kumar S. (2007). MEGA4: molecular evolutionary genetics analysis (mega) software version 4.0.. *Mol Biol Evol* **24**: 1596–1599. msm092..

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. (2007). The human microbiome project. *Nature* **449**: 804–810.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial
genomes from the environment. *Nature* **428**: 37–43.

Voget S, Steele HL, Streit WR. (2006). Characterization of a metagenome-derived halotolerant cellulase. *J Biotechnol* **126**: 26–36.

Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke W *et al.* (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinform* **7**: 142.

Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583.

Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H *et al.* (2006). Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol* **4**: e188.

Yan T, Fields MW, Wu L, Zu Y, Tiedje JM, Zhou J. (2003). Molecular diversity and characterization of nitrite reductase gene fragments (*nirK* and *nirS*) from nitrate- and uranium-contaminated groundwater. *Environ Microbiol* **5**: 13–24.

Yang Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.

Yang Z. (2003). Adaptive molecular evolution. In: Balding DJ, Bishop M, Cannings C (eds) *Handbook of Statistical Genetics*, 2nd edn. John Wiley and Sons, Ltd: New York, NY, pp 229–254.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The sorcerer ii global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)