Lawrence Berkeley National Laboratory

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

Permalink
https://escholarship.org/uc/item/0j9526fb

Authors
Hemme, C.L.
Deng, Y.
Gentry, T.J.
et al.

Publication Date
2010-09-01

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 Green-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, University of Oklahoma, Norman, OK, USA, 73019;* [2]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA, 37831-6038;* [3]*Department of Soil Sciences, Texas A&M University, College Station, TX, USA, 77843;* [4]*Department of Microbiology, Montana State University, Bozeman, MT, USA, 59717;* [5]*U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA, USA, 94598;* [6]*Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, 94270;* [7]*Center for Microbial Ecology, Michigan State University, East Lansing, MI, USA, 48824*

[†]corresponding author

Dr. Jizhong Zhou

Institute for Environmental Genomics

University of Oklahoma

Norman, OK 73019

Phone: (405) 325-6073

Fax: (405) 325-3442

Email: jzhou@rccc.ou.edu

Subject Category: Genomic and Computational Biology, Microbiology

Total Character Count: 37731/60000

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) have 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 over-abundant in key genes conferring resistance to specific stresses including nitrate, heavy metals and acetone. Evolutionary analysis indicates that lateral gene transfer could be a key mechanism in rapidly responding and adapting 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. (170/175 words)

Introduction

Microorganisms are the most abundant and diverse group of life on the planet and play an integral role 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 (Handelsman et al, 2007; He et al, 2007; Riesenfeld et al, 2004). 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 & 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 have not to date 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

studies (Allen & Banfield, 2005).   In this study, a microbial community from highly uranium-contaminated groundwater was sequenced using a random shotgun sequencing-based strategy with the goal of addressing the following questions: (i) 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? (iii) What are the molecular mechanisms responsible for such environmental changes? Results reveal novel insights into microbial community diversity, structure and function in a contaminated ecosystem and mechanisms by which microbial communities adapt to extreme levels of contamination.

Results

Overview of the metagenomic sequencing

(i) Phylogenetic diversity of the sampling site. Groundwater from well FW106 at Oak Ridge Environmental Remediation Sciences Program (ERSP) Field Research Center (FRC) is highly acidic (pH 3.7), and contaminated with extremely high levels of uranium (among the highest in the U.S.), nitrate, technetium and various organic contaminants (Table S1). Microscopic analysis suggests a simple community structure with 2-3 different cell types dominating the sample (Fig. S1). Similarly, SSU rRNA gene-based phylogenetic analysis reveals very low phylogenetic diversity with a total of 13 operational taxonomic units (OTU's) from 619 sequences at the 98% sequence identity cutoff, with ~87% of these sequences corresponding to the BFXI557 γ-proteobacterial

clone (Fig. S2). The community is composed primarily of γ- and β-*proteobacteria* and dominated by *Rhodanobacter*-like γ-proteobacterial and *Burkholderia*-like β-proteobacterial species (Fig. S2).

(ii) 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) (Table S2, Fig. S2). Contigs from the pga assembly were binned using PhyloPythia (McHardy et al, 2007) (Table 1; Fig. 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 (Fig. S3); however, the lack of closely related reference genomes complicates phylogenetic assignment of this metagenome. While 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 12335 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 (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 using the IMG phylogenetic profiling tool (Fig. 1B). While 16S rRNA analysis and field

experiments show dominance of the community by γ-proteobacteria species, β-proteobacteria constituted the largest reservoir of assigned functional genes (18%) followed by γ- (12%) and α-proteobacteria (3%) (Fig. 1B). The dominant lineages in FW106 based on protein assignment are *Burkholderiaceae*, *Xanthomonadaceae* and *Comamonadaceae*, consistent with previous analyses (Fig. 1B).

(iii) 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 (Fig. S4). Overabundance of defense and repair mechanisms for dealing with stress-induced damage as well as 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 driven by specific contaminants such as nitrate and heavy 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) (Fig. 2). Abundance profiling of FW106 with other metagenomes using the IMG abundance profiling tools show similar results (results not shown). Accumulation of genes involved in resistance and stress response mechanisms thus appears to be a basic survival strategy employed by the community in response to the specific contaminants in FW106.

Metabolic Reconstruction of FW106 γI

To better understand stress mechanisms involved in the survival of FW106 microbial community on a genomic scale and to gain a comprehensive view of the metabolic capabilities of the community, metabolic reconstruction was performed 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 (Fig. 3) and a partial reconstruction of FW106 βI (data not shown). While 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 (Table S1).

Reconstruction of central carbon pathways and identification of carbon transport systems suggests the community subsists primarily on simple mono- and disaccharides, including cellulosic degradation products (e.g. 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 (Fig. 2). Complete glycolytic, TCA, pentose phosphate, Entner-Doudroff and methylglyoxal 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 (e.g. acetone, 1, 2-dichloroethene, methanol and formaldehyde). Pyruvate dehydrogenase

8

complex components are present but not fermentative pyruvate conversion enzymes (e.g. 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) employs a complete denitrification pathway for the conversion of nitrate and nitrite to N_2 (Fig. 3, 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) 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 (SRB) 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 Table S3. Adaptations observed for specific stressors are described as follows:

(i) 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 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 & Hsu, 1991). Nitrite is then transported to the periplasm, again by NarK, where it is ultimately converted to N_2 via denitrification (Fig. 3, brown pathways).

(ii) pH stress. Metabolic reconstruction suggests several possible mechanisms of acid resistance. 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 via 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 (Fig. 3, orange pathways). Additional general stress mechanisms implicated in acid resistance (*rpoS*, *gshB*) were also identified. While it is difficult to elucidate the acid stress response using genomic data alone, these mechanisms may serve as the functional core of acid stress response by the FW106 bacteria.

(iii) Organic solvent stress. Degradation of organic contaminants typically requires specialized multistep pathways specific to a given class of compounds (Horvath, 1972). The FW106 community employs several such mechanisms to deal with specific organic contaminants present in the FW106 environment. In particular, FW106 γ1 utilizes pathways for the degradation of 1,2-dichloroethene and acetone, major contaminants of the site (Fig. 3, cyan pathways). 1,2-dichloroethene is a degradation product of tetrachloroethene which analysis 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 via the butyrate pathway, though not all of the necessary genes (e.g. butanol dehydrogenase) are identified (Fig. 3). In contrast, no complete pathways are identified for degradation of other major organic contaminants of the site, including aromatic compounds. More general response mechanisms identified in the metagenome, such as the highly abundant AcrA/CzcA-like RFD multidrug efflux proteins, may compensate for the lack of specific degradation pathways by exporting toxic organics from the cell.

(iv) Heavy metal stress. In contrast to organic contaminants, the metabolic mechanisms for resistance to heavy metal ions are relatively simple, typically involving: (a) conversion of the ion to a less toxic form followed by efflux (e.g. Hg^{2+}); (b) export of the metal ion to the periplasm followed by reduction to a lower oxidation state and decreased solubility of the ion (e.g. U^{6+}, Cr^{6+}, etc.); and (c) export of the ion from the cell entirely (e.g. Co^{2+}, Cd^{2+}, Zn^{2+}, etc.) (Silver & Phung, 1996). Many of the genes imparting these activities are known to be plasmid-borne and may easily be transferred between species (Silver & 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) (Fig. 3, red pathways; Table S3). Heavy metals represent a major stress on the community and the abundance and diversity of metal efflux mechanisms suggests that adaptation to metal stress is of particular importance to community survival and has been a major factor in shaping the FW106 microbial community composition and structure.

Evolutionary mechanisms of stress adaptation

Positive selection, gene duplication and lateral gene transfer (LGT) are three main evolutionary mechanisms that drive evolution, but debate remains regarding the relative importance of these processes in microbial genome and community evolution (Ge et al, 2005; Smets & 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.

(i) Positive selection. The high concentrations of multiple contaminants at FW106 are expected to exert strong selective pressures on the community. Metagenome-wide pairwise *dN/dS* analyses of FW106 genes compared to closely-related reference genes from GenBank was conducted using the Nei-Gojobori (Nei & 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 single nucleotide polymorphisms (SNPs) are identified from the assembled FW106 read libraries, corresponding to ~1.2 SNP/kb. 2701 of these SNPs 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 & Kumar, 2000), followed by gradual accumulation of nearly neutral mutations. To further differentiate between these

models, five representative genes of interest were directly amplified from FW106 metagenomic DNA, sequenced, and used for population genetics analysis (Table S4). All five loci showed no SNPs in the assembled metagenome but do exhibit a range of diversity when sequenced directly (3-186 segregating sites). 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 S4). ZZ values for all loci suggest a low rate of recombination. 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 such 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 4 of the 5 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 play little role in the microbial community evolution at the genetic level.

(ii) Lateral gene transfer. LGT has been suggested to be the primary evolutionary mechanism in stressed soil communities leading to adaptive strains in the short term (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. SIGI-HMM (Langille et al, 2008; Waack et al, 2006) analysis identifies 277 (~7%) genes from major scaffolds (scaffold >100 kb, 3901 genes in total) as putative alien genes (Table S5). A manual survey of mobile elements (e.g., transposons, insertion elements, and integrases) suggests a rate of ~12 transpositions/Mb in the FW106 community. This is within the observed range of *Xanthomonas* species, the closest sequenced relatives of FW106 γI (Table S6). 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 (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 (Fig. S5).

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 (e.g. %GC, codon bias, etc.) compared to the genomic background (Lawrence & Ochman, 1997). Several methods are employed to identify recently acquired GIs (genomic islands) in the major scaffolds (>100 kb) and a comprehensive list of putative transferred genes is provided in Table S5. Where statistical methods result in ambiguity, phylogenetic methods are employed 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 (Fig. 4). 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 & 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 (e.g. %GC) in the operon from the genomic background (Figure S6; Table S7). 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 (Fig. 4; Fig. S7). Finally, both operons are associated with transposons and other mobile elements (Fig. 4). 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 & Phung, 1996). Eight partial or complete *mer* operons as well as additional *mer* operon genes were identified in the FW106 metagenome (Fig. S8) and at least four of these clusters are in the metagenome. 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 play a critical role 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 (Fig. S9), which are known to be 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 metagenomics analysis (Handelsman et al, 2007).   In this study, metagenomics 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 metagenomics analysis focusing on the responses and adaptation of groundwater microbial communities to human-induced environmental change. Since groundwater is a key limiting resource and its restoration following 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 OTU's and dominated by denitrifying γ- and β-proteobacterial species. Previous studies have observed approximately ~160 (97% cut-off) OTU's pristine groundwater from the FRC background site (FW300, 2 km away (Fields et al, 2005)). These results show 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 (Fields et al, 2005; Hwang et al, 2008; Palumbo et al, 2004; Yan et al, 2003), 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 (e.g., *Geobacter, Anaeromyxobacter, Desulfovibrio, Desulfitobacterium*) have been observed in various FRC sites (Brodie et al, 2006; Hwang et al, 2008; Petrie et al, 2003). However, metabolic reconstruction based on metagenomics sequencing suggests that the FW106 community has retained denitrification activity, but not dissimilatory sulphate reduction, metal reduction,

nitrification, and methane oxidation activities. However, due to potential under-sampling 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 election acceptor for the community but also as the primary source of biological nitrogen. Furthermore, the high nitrate concentrations favour 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 (e.g. acetone). Finally, toxic heavy metal stress has resulted in accumulation of multiple heavy metal resistance genes, particularly those for divalent cation efflux (*czcABC, czcD*), mercuric resistance and possibly cytochrome-mediated dissimilatory metal reduction (cytochrome c_{553}). Thus, prolonged exposure to high concentrations of mixed contaminants has had a profound affect on the structure and metabolic activities of the FW106 community.

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 over-abundance of key genes conferring resistance to specific contaminants. Nitrate, heavy metals (e.g. divalent cations, mercury) and organic solvents (e.g. chlorinated hydrocarbons, acetone) in particular have played key roles in shaping the genome and community structure of FW106. Although the majority of microbial populations may have gone extinct following 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 suggests 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 clarify the adaptive mechanisms 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 play important role 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. However, the currently available FRC metagenome sequence data is insufficient to resolve this issue. Whole genome sequencing of similar isolates from other contaminated and non-contaminated sites will be needed to test this hypothesis.

Materials and Methods Summary

Microbial biomass was collected from ~1,700 L groundwater from well FW106 by filtration. This well is acidic with pH of 3.7, and highly contaminated with uranium, technetium-99, nitrate and a variety of organic contaminants including acetone, 1,2-dichloroethene and benzoic acid (Table S1). High molecular weight community DNA

was extracted using grinding, freezing-thawing SDS-based methods (Zhou et al, 1996) and the purified DNA was used for random shotgun sequencing (Tyson et al, 2004). Double-ended sequencing reactions were performed using PE BigDye terminator chemistry and resolved using PRISM 3730 capillary DNA sequencer. ~70 Mb of raw sequence yielded ~53 Mb Q20 sequence in three clone libraries (20.04 Mb from small insert (3 kb) pUC library, 23.13 Mb from medium insert (8kb) pMCL, 9.27 Mb from large insert (40kb) pCCiFos). Contig assembly was conducted alternatively using i) Phrap and ii) Lucy (vector and quality trimming) (Chou & Holmes, 2001) and the Paracel Genome Assembler (pga) (Paracel, Pasadena, CA), resulting in 9,554,544 bp assembled into 6079 contigs ranging in size from 100 bp to 575 kb from the pga assembly. pga contigs were binned using PhyloPythia (McHardy et al, 2007) and genes were assigned to phylogenetic taxa using the IMG phylogenetic profiling tools. Gene prediction, functional assignment and metabolic reconstruction were performed automatically using internal JGI protocols. SNPs were detected using ad hoc computational methods using BioPerl (Stajich et al, 2002). Oligonucleotide primers for all described experiments were designed based on the assembled FW106 metagenomic sequence and population genetics parameters were determined using DnaSP (Rozas & Rozas, 1999). Laterally transferred genes were detected using a combination of composition-based and phylogenetic methods. Pairwise analyses of major scaffold genes were conducted using PAML (Yang, 1997). Phylogenetic analyses were conducted using MEGA 4.0 (Tamura et al, 2007) for functional gene analysis and with ARB (Ludwig et al, 2004) for 16S analysis. Metagenomic sequences are deposited in the JGI-IMG database

(Markowitz et al, 2007).   Details for all methods are provided in Supplementary

Materials and Methods.

References

Allen EE, Banfield JF (2005) Community Genomics in Microbial Ecology and Evolution. *Nature Reviews Microbiology* **3:** 489-498

Bearson S, Bearson B, Foster JW (1997) Acid stress responses in enterobacteria. *FEMS Microbiology Letters* **147:** 173-180

Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, Hazen TC, Richardson PM, Herman DJ, Tokunaga TK, Wan JM, Firestone MK (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, Sørensen SJ (2008) Acclimation of subsurface microbial communities to mercury. *FEMS Microbiology Ecology* **65:** 145-155

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM (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. *Canadian Journal of Zoology* **50:** 559

Fields MW, Bagwell CE, Carroll SL, Yan T, Liu X, Watson DB, Jardine PM, Criddle CS, Hazen TC, Zhou J (2006) Phylogenetic and Functional Biomarkers as Indicators of Bacterial Community Responses to Mixed-Waste Contamination. *Environ Sci Technol* **40:** 2601-2607

Fields MW, Yan T, Rhee SK, Carroll SL, Jardine PM, Watson DB, Criddle CS, Zhou J (2005) Impacts on microbial communities and cultivable isolates from groundwater contaminated with high levels of nitric acid-uranium waste. *FEMS Microbiology Ecology* **53:** 417-428

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 Biology* **3:** e316

Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (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, Doolittle WF, Fraser-Liggett CM, Godzik A, Gordon JI, Riley M, Schmid MB, Reid AH. (2007) Committee on Metagenomics: Challenges and Functional Applications. National Academy of Sciences, Washington, D.C.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, Criddle C, Zhou J (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. *Bacteriological Reviews* **36:** 146-155

Hwang C, Wu W, Gentry TJ, Carley J, Corbin GA, Carroll SL, Watson DB, Jardine PM, Zhou J, Criddle CS, Fields MW (2008) Bacterial community succession during in situ uranium bioremediation: spatial similarities along controlled flow paths. *ISME J* **doi:10.1038/ismej.2008.77**

Langille M, Hsiao W, Brinkman F (2008) Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* **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, Buchner A, Lai T, Steppi S, Jobb G, Forster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Konig A, Liss T 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, Chen IMA, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC (2007) IMG/M: a data management and analysis system for metagenomes. *Nucl Acids Res***:** 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 Meth* **4:** 63-72

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

Palumbo AV, Schryver JC, Fields MW, Bagwell CE, Zhou JZ, Yan T, Liu X, Brandt CC (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. *ES&T* **In Press**

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 Biology and Biochemistry* **34:** 285-296

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics* **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

Silver S, Phung LT (1996) Bacterial Heavy Metal Resistance: New Surprises. *Annual Review of Microbiology* **50:** 753-789

Sluis MK, Ensign SA (1997) Purification and characterization of acetone carboxylase from *Xanthobacter* strain Py2. *Proceedings of the National Academy of Sciences* **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, Johnson JR, Dykhuizen DE (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, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD et al (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Res* **12:** 1611-1618

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol Biol Evol***:** 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, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (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. *Journal of Biotechnology* **126:** 26-36

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

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences* **95:** 6578-6583

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. *Environmental Microbiology* **5:** 13-24

Yang Z (1997) PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood. *Computer Applications in BioSciences* **13:** 555-556

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia J-M, Soergel DA, Zhai Y et al (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biology* **5:** e16

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62:** 316-322

Acknowledgements

Authors' Contributions

All authors contributed intellectual input and assistance to this study and manuscript 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.

Figure Legends

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.

Figure 2. 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.

Figure 3. 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).

Figure 4. Putative acetone carboxylation operons of FW106. 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: R, Fis-type helix-turn-helix activator of acetoin/glycerol metabolism; α/β/γ, subunits of acetone carboxylase; aCat, acetyl-CoA acetyltransferase; aCs, acyl-CoA synthetase; PPII, Uncharacterized protein related to plant photosystem II stability/assembly factor; RND, predicted exporters of the RND superfamily; DH, alcohol dehydrogenase.

A)

α–Proteobacteria (0.2%)

Archaea and Eukarya (1.3%)

β–Proteobacteria (26.1%)

γ–Proteobacteria (38.0%)

Unassigned Bacteria (27.6%)

B)

Deltaproteobacteria - 1%

γ–Proteobacteria (11.7%)

Pseudomonadaceae (1.2%)

Xanthomonadaceae (7.2%)

Unassigned (64%)

Bradyrhizobiaceae (0.5%)

Burkholderiaceae (9.6%)

Comamonadaceae (1.8%)

Oxalobacteraceae (3.5%)

Rhodocyclaceae (1.4%)

β–Proteobacteria (18.2%)

α–Proteobacteria (2.8%)

A)

FW106
ucgw2_orm_C1041
(2007099749-64)

Polaromonas naphthalenivorens
CJ2 (639833942-38)

Ralstonia eutropha
JMP134 megaplasmid
(637689043-52)

Dechloromonas
aromatica **RCB**
(637679474-82)

Azoarcus **sp. EbN1 (637609911-23)**

Xanthobacter autotrophicus **Py2**
(639080683-8)

B)

FW106 ucgw2_orm_C1060
(2007101632-8)

Geobacter uraniumreducens **Rf4**
(639028619-30)

Polaromonas
naphthalenivorens
CJ2 (638955670-7)

Burkholderia **sp. 383**
(637766052-9)

Xanthobacter
autotrophicus **Py2**
(639079622-31)

I)   **Table 1. Binning of metagenomic contigs by PhyloPythia**

Domain	Phylum	Class	# Contigs	% Total Contigs	bp Sequence	% Total Sequence
Archaea	Crenarcheota	Thermoprotei	1	0.02	721	0.01
Archaea	Crenarcheota	Unassigned	1	0.02	845	0.01
Archaea	Euryarcheota	Unassigned	16	0.26	13720	0.14
Archaea	Unassigned	Unassigned	34	0.56	27790	0.29
Bacteria	Actinobacteria	Actinobacteria	9	0.15	19005	0.20
Bacteria	Actinobacteria	Unassigned	52	0.86	61095	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	11581	0.12
Bacteria	Firmicutes	Unassigned	9	0.15	8441	0.09
Bacteria	Proteobacteria	Alphaproteobacteria	9	0.15	15552	0.16
Bacteria	Proteobacteria	Betaproteobacteria	1659	27.29	2490010	26.06
Bacteria	Proteobacteria	Epsilonproteobacteria	2	0.03	1669	0.02
Bacteria	Proteobacteria	Gammaproteobacteria	84	1.38	3629419	38.00
Bacteria	Proteobacteria	Unassigned	450	7.40	552980	5.79
Bacteria	Unassigned	Unassigned	3471	57.10	2636705	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	78167	0.82

1   **Supplementary Materials**

Materials and Methods

A)   Site description and sampling.

The FW106 well from which groundwater was obtained is located in Area 3 of the Oak Ridge Field Research Center (FRC) near the Y-12 National Security Complex in Oak Ridge, TN (http://www.esd.ornl.gov/orifrc/).   This site lies in the path of a highly-contaminated groundwater plume originating from the original S-3 Waste Disposal Ponds.   Because of the proximity of the well to the S-3 ponds, uranium contamination levels are among the highest reported in the world.   Contaminants present in FW106 groundwater are listed in Table S1 (compared to pristine FRC groundwater from well FW301) and include high concentrations of nitric acid (pH ~3.7), radionucleotides (technetium-99 and uranium) and volatile organics.

To obtain sufficient biomass for sequencing, the FW106 well was extensively purged (several well volumes of water removed), ~1700 liters of groundwater was pumped from the matrix surrounding the screened area (~10-4m depth) using peristaltic pumps and was passed through sintered metal (T. J. Phelps, unpublished) (589 L) or 0.2 µm Supor$^{®}$ (Pall Corporation) filters (1126 L) to collect the biomass. Microbial cells were counted using LIVE/DEAD® stain (Invitrogen) and fluorescent microscopy according to the manufacturer's recommendations.   Cells were recovered from the filters by shaking and/or brief sonification and were pelleted by ultracentrifugation. The pH of the pellet was adjusted to 7.0 prior to DNA extraction (Zhou et al, 1996).   Direct bacterial counts were between 10^4 and 10^5 cells/ml and approximately 300 µg of DNA was obtained from the extracted groundwater.   Recovered DNA was treated with RNAse (Zhou et al, 1996) and sent to the DOE Joint Genome Institute (JGI) for constructing SSU rDNA clone and

three genome libraries for sequencing. The SSU rRNA gene sequences were initially processed within BioEdit (v. 7.0.5.3) (Hall, 2001) by aligning the sequences using ClustalW, trimming to a shared ~1,300 bp region, and generating a distance matrix using DNADist. DOTUR (Schloss & Handelsman, 2004) was then used to categorize the sequences into operational taxonomic units (OTUs), based on the distance matrix. 13 OTUs were defined at the 98% cutoff with the majority of OTUs mapping to *Rhodanobacter*-like γ-proteobacteria and the remainder to *Azoarcus*-like β-proteobacteria.

B)   Shotgun Sequencing, Assembly and Binning.

Metagenomic sequencing was conducted at JGI using random shotgun sequencing. ~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 (8kb) pMCL, 9.27 Mb large insert (40kb) pCCiFos). The sequencing reads (66220) were assembled into 421 contigs w/ >20 reads (2770 contigs total, ~8.3 Mb assembled DNA) using Phrap 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, alignment discrepancies beyond those expected for random sequencing errors were allowed if they were consistent with end-pairing constraints. A second assembly of the FW106 metagenome was conducted using Lucy (vector and quality trimming) (Chou & Holmes, 2001) and the Paracel Genome Assembler (pga) (Paracel, Pasadena, CA).   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 using the IMG pipeline. The pga assembly and

associated IMG annotation are available at the IMG/m database (Markowitz et al, 2007) and the FW106 read library has been deposited in GenBank (accession number pending).

Previous analyses and visual inspection of recovered biomass suggest a community dominated by a few closely-related γ-proteobacteria*l* populations. Consistent with this observation, metagenomic bins could not be defined by differences in GC content. Preliminary taxonomic bins were defined for the Phrap assembly using 16S rRNA gene sequences derived from the previously described OTU analysis and from 16S rRNA gene fragments identified directly from the metagenome. Contigs associated with these genes through scaffold assembly were in turn added to the appropriate bins. Once the preliminary bins were established, additional conserved anchor genes (23S rRNA, *recA*, *rpoB*, *gyrB*, *fusA* and *ileS*) were identified. The contigs containing these genes and contigs associated through scaffolding assigned to the appropriate bins. Finally, the remaining contigs were assigned by comparing the contig sequences to the GenBank nr database via BLAST where possible. The pga assembly was binned using PhyloPythia (McHardy et al, 2007). While the PhyloPythia binning still showed dominance by γ-proteobacteria followed by β-proteobacteria, the proportion of contigs sorting to β-proteobacterial and taxonomically unassigned bins was higher than that observed for the manual binning method. In both binning methods, the largest assembled contigs tended to bin to the γ-proteobacteria.

C)   Sequence Analysis and Metabolic Reconstruction.

Metabolic reconstruction of the community was performed automatically as part of the IMG annotation pipeline. Manual examination of the community metabolic network was used to fill in gaps in pathways (when possible) and to assign metabolic activities to specific bins. Abundance profiles for FW106 genes were determined by calculating odds

ratios for genes assigned to general COG and KEGG functional categories and to specific geochemically-relevant COGs. Odds ratios were calculated (i.e. for COG categories) as follows:

$$Odds\_Ratio = \frac{A/B}{C/D}$$

Where:

A = #FW106 genes assigned to a specific COG category

B = #FW106 genes assigned to all COG categories

C = # genes from all sequenced bacteria assigned to a specific COG category

D = # genes from all sequenced bacteria assigned to all COG categories

p-values were calculated for each odds ratio using one- and two-tailed Fisher's Exact Test to determine significant deviations from equilibrium (odds ratio = 1) (Rosner, 2005) . Values were plotted as ln (odds ratio) to better visualize positive and negative trends in the data. Abundance profiles of FW106 versus other metagenomes using the IMG abundance profiling tools showed similar results, suggesting little bias in the bacterial isolate genome database due to overrepresentation of pathogenic species.

D)   Evolutionary Analyses.

Evolutionary analyses were conducted using *ad hoc* Perl scripts (Deng *et al.*, unpublished results) except where explicitly stated. A nucleotide change was classified as a single nucleotide polymorphism (SNP) in a manner similar to that described in Wu et al., 2006:

(i) sequence quality score >20 in both the contig and read sequences; (ii) >4-fold coverage in the affected assembly column; (iii) there were differences among the other reads at that position; and (iv) the change was flanked by 3 invariable nucleotides on each side(Wu et al, 2006). Genes potentially under positive selection were identified by pairwise comparison of each FW106 CDS to the best BLAST (Altschul et al, 1997) match from GenBank (blastn, default options, local alignment >70% gene sequence) with the Nei-Gojobori method (Nei & Gojobori, 1986) implemented in BioPerl (Stajich et al, 2002) and the maximum likelihood methods implemented in PAML (Yang, 1997). ω values >1 indicate putative positive selection.

Laterally transferred genes, particularly those which have occurred recently and/or from a phylogenetically distinct donor, often display sequences characteristics from those of the host genome, including G+C content, dinucleotide frequencies, codon usage bias, and etc. Various statistical methods including iterative discriminant and hidden Markov model methods have been developed to identify regions based on these deviations from the genomic background characteristics. Potential problems with this method occur when alien regions display sequence characteristics indistinguishable from the genomic background (e.g. lateral transfer within similar populations) or when sufficient time has passed such that amelioration has occurred, i.e. the alien region has assumed the characteristics of the genomic background. Alternatively, phylogenetic methods involving comparison of gene family trees to well-defined species trees can reveal deviations in the gene family caused by lateral gene transfer. This method can detect more ancient transfers or alien genes that are indistinguishable by statistical methods, but may not be able to distinguish between closely-related orthologs and genes transferred between similar populations. For recent LGT events occurring in the FW106 population over the 50 years since introduction of contamination, amelioration of

transferred regions should be minimal and thus is not a factor in this analysis. A more likely scenario in a low-diversity environment such as FW106 groundwater is the transfer of genes between closely related populations. In this scenario, visual examination of the genome may be sufficient to identify potential LGT events between similar populations. In many cases, laterally transferred genes are associated with transposable elements or phage operon genes; though indistinguishable by statistical or phylogenetic means, the association of gene clusters with transposons and phage elements is a strong indication of lateral transfer. A final issue unique to metagenomic sequences is that the possibility of inaccurate assembly of the metagenomic reads can produce false instances of LGT. This problem is expected to be minimized in a low diversity community where assembly of long (kb-Mb range) contigs/scaffolds at high read depths is possible. Comparison of multiple assemblies of the FW106 metagenome as well as multiple PCR experiments using FW106 metagenomic DNA as template with primers designed from the assembled FW106 metagenome suggest the assembly does in fact accurately reflect the true genomic organization of the constituent species (see below).

Multiple complementary methods were employed in this study to detect putative genomic islands (GI) resulting from lateral transfer. First, an iterative discriminant analysis designed to detect deviations from background GC content, dinucleotide frequencies and codon usage was conducted for the major FW106 scaffolds (>100 Mb) (Tu & Ding, 2003). A second method, SIGI-HMM (as implemented in the Colombo program) employs a hidden Markov model (HMM)-based method for identifying regions of aberrant codon usage while minimizing false positives (Merkl, 2004; Waack et al, 2006). Third, selected GI's detected by these methods were verified by visual inspection of the GC content of these regions in Artemis (Berriman & Rutherford, 2003) (2.5 SD cutoff) and by synteny with phage genes, transposons and other known mobile elements. In

addition, protein phylogenies of the putatively selected genes of interest (i.e. geochemical resistance genes) were constructed and compared to 16S rRNA species trees that included FW106 16S rRNA genes. 16S rRNA gene-based phylogenetic trees were constructed by adding FW106 sequences to the GreenGenes (DeSantis et al, 2006) 16S dataset using the parsimony insertion function of ARB (DeSantis et al, 2006). Deviations from the 16S rRNA gene tree suggest possible LGT events and putative donor taxa. Emphasis in the text is given for LGT events (e.g. acetone carboxylase) predicted by multiple analysis methods.

E) Population Genetics Analysis

Degenerate PCR primers were constructed to probe the diversity of *narK* and *czcD* genes directly from FW106 metagenomic DNA. *Taq* DNA polymerase from Invitrogen was used for amplification and the PCR reaction mixture was: 10 μl of 10X PCR buffer minus Mg; 2 μl of 10 mM dNTP mixture; 3 μl of 50 mM $MgCl_2$; 5 μl of primer mix (10 μM each); 100 ng of template DNA from FW106; 0.5 μl of *Taq* DNA polymerase (5U/μl) and Nanopure water to adjust the final volume to 100 μl. PCR amplification was performed with a GeneAmp PCR system 9600 thermal cycler (Applied Biosystems) and subjected to a 5 min denaturation step at 95°C followed by 35 cycles at 94°C for 45 s, 55°C for 30 s and 72°C for 1.5 min. The reaction mixture was then held at 72°C for 15 min. Amplicons were gel purified by a Qiagen kit according to the manufacturer's manual and subsequently cloned into the easy sequencing TA-cloning kit. 96 colonies from each clone set were picked and sequenced by the laboratory of Dr. Bruce Roe (University of Oklahoma) as previously described (Elshahed et al, 2003).

Codon-based multisequence alignments of each sequence set were constructed using the ClustalW (Stajich et al, 2002) algorithm implemented in MEGA 3.1 (Kumar et al, 2004).

Single nucleotide indels resulting in frameshifts were assumed to be the result of sequencing error and were manually corrected within the alignments. Phylogenetic trees for each alignment were constructed in MEGA 3.1 (Kumar et al, 2004) using the neighbor-joining algorithm (Tamura-Nei, heterogeneous lineages, $\gamma=2$, 500 bootstrap replicates, complete gap deletion).   Sequences in clusters representing putative populations were extracted, realigned, and analyzed using the population genetics algorithms implemented in DnaSP 4.0 (Rozas & Rozas, 1999). Statistics calculated for each dataset include Tajima's D, Fu and Li's D*/F*/D/F, Fu's F_S, Fay and Wu's H and standard population genetics parameters (π, θ_w, S, ω, ZZ, etc.). The parameters were then analyzed to identify deviations from neutrality consistent with the effects of selection or demographic effects as described in the literature (Charlesworth et al, 1995; Kim, 2006; Nei & Kumar, 2000; Tajima, 1989).

Supplementary Tables

Table S1. Geochemistry of FRC sites FW301 (uncontaminated background) and FW106. Contaminant concentrations obtained from http://www.esd.ornl.gov/orifrc/.

	FW301	FW106
pH	~7	3.7
NO_3^- (mg/L)	1.5	2331
SO_4^{2-} (mg/L)	6.3	1997
Uranium (mg/L)	>0.0001	51
Technetium-99 (pCi/L)	-	3700
cis-1,2-Dichloroethene (µg/mL)	5	1216
1,2-Dichloroethene (µg/mL)	5	1153
Tetrachloroethene (µg/mL)	5	810
1-Butanol (µg/mL)	-	475
Acetone (µg/mL)	10	823
Benzoic Acid (µg/mL)	-	1400
Sodium (mg/L)	1.96	826
Chloride (mg/L)	1.125	465
Magnesium (mg/L)	2.58	45.7
Dissolved Oxygen (mg/L)	-	0.26

Table S2. Metagenome statistics of FW106

		% Total
DNA # bases		
Total bases	9554544	100.00%
DNA coding # bases	8076611	84.53%
DNA G+C # bases	6011119	63.20%
Scaffolds[a]	5698	-
Genes		
Total # Genes	12420	100.00%
RNA Genes	85	0.68%
rRNA Genes	7	0.06%
5S rRNA	1	0.01%
16S rRNA	3	0.02%
23S rRNA	3	0.02%
tRNA Genes	78	0.63%
Protein coding genes	12335	99.32%
Genes w/ function prediction	8689	69.96%
Genes w/o function prediction	3646	29.36%
Genes assigned to enzymes	1692	13.62%
Genes connected to KEGG pathways	1423	11.46%
Genes in COGs	7961	64.10%
Genes in Pfam	7410	59.66%

Table S3. Geochemical resistance genes identified in FW106 metagenome. A total of 444 genes (~4.7%) were identified.

Scaffold	ORF ID	IMG GOID	Gene Nam	Function	COG Grou	COG	Relevance
2760	6754	2005743887	-	Na+/H+ Antiporter	-	-	Acid Resistance
1029	1591	2005738626	ackA	Acetate Kinase	C	COG0282	Acid Resistance
1776	2898	2005739939	ackA	Acetate Kinase	C	COG0282	Acid Resistance
2764	7419	2005744555	ackA	Acetate Kinase	C	COG0282	Acid Resistance
2767	8282	2005745430	acs	Acetate--CoA Ligase	I	COG0365	Acid Resistance
2768	8451	2005745603	acs	Acetate--CoA Ligase	I	COG0365	Acid Resistance
2768	8458	2005745610	acs	Acetate--CoA Ligase	I	COG0365	Acid Resistance
2769	8643	2005745807	fdhA	Formate Dehydrogenase, Alpha Subunit	R	COG3383	Acid Resistance
2769	8644	2005745808	fdhA	Formate Dehydrogenase, Alpha Subunit	R	COG3383	Acid Resistance
696	972	2005737994	gshB	Glutathione Synthase	J	COG0189	Acid Resistance
2769	8709	2005745875	gshB	Glutathione Synthase	J	COG0189	Acid Resistance
1394	2241	2005739278	kefB	Kef-type K+ transport systems, membrane components	P	COG0475	Acid Resistance
2747	5594	2005742681	kefB	Kef-type K+ transport systems, membrane components	P	COG0475	Acid Resistance
2770	9016	2005746184	kefB	Kef-type K+ transport systems, membrane components	P	COG0475	Acid Resistance
8	11	2005737010	LDH	Lactate dehydrogenase and related dehydrogenases	CHR	COG1052	Acid Resistance
2759	6628	2005743743	LDH	Lactate dehydrogenase and related dehydrogenases	CHR	COG1052	Acid Resistance
1759	2867	2005739908	LDH	Malate/lactate dehydrogenases	C	COG0039	Acid Resistance
2747	5596	2005742683	LDH	Malate/lactate dehydrogenases	C	COG0039	Acid Resistance
2750	5723	2005742810	LDH	L-lactate dehydrogenase (FMN-dependent)	C	COG1304	Acid Resistance
2764	7445	2005744581	LDH	L-lactate dehydrogenase (FMN-dependent)	C	COG1304	Acid Resistance
2769	8830	2005745998	LDH	L-lactate dehydrogenase (FMN-dependent)	C	COG1304	Acid Resistance
2747	5593	2005742680	nhaD/arsB	Na+/H+ antiporter NhaD and related arsenite permeases	P	COG1055	Acid Resistance
2770	8931	2005746099	nhaP	NhaP-type Na+/H+ and K+/H+ antiporters	P	COG0025	Acid Resistance
1321	2121	2005739158	rpoS	RNAP Sigma-38 Factor	K	COG0568	Acid Resistance
2504	4483	2005741549	rpoS	RNAP Sigma-38 Factor	K	COG0568	Acid Resistance
1902	3160	2005740203	-	Heavy Metal Sensor Kinase	T	COG0642	Heavy Metals
2761	6975	2005744108	-	Heavy Metal Sensor Kinase	T	COG0642	Heavy Metals
2761	6976	2005744109	-	Heavy Metal Response Regulator	K	COG0745	Heavy Metals
379	500	2005737521	-	Heavy Metal-Resistance Transcriptional Regulator	K	COG0789	Heavy Metals
521	690	2005737712	-	Heavy Metal-Resistance Transcriptional Regulator	K	COG0789	Heavy Metals
2758	6361	2005743466	-	Heavy Metal-Binding Protein	R	COG3019	Heavy Metals
46	56	2005737055	-	Multicopper Oxidase	Q	COG2132	Heavy Metals
2515	4511	2005741577	-	Multicopper Oxidase	Q	COG2132	Heavy Metals
2558	4601	2005741669	-	Multicopper Oxidase	Q	COG2132	Heavy Metals
2765	7619	2005744757	-	Multicopper Oxidase	Q	COG2132	Heavy Metals
2766	8009	2005745155	-	Multicopper Oxidase	Q	COG2132	Heavy Metals
2770	9179	2005746349	-	Multicopper Oxidase	Q	COG2132	Heavy Metals
1819	2986	2005740027	ACR3	Arsenate Efflux	P	COG0798	Heavy Metals
2765	7628	2005744766	ACR3	Arsenate Efflux	P	COG0798	Heavy Metals
2764	7494	2005744630	acrB	Cation/multidrug efflux pump	V	COG0841	Heavy Metals
2764	7495	2005744631	acrB	Cation/multidrug efflux pump	V	COG0841	Heavy Metals
2765	7646	2005744784	acrB	Cation/multidrug efflux pump	V	COG0841	Heavy Metals
2766	7967	2005745111	acrB	Cation/multidrug efflux pump	V	COG0841	Heavy Metals
911	1379	2005738406	arsB	Na+/H+ antiporter NhaD and related arsenite permeases	P	COG1055	Heavy Metals
2764	7403	2005744539	arsB	Na+/H+ antiporter NhaD and related arsenite permeases	P	COG1055	Heavy Metals
2764	7404	2005744540	arsB	Na+/H+ antiporter NhaD and related arsenite permeases	P	COG1055	Heavy Metals
2766	7832	2005744972	arsB	Na+/H+ antiporter NhaD and related arsenite permeases	P	COG1055	Heavy Metals
1819	2987	2005740028	arsC	Arsenate Reductase (Glutaredoxin)	T	COG0394	Heavy Metals
2765	7627	2005744765	arsC	Arsenate Reductase (Glutaredoxin)	T	COG0394	Heavy Metals
2765	7629	2005744767	arsC	Arsenate Reductase (Glutaredoxin)	T	COG0394	Heavy Metals
1819	2988	2005740029	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
2254	3883	2005740935	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
2763	7213	2005744349	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
2765	7602	2005744740	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
2765	7624	2005744762	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
2765	7626	2005744764	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
2768	8523	2005745675	arsR	Arsenate Resistance Transcriptional Regulator	K	COG0640	Heavy Metals
940	1432	2005738461	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
1909	3174	2005740217	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
1974	3303	2005740346	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2152	3669	2005740717	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2348	4110	2005741170	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2622	4770	2005741838	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2680	4961	2005742031	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2711	5102	2005742172	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2753	5911	2005742998	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2753	5911	2005742998	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2758	6442	2005743549	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2760	6703	2005743836	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2761	6936	2005744069	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2765	7574	2005744712	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2765	7601	2005744739	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2765	7682	2005744820	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2765	7682	2005744820	cadA/zntA/	Heavy Metal-Translocating ATPase	P	COG2217	Heavy Metals
2348	4109	2005741169	cadR	Cd/Pb-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
847	1259	2005738281	chrA	Chromate Efflux	P	COG2059	Heavy Metals
2753	5878	2005742965	chrA	Chromate Efflux	P	COG2059	Heavy Metals
2766	8010	2005745156	chrA	Chromate Efflux	P	COG2059	Heavy Metals
2753	5875	2005742962	chrB	Chromate Efflux	S	COG4275	Heavy Metals
2753	5879	2005742966	chrB	Chromate Efflux	S	COG4275	Heavy Metals
69	89	2005737090	copA	Copper Resistance Protein	Q	COG2132	Heavy Metals
2661	4894	2005741962	copA	Copper Resistance Protein	Q	COG2132	Heavy Metals
2702	5048	2005742118	copA	Copper Resistance Protein	Q	COG2132	Heavy Metals
2736	5301	2005742373	copA	Copper Resistance Protein	Q	COG2132	Heavy Metals
2761	6933	2005744066	copA	Copper Resistance Protein	Q	COG2132	Heavy Metals
2766	8026	2005745172	copA	Copper Resistance Protein	Q	COG2132	Heavy Metals
2702	5047	2005742117	copB	Copper Resistance Protein	P	COG3667	Heavy Metals
2736	5299	2005742371	copB	Copper Resistance Protein	P	COG3667	Heavy Metals
2761	6934	2005744067	copB	Copper Resistance Protein	P	COG3667	Heavy Metals
2766	8025	2005745171	copB	Copper Resistance Protein	P	COG3667	Heavy Metals
2711	5101	2005742171	copG	Heavy Metal-Binding Protein	R	COG3019	Heavy Metals
2081	3525	2005740568	copR	Heavy Metal-Response Regulator	K	COG0745	Heavy Metals
2710	5098	2005742168	copR	Heavy Metal Response Regulator	K	COG0745	Heavy Metals
2766	8001	2005745147	copR	Heavy Metal Response Regulator	K	COG0745	Heavy Metals
2710	5099	2005742169	copS	Heavy Metal Sensor Kinase	T	COG0642	Heavy Metals
2766	8000	2005745146	copS	Heavy Metal Sensor Kinase	T	COG0642	Heavy Metals
2620	4765	2005741833	copZ	Heavy Metal Transport/Detoxification	P	COG2608	Heavy Metals

2753	5910	2005742997	copZ	Heavy Metal Transport/Detoxification	P	COG2608	Heavy Metals
2753	5912	2005742999	copZ	Heavy Metal Transport/Detoxification	P	COG2608	Heavy Metals
2765	7576	2005744714	copZ	Heavy Metal Transport/Detoxification	P	COG2608	Heavy Metals
2765	7578	2005744716	copZ	Heavy Metal Transport/Detoxification	P	COG2608	Heavy Metals
1352	2168	2005739205	corA	Mg/Co Transport	P	COG0598	Heavy Metals
1959	3274	2005740317	corA	Mg/Co Transport	P	COG0598	Heavy Metals
2760	6817	2005743950	corA	Mg/Co Transport	P	COG0598	Heavy Metals
2680	4960	2005742030	cueR	Cu(I)-responsive transcriptional regulator	K	COG0789	Heavy Metals
160	196	2005737198	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
655	904	2005737926	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
816	1195	2005738217	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1063	1657	2005738692	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1277	2039	2005739074	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1692	2751	2005739792	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1831	3012	2005740053	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1857	3065	2005740108	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1880	3114	2005740157	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1890	3136	2005740179	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
1932	3217	2005740260	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2192	3753	2005740801	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2194	3757	2005740805	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2297	3976	2005741028	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2310	4007	2005741061	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2310	4008	2005741062	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2636	4810	2005741878	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2707	5080	2005742150	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2728	5218	2005742290	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2734	5274	2005742346	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2734	5283	2005742355	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2761	6959	2005744092	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2761	6971	2005744104	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2762	7100	2005744233	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
123	153	2005737154	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
431	574	2005737595	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2695	5020	2005742090	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2711	5105	2005742175	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2728	5218	2005742290	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2765	7591	2005744729	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2765	7592	2005744730	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2766	8015	2005745161	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2770	9257	2005746427	czcA	Cd/Co/Zn Efflux	P	COG3696	Heavy Metals
2761	6972	2005744105	czcB	Cd/Co/Zn Efflux	-	-	Heavy Metals
2762	7101	2005744234	czcB	Cd/Co/Zn Efflux	-	-	Heavy Metals
118	146	2005737147	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
143	177	2005737178	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
160	195	2005737197	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
194	245	2005737247	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
1262	2012	2005739047	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
1800	2948	2005739989	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
1856	3063	2005740106	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2105	3573	2005740616	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2603	4719	2005741787	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2636	4811	2005741879	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2695	5022	2005742092	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2707	5079	2005742149	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2711	5104	2005742174	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2728	5217	2005742289	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2734	5275	2005742347	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2754	6007	2005743094	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2761	6958	2005744091	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2764	7496	2005744632	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2765	7593	2005744731	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2765	7645	2005744783	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2765	7679	2005744817	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2766	7968	2005745112	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2766	8016	2005745162	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2768	8526	2005745678	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2770	9258	2005746428	czcB	Membrane-Fusion Protein	M	COG0845	Heavy Metals
2707	5078	2005742148	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2711	5103	2005742173	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2728	5216	2005742288	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2734	351	2005737357	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2761	6957	2005744090	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2761	6973	2005744106	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2762	7102	2005744235	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2765	7594	2005744732	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2765	7644	2005744782	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2766	7969	2005745113	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2766	8017	2005745163	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
2770	9259	2005746429	czcC	Outer Membrane Efflux Protein	U	COG1538	Heavy Metals
273	360	2005737366	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2090	3544	2005740587	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2584	4673	2005741741	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2715	5128	2005742198	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2715	5130	2005742200	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2724	5182	2005742254	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2730	5238	2005742310	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2734	5269	2005742341	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2734	5271	2005742343	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2734	5272	2005742344	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2759	6614	2005743729	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2765	7587	2005744725	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2765	7589	2005744727	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2766	8006	2005745152	czcD	Cd/Co/Zn Efflux	P	COG1230	Heavy Metals
2766	7979	2005745125	emrE	Toxic Cation Efflux	P	COG2076	Heavy Metals
239	311	2005737317	eriC	Chloride Channel	P	COG0038	Heavy Metals
2041	3434	2005740477	eriC	Chloride Channel	P	COG0038	Heavy Metals
2756	6135	2005743226	eriC	Chloride Channel	P	COG0038	Heavy Metals

2760	6711	2005743844	eriC	Chloride Channel	P	COG0038	Heavy Metals
2760	6712	2005743845	eriC	Chloride Channel	P	COG0038	Heavy Metals
2762	7175	2005744310	eriC	Chloride Channel	P	COG0038	Heavy Metals
2767	8265	2005745413	eriC	Chloride Channel	P	COG0038	Heavy Metals
1895	3144	2005740187	fur	Ferric Iron Uptake Regulator	P	COG0735	Heavy Metals
2758	6402	2005743507	fur	Ferric Iron Uptake Regulator	P	COG0735	Heavy Metals
2759	6616	2005743731	fur	Ferric Iron Uptake Regulator	P	COG0735	Heavy Metals
2730	5237	2005742309	hmrR	Heavy Metal-Resistance Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7683	2005744821	hmrR	Heavy Metal-Resistance Transcriptional Regulator	K	COG0789	Heavy Metals
190	238	2005737240	irlS	Heavy Metal Sensor Kinase	T	COG0642	Heavy Metals
521	689	2005737711	merA	Mercuric Reductase	C	COG1249	Heavy Metals
831	1228	2005738250	merA	Mercuric Reductase	C	COG1249	Heavy Metals
1261	2011	2005739046	merA	Mercuric Reductase	C	COG1249	Heavy Metals
1392	2238	2005739275	merA	Mercuric Reductase	C	COG1249	Heavy Metals
1543	2502	2005739541	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2727	5212	2005742284	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2736	5313	2005742385	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2753	5905	2005742992	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2765	7631	2005744769	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2765	7666	2005744804	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2767	8109	2005745257	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2768	8418	2005745570	merA	Mercuric Reductase	C	COG1249	Heavy Metals
2765	7641	2005744779	merA	Mercuric Reductase	P	COG2608	Heavy Metals
2753	5903	2005742990	merC	Mercuric Transport Inner Membrane Protein	-	-	Heavy Metals
2766	8012	2005745158	merC	Mercuric Transport Inner Membrane Protein	-	-	Heavy Metals
2767	8108	2005745256	merC	Mercuric Transport Inner Membrane Protein	-	-	Heavy Metals
2753	5906	2005742993	merD	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7640	2005744778	merD	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2727	5209	2005742281	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
2734	5277	2005742349	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
2736	5316	2005742388	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
2753	5902	2005742989	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
2765	7642	2005744780	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
2765	7665	2005744803	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
2767	8107	2005745255	merP	Mercuric-Binding Periplasmic Protein	P	COG2608	Heavy Metals
323	427	2005737437	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
1129	1780	2005738815	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2680	4960	2005742030	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2719	5146	2005742216	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2734	5279	2005742351	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2753	5894	2005742981	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2753	5913	2005743000	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2760	6746	2005743879	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7572	2005744710	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7583	2005744721	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7652	2005744790	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7671	2005744809	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2765	7683	2005744821	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2766	7922	2005745064	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2766	8030	2005745176	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2767	8105	2005745253	merR	Mercuric-Responsive Transcriptional Regulator	K	COG0789	Heavy Metals
2734	5278	2005742350	merT	Mercuric Transport Protein	-	-	Heavy Metals
2753	5901	2005742988	merT	Mercuric Transport Protein	-	-	Heavy Metals
2765	7643	2005744781	merT	Mercuric Transport Protein	-	-	Heavy Metals
2765	7664	2005744802	merT	Mercuric Transport Protein	-	-	Heavy Metals
2767	8106	2005745254	merT	Mercuric Transport Protein	-	-	Heavy Metals
636	875	2005737897	mgtE	Mg/Co/Ni Transport	P	COG2239	Heavy Metals
2760	6683	2005743816	mgtE	Mg/Co/Ni Transport	P	COG2239	Heavy Metals
2761	6928	2005744061	MMT1	Co/Zn/Cd Transport	P	COG0053	Heavy Metals
1628	2646	2005739685	terC	Tellerium Resistance	P	COG0861	Heavy Metals
2758	6421	2005743526	terC	Tellerium Resistance	P	COG0861	Heavy Metals
135	167	2005737168	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
443	589	2005737610	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
504	674	2005737696	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
671	932	2005737954	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
941	1434	2005738463	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
941	1435	2005738464	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
1336	2145	2005739182	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2091	3545	2005740588	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2273	3920	2005740972	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2316	4026	2005741080	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2397	4222	2005741284	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2734	5276	2005742348	-	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2766	7939	2005745083	amtB	Ammonia Transport	P	COG0004	Nitrogen Metabolism
2768	8425	2005745577	amtB	Ammonia Transport	P	COG0004	Nitrogen Metabolism
677	940	2005737962	anr	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
2308	4001	2005741053	anr	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
2558	4600	2005741668	anr	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
2764	7507	2005744643	anr	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
2769	8897	2005746065	anr	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
1136	1792	2005738827	crp	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
467	616	2005737638	czcC	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
1412	2274	2005739311	dnrD	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
1412	2274	2005739311	dnrD	Anaerobic Regulator Protein	T	COG0664	Nitrogen Metabolism
2509	4498	2005741564	dnrO	Anaerobic Regulator Protein	-	-	Nitrogen Metabolism
1938	3228	2005740271	fdhA	Formate Dehydrogenase, Alpha Subunit	C	COG0243	Nitrogen Metabolism
920	1396	2005738423	gdhA	NAD-Specific Glutamate Dehydrogenase	E	COG2902	Nitrogen Metabolism
2119	3602	2005740645	gdhA	NAD-Specific Glutamate Dehydrogenase	E	COG2902	Nitrogen Metabolism
2371	4158	2005741218	gdhA	NAD-Specific Glutamate Dehydrogenase	E	COG2902	Nitrogen Metabolism
2754	6005	2005743092	gdhA	NAD-Specific Glutamate Dehydrogenase	E	COG2902	Nitrogen Metabolism
2754	6006	2005743093	gdhA	NAD-Specific Glutamate Dehydrogenase	E	COG2902	Nitrogen Metabolism
2755	6054	2005743143	gdhA	NAD-Specific Glutamate Dehydrogenase	E	COG2902	Nitrogen Metabolism
1069	1670	2005738705	glnA	Glutamine Synthetase	E	COG0174	Nitrogen Metabolism
2755	6069	2005743158	glnA	Glutamine Synthetase	E	COG0174	Nitrogen Metabolism
2768	8423	2005745575	glnA	Glutamine Synthetase	E	COG0174	Nitrogen Metabolism
112	137	2005737138	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism
2734	5273	2005742345	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism
2751	5798	2005742885	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism

2761	6960	2005744093	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism
2765	7590	2005744728	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism
2766	7940	2005745084	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism
2768	8426	2005745578	glnB	Nitrogen Regulatory Protein PII	E	COG0347	Nitrogen Metabolism
105	127	2005737128	glnD	Protein-PII Uridylyltransferase	O	COG2844	Nitrogen Metabolism
690	964	2005737986	glnD	Protein-PII Uridylyltransferase	O	COG2844	Nitrogen Metabolism
2766	7908	2005745050	glnD	Protein-PII Uridylyltransferase	O	COG2844	Nitrogen Metabolism
1064	1659	2005738694	gltB	Glutamate Synthase Domain I	E	COG0067	Nitrogen Metabolism
1992	3339	2005740382	gltB	Glutamate Synthase Domain I	E	COG0067	Nitrogen Metabolism
2675	4948	2005742016	gltB	Glutamate Synthase Domain I	E	COG0067	Nitrogen Metabolism
2768	8461	2005745613	gltB	Glutamate Synthase Domain II	E	COG0069	Nitrogen Metabolism
846	1257	2005738279	gltD	Glutamate synthase (NADPH)	E	COG0493	Nitrogen Metabolism
1306	2095	2005739132	gltD	Glutamate synthase (NADPH)	E	COG0493	Nitrogen Metabolism
2768	8462	2005745614	gltD	Glutamate synthase (NADPH)	E	COG0493	Nitrogen Metabolism
128	158	2005737159	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
242	315	2005737321	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
328	436	2005737446	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
1376	2212	2005739249	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
2483	4429	2005741495	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
2760	6808	2005743941	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
2769	8918	2005746086	narG	Nitrate Reductase Alpha Subunit	C	COG5013	Nitrogen Metabolism
2483	4428	2005741494	narH	Dissimilatory Nitrate Reductase Beta Subunit	C	COG1140	Nitrogen Metabolism
2760	6807	2005743940	narH	Dissimilatory Nitrate Reductase Beta Subunit	C	COG1140	Nitrogen Metabolism
2769	8917	2005746085	narH	Dissimilatory Nitrate Reductase Beta Subunit	C	COG1140	Nitrogen Metabolism
2760	6805	2005743938	narI	Dissimilatory Nitrate Reductase Gamma Subunit	C	COG2181	Nitrogen Metabolism
2769	8915	2005746083	narI	Dissimilatory Nitrate Reductase Gamma Subunit	C	COG2181	Nitrogen Metabolism
2760	6806	2005743939	narJ	Dissimilatory Nitrate Reductase Delta Subunit	C	COG2180	Nitrogen Metabolism
2769	8916	2005746084	narJ	Dissimilatory Nitrate Reductase Delta Subunit	C	COG2180	Nitrogen Metabolism
968	1487	2005738522	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
1320	2119	2005739156	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
1320	2120	2005739157	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
1376	2211	2005739248	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
1477	2388	2005739425	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
1477	2389	2005739426	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
1492	2412	2005739449	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
2042	3436	2005740479	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
2769	8906	2005746074	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
2769	8912	2005746080	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
2769	8913	2005746081	narK	Nitrate/Nitrite Antiporter	P	COG2223	Nitrogen Metabolism
2326	4056	2005741110	narX/narQ	Nitrate/Nitrite Sensor Protein	T	COG3850	Nitrogen Metabolism
2767	8088	2005745234	nasA	Assimilatory Nitrate Reductase Alpha Subunit	C	COG0243	Nitrogen Metabolism
800	1161	2005738183	nasA	Assimilatory Nitrate Reductase Alpha Subunit	R	COG3383	Nitrogen Metabolism
800	1159	2005738181	nirB	NAD(P)H Nitrite Reductase	C	COG1251	Nitrogen Metabolism
1495	2419	2005739456	nirB	NAD(P)H Nitrite Reductase	C	COG1251	Nitrogen Metabolism
1830	3009	2005740050	nirB	NAD(P)H Nitrite Reductase	C	COG1251	Nitrogen Metabolism
2767	8089	2005745235	nirB	NAD(P)H Nitrite Reductase	C	COG1251	Nitrogen Metabolism
2767	8091	2005745237	nirB	NAD(P)H Nitrite Reductase	C	COG1251	Nitrogen Metabolism
800	1160	2005738182	nirD	NAD(P)H Nitrite Reductase (Ferredoxin) Small Subunit	R	COG2146	Nitrogen Metabolism
2766	7949	2005745093	nirD	NAD(P)H Nitrite Reductase (Ferredoxin) Small Subunit	R	COG2146	Nitrogen Metabolism
2767	8090	2005745236	nirD	NAD(P)H Nitrite Reductase (Ferredoxin) Small Subunit	R	COG2146	Nitrogen Metabolism
2770	9111	2005746281	nirD	NAD(P)H Nitrite Reductase (Ferredoxin) Small Subunit	R	COG2146	Nitrogen Metabolism
763	1091	2005738113	nirK	Nitrite Reductase (Copper), NO-Forming	Q	COG2132	Nitrogen Metabolism
2769	8904	2005746072	nirK	Nitrite Reductase (Copper), NO-Forming	Q	COG2132	Nitrogen Metabolism
544	721	2005737743	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
777	1117	2005738139	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2137	3636	2005740684	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2750	5392	2005742779	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2765	7680	2005744818	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2766	7985	2005745131	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2768	8511	2005745663	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2769	8881	2005746049	nodT	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
291	387	2005737393	norB	Nitric Oxide Reductase Large Subunit	P	COG3256	Nitrogen Metabolism
823	1209	2005738231	norB	Nitric Oxide Reductase Large Subunit	P	COG3256	Nitrogen Metabolism
1319	2118	2005739155	norB	Nitric Oxide Reductase Large Subunit	P	COG3256	Nitrogen Metabolism
1453	2341	2005739378	norB	Nitric Oxide Reductase Large Subunit	P	COG3256	Nitrogen Metabolism
2767	8272	2005745420	norB	Nitric Oxide Reductase Large Subunit	P	COG3256	Nitrogen Metabolism
2769	8891	2005746059	norB	Nitric Oxide Reductase Large Subunit	P	COG3256	Nitrogen Metabolism
1962	3281	2005740324	nosD	Nitrous Oxide Accessory Protein	P	COG3420	Nitrogen Metabolism
2442	4331	2005741393	nosD	Nitrous Oxide Accessory Protein	P	COG3420	Nitrogen Metabolism
2764	7391	2005744527	nosD	Nitrous Oxide Accessory Protein	P	COG3420	Nitrogen Metabolism
2764	7390	2005744526	nosF	ABC Transport ATP-Binding Protein	V	COG1131	Nitrogen Metabolism
651	899	2005737921	nosL	Nitrous Oxide Reduction	C	COG4314	Nitrogen Metabolism
2764	7388	2005744524	nosL	Nitrous Oxide Reduction	C	COG4314	Nitrogen Metabolism
2764	7393	2005744529	nosR	Nitrous Oxide Reductase Transcriptional Regulator	K	COG3901	Nitrogen Metabolism
2761	6917	2005744050	nosY	Nitrous Oxide Reductase Transport	R	COG1277	Nitrogen Metabolism
2764	7389	2005744525	nosY	Nitrous Oxide Reductase Transport	R	COG1277	Nitrogen Metabolism
2770	8965	2005746133	nosY	Nitrous Oxide Reductase Transport	R	COG1277	Nitrogen Metabolism
2071	3504	2005740547	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2442	3504	2005740547	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2442	4332	2005741394	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2442	4332	2005741394	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2753	5877	2005742964	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2764	7392	2005744528	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2764	7392	2005744528	nosZ	Nitrous Oxide Reductase Precursor	C	COG4263	Nitrogen Metabolism
2768	8422	2005745574	ntrB	Signal Transduction Histidine Kinase, Nitrogen-Specific	T	COG3852	Nitrogen Metabolism
856	1276	2005738299	ntrC	Nitrogen Regulation Protein NR(I)	T	COG2204	Nitrogen Metabolism
2155	3676	2005740724	ntrC	Nitrogen Regulation Protein NR(I)	T	COG2204	Nitrogen Metabolism
2768	8421	2005745573	ntrC	Nitrogen Regulation Protein NR(I)	T	COG2204	Nitrogen Metabolism
1299	2082	2005739117	tolC	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2762	7190	2005744325	tolC	Outer Membrane Efflux Protein	R	COG1277	Nitrogen Metabolism
2754	5974	2005743061	acaB	Acetyl-CoA Acetyltransferase	I	COG0183	Organic Solvents
1617	2625	2005739664	acoR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2769	8609	2005745771	acoR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2769	8646	2005745810	acoR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
1576	2556	2005739595	acxA	Acetone Carboxylase, Alpha Subunit	Q	COG0146	Organic Solvents
2750	5738	2005742825	acxA	Acetone Carboxylase, Alpha Subunit	Q	COG0146	Organic Solvents
2754	5976	2005743063	acxA	Acetone Carboxylase, Alpha Subunit	Q	COG0146	Organic Solvents
2755	6092	2005743181	acxA	Acetone Carboxylase, Alpha Subunit	Q	COG0146	Organic Solvents

1107	1744	2005738779 acxB	Acetone Carboxylase, Beta Subunit	Q	COG0145	Organic Solvents
2750	5735	2005742822 acxB	Acetone Carboxylase, Beta Subunit	Q	COG0145	Organic Solvents
2750	5737	2005742824 acxB	Acetone Carboxylase, Beta Subunit	Q	COG0145	Organic Solvents
2754	5977	2005743064 acxB	Acetone Carboxylase, Beta Subunit	Q	COG0145	Organic Solvents
2755	6091	2005743180 acxB	Acetone Carboxylase, Beta Subunit	Q	COG0145	Organic Solvents
2750	5736	2005742823 acxC	Acetone Carboxylase, Gamma Subunit	Q	COG4647	Organic Solvents
2754	5975	2005743062 acxC	Acetone Carboxylase, Gamma Subunit	Q	COG4647	Organic Solvents
2076	3514	2005740557 acxR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2154	3675	2005740723 acxR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2551	4583	2005741649 acxR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2747	5572	2005742657 acxR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2750	5719	2005742806 acxR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2754	5980	2005743067 acxR	Transcriptional Activator of Acetoin/Glycerol Metabolism	K	COG3284	Organic Solvents
2765	7636	2005744774 adc	Acetoacetate Decarboxylase	Q	COG4689	Organic Solvents
2765	7650	2005744788 adc	Acetoacetate Decarboxylase	Q	COG4689	Organic Solvents
732	1039	2005738061 adhC	Bifunctional Glutathione-Dependent Formaldehyde Dehydrogenase and Alcohol Dehydrogenase III	C	COG1062	Organic Solvents
2754	5973	2005743060 caiC	Acyl-CoA Synthetase	Q	COG0318	Organic Solvents
2413	4264	2005741326 dehH	Haloacetate Dehalogenase	R	COG0596	Organic Solvents
1129	1779	2005738814 dhaA	Haloalkane Dehydrogenase	R	COG0596	Organic Solvents
2767	8084	2005745230 fghA	S-Formylglutathione Hydrolase	R	COG0627	Organic Solvents
2769	8660	2005745824 moxJ	Methanol Oxidation Protein	R	COG0666	Organic Solvents
1112	1752	2005738787 moxR	MoxR-like ATPases	R	COG0714	Organic Solvents
1832	3014	2005740055 moxR	MoxR-like ATPases	R	COG0714	Organic Solvents
2215	3807	2005740857 moxR	MoxR-like ATPases	R	COG0714	Organic Solvents
2759	6572	2005743685 moxR	MoxR-like ATPases	R	COG0714	Organic Solvents
2770	9127	2005746297 moxR	MoxR-like ATPases	R	COG0714	Organic Solvents
1990	3337	2005740380 mxaF	Methanol Dehydrogenase, Heavy Chain	G	COG4993	Organic Solvents
2769	8661	2005745825 mxaF	Methanol Dehydrogenase, Heavy Chain	G	COG4993	Organic Solvents
966	1481	2005738516 apsK/cycC	APS Sulfurylase	P	COG0529	Sulfate
1865	3078	2005740121 cysA	ABC Sulfate Transporter ATP-Binding Protein	E	COG1118	Sulfate
555	736	2005737758 cysD	Sulfate Adenylyltransferase, Small Subunit	H	COG0175	Sulfate
2473	4403	2005741467 cysD	Sulfate Adenylyltransferase, Small Subunit	H	COG0175	Sulfate
2473	4404	2005741468 cysD	Sulfate Adenylyltransferase, Small Subunit	H	COG0175	Sulfate
2766	7850	2005744990 cysD	Sulfate Adenylyltransferase, Small Subunit	H	COG0175	Sulfate
2473	4405	2005741469 cysH	5' Adenylylsulfate APS Reductase	H	COG0175	Sulfate
2744	5479	2005742559 cysH	5' Adenylylsulfate APS Reductase	H	COG0175	Sulfate
2092	3548	2005740591 cysI	Sulfite Reductase (ferredoxin)	P	COG0155	Sulfate
2744	5480	2005742560 cysI	Sulfite Reductase (ferredoxin)	P	COG0155	Sulfate
1217	1929	2005738964 cysJ	Sulfite Reductase (flavoprotein)	P	COG0369	Sulfate
1417	2282	2005739319 cysJ	Sulfite Reductase (flavoprotein)	P	COG0369	Sulfate
2744	5481	2005742561 cysJ	Sulfite Reductase (flavoprotein)	P	COG0369	Sulfate
2767	8278	2005745426 cysJ	Sulfite Reductase (flavoprotein)	P	COG0369	Sulfate
234	303	2005737309 cysK	Cysteine Synthase	E	COG0031	Sulfate
1636	2661	2005739700 cysK	Cysteine Synthase	E	COG0031	Sulfate
2463	4378	2005741442 cysK	Cysteine Synthase	E	COG0031	Sulfate
2757	6241	2005743344 cysK	Cysteine Synthase	E	COG0031	Sulfate
2759	6601	2005743716 cysK	Cysteine Synthase	E	COG0031	Sulfate
2761	6998	2005744131 cysK	Cysteine Synthase	E	COG0031	Sulfate
2766	7823	2005744963 cysK	Cysteine Synthase	E	COG0031	Sulfate
965	1480	2005738515 cysN	Sulfate Adenylyltransferase, Large Subunit	P	COG2895	Sulfate
2473	4402	2005741466 cysN	Sulfate Adenylyltransferase, Large Subunit	P	COG2895	Sulfate
2766	7851	2005744991 cysN	Sulfate Adenylyltransferase, Large Subunit	P	COG2895	Sulfate
1865	3080	2005740123 cysU	ABC Sulfate Transporter Permease Protein	O	COG0555	Sulfate
1865	3079	2005740122 cysW	ABC Sulfate Transporter Permease Protein	P	COG4208	Sulfate

Table S4. Population genetics analysis of sequenced FW106 genes. 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.

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
# 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
$\pi_S{}^e$	0.00026	0.00168	0.05985	0.01539	0.01663
$\pi_a{}^f$	0.00020	0.00236	0.00819	0.00336	0.00262
$\pi_a/\pi_S{}^g$	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
$\theta_w{}^k$	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 & Li's D*	-2.69*	-5.53**	-2.50*	-4.06*	-1.70 (NS)

Fu & 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 & Li's D^m	-2.36 (NS)	-4.38*	-1.81 (NS)	-3.13**	-0.60 (NS)
Fu & Li's F^m	-2.47*	-4.43**	-2.26 (NS)	-3.03**	-0.66 (NS)
Fay & Wu's H^m	0.13 (NS)	-10.41***	-100.07*	-7.17 (NS)	-5.49*

[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] **# segregating sites**

[d] **# synonymous and nonsynonymous substitutions**

[e] **Synonymous nucleotide diversity**

[f] **Nonsynonymous nucleotide diversity**

[g] **Intraspecific diversity**

[h] **Interspecific divergence**

[i] **Average # 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 utilizing outlier sequences**

Table S5. Putative alien (laterally transferred) genes identified by SIGI-HMM.

Scaffold	Contig	IMG Gene	Scaffold Length	COG Category	Function
1	2747	5559	198606	-	phage-related conserved hypothetical protein
1	2747	5560	198606	-	Putative gene predicted by FgeneshB
1	2747	5563	198606	-	Putative gene predicted by FgeneshB
1	2758	6361	198606	-	Putative gene predicted by FgeneshB
1	2758	6362	198606	-	Putative gene predicted by FgeneshB
1	2758	6363	198606	-	hypothetical protein
1	2758	6364	198606	C	cytochrome c, class I
1	2758	6365	198606	C	cytochrome c oxidase
1	2758	6366	198606	-	Cytochrome C
1	2758	6367	198606	-	putative cytochrome c1 precursor protein
1	2758	6377	198606	-	Putative gene predicted by FgeneshB
1	2758	6378	198606	R	Secretion chaperone CsaA
1	2758	6434	198606	O	thioredoxin reductase
1	2758	6462	198606	K	helix-turn-helix motif
1	2758	6463	198606	-	Putative gene predicted by FgeneshB
1	2758	6464	198606	L	DNA adenine methylase
1	2758	6465	198606	-	Putative gene predicted by FgeneshB
220	2748	5621	371213	-	Putative gene predicted by FgeneshB
220	2748	5624	371213	G	ketose-bisphosphate aldolases
220	2748	5625	371213	G	Myo-inositol catabolism IolB region
220	2748	5626	371213	G	PfkB
220	2748	5627	371213	E	thiamine pyrophosphate enzyme, central region
220	2748	5630	371213	-	Putative gene predicted by FgeneshB
220	2748	5631	371213	K	periplasmic binding protein/LacI transcriptional regulator
220	2748	5633	371213	R	oxidoreductase-like
220	2748	5634	371213	G	Xylose isomerase-like TIM barrel
220	2748	5635	371213	R	oxidoreductase-like
220	2748	5636	371213	-	transposase
220	2752	5810	371213	-	Putative gene predicted by FgeneshB
220	2752	5811	371213	-	hypothetical protein
220	2752	5812	371213	-	Putative gene predicted by FgeneshB
220	2752	5813	371213	-	hypothetical protein
220	2752	5814	371213	-	hypothetical protein
220	2752	5815	371213	-	hypothetical protein
220	2752	5816	371213	-	hypothetical protein
220	2752	5817	371213	-	hypothetical protein
220	2752	5818	371213	-	hypothetical protein
220	2752	5819	371213	-	Putative gene predicted by FgeneshB
220	2752	5820	371213	-	Putative gene predicted by FgeneshB
220	2752	5821	371213	-	Putative gene predicted by FgeneshB
220	2752	5822	371213	-	Putative gene predicted by FgeneshB
220	2752	5823	371213	-	Putative gene predicted by FgeneshB
220	2752	5824	371213	-	Putative gene predicted by FgeneshB
220	2752	5825	371213	-	hypothetical protein
220	2752	5826	371213	-	Putative gene predicted by FgeneshB
220	2752	5827	371213	-	hypothetical protein
220	2752	5857	371213	G	PfkB
220	2752	5858	371213	G	RbsD or FucU transport
220	2752	5861	371213	G	periplasmic binding protein/LacI transcriptional regulator
220	2752	5862	371213	F	deoxyribose-phosphate aldolase
220	2752	5863	371213	C	Betaine-aldehyde dehydrogenase
220	2752	5864	371213	C	aldehyde dehydrogenase
220	2752	5865	371213	-	Putative gene predicted by FgeneshB
220	2752	5866	371213	R	Oxidoreductase-like
220	2768	8376	371213	-	Putative gene predicted by FgeneshB
220	2768	8404	371213	-	Putative gene predicted by FgeneshB
220	2768	8405	371213	-	conserved hypothetical protein
220	2768	8406	371213	-	Putative gene predicted by FgeneshB
220	2768	8407	371213	-	Putative gene predicted by FgeneshB
220	2768	8408	371213	-	hypothetical protein
220	2768	8409	371213	-	conserved hypothetical protein
220	2768	8410	371213	-	conserved hypothetical protein
220	2768	8411	371213	-	Putative gene predicted by FgeneshB
220	2768	8414	371213	V	restriction modification system DNA specificity domain
220	2768	8445	371213		Type I restriction modification system methyltransfe
220	2768	8446	371213	V	Restriction endonuclease S subunits-like
220	2768	8501	371213	-	Cold shock proteins
220	2768	8502	371213	L	exodeoxyribonuclease III (xth)
220	2768	8503	371213	R	YgfB and YecA
220	2768	8551	371213	-	Putative gene predicted by FgeneshB
220	2768	8599	371213	-	Putative gene predicted by FgeneshB
215	2732	5253	911819	M	Rhamnan synthesis F
215	2732	5254	911819	-	Putative gene predicted by FgeneshB
215	2732	5255	911819	M	glycosyl transferase, family 2
215	2737	5325	911819	L	hypothetical protein
215	2737	5326	911819	-	hypothetical protein
215	2737	5327	911819	M	glycosyl transferase, family 2
215	2737	5328	911819	-	Putative gene predicted by FgeneshB
215	2737	5329	911819	G	Xylose isomerase-like TIM barrel
215	2737	5330	911819	G	hypothetical protein
215	2737	5331	911819	R	FAD dependent oxidoreductase
215	2737	5332	911819	M	glycosyl transferase, family 2
215	2737	5333	911819	-	Putative gene predicted by FgeneshB
215	2737	5335	911819	-	Putative gene predicted by FgeneshB
215	2737	5336	911819	-	Putative gene predicted by FgeneshB
215	2750	5743	911819	-	Putative gene predicted by FgeneshB
215	2750	5744	911819	R	conserved hypothetical protein
215	2750	5745	911819	R	exporters of the RND superfamily
215	2750	5747	911819	R	exporters of the RND superfamily
215	2750	5748	911819	-	putative regulatory protein, LysR

215	2750	5749	911819	-	Putative gene predicted by FgeneshB
215	2756	6120	911819	K	transcription elongation factor GreA/GreB region
215	2756	6122	911819	R	exporters of the RND superfamily
215	2756	6123	911819	-	putative regulatory protein, LysR
215	2756	6124	911819	-	conserved hypothetical protein
215	2756	6153	911819	-	Putative gene predicted by FgeneshB
215	2761	6903	911819	-	Putative gene predicted by FgeneshB
215	2761	6904	911819	M	Cyclopropane-fatty-acyl-phospholipid synthase
215	2761	6931	911819	-	hypothetical protein
215	2761	6932	911819	-	Putative gene predicted by FgeneshB
215	2761	6949	911819	-	Putative gene predicted by FgeneshB
215	2761	6951	911819	-	hypothetical protein
215	2761	6952	911819	C	cytochrome c, class I
215	2761	6953	911819	-	hypothetical protein
215	2761	6954	911819	-	Putative WD-repeat containing protein
215	2761	6955	911819		Cytochrome c5
215	2761	6956	911819	H	protoporphyrinogen oxidase
215	2761	6961	911819	-	Putative gene predicted by FgeneshB
215	2761	6962	911819	-	hypothetical protein
215	2761	6963	911819	O	cytochrome c biogenesis protein, transmembrane region
215	2761	6985	911819	S	conserved hypothetical protein
215	2761	6986	911819	L	Resolvase-like
215	2761	6987	911819	-	Putative gene predicted by FgeneshB
215	2761	6988	911819	-	hypothetical protein
215	2761	6989	911819	L	ATP-dependent exoDNAse (exonuclease V) alpha subunit - helicase superfamily I member-like
215	2761	6990	911819	-	Putative gene predicted by FgeneshB
215	2761	6991	911819	-	Replication protein A
215	2761	7030	911819	-	hypothetical protein
215	2761	7031	911819	M	glycosyl transferase, group 1
215	2764	7446	911819	-	Putative gene predicted by FgeneshB
215	2764	7447	911819	Q	DSBA oxidoreductase
215	2764	7476	911819	-	hypothetical protein
215	2764	7477	911819	-	hypothetical protein
215	2764	7478	911819	-	conserved hypothetical protein
215	2764	7479	911819	-	hypothetical protein
215	2764	7480	911819	-	Putative gene predicted by FgeneshB
215	2764	7481	911819	-	Putative gene predicted by FgeneshB
215	2764	7482	911819	-	Similarities with unknown protein
215	2764	7483	911819	-	Putative gene predicted by FgeneshB
215	2764	7487	911819	O	AAA ATPase, central region
215	2764	7488	911819	-	putative modification methylase
215	2764	7489	911819	R	conserved hypothetical protein
215	2764	7490	911819	-	Bll0881 protein
215	2764	7491	911819	-	hypothetical protein
215	2764	7539	911819	-	Putative gene predicted by FgeneshB
215	2764	7540	911819	-	hypothetical protein
215	2764	7541	911819	-	conserved hypothetical protein
215	2764	7542	911819	-	hypothetical protein
215	2764	7543	911819	-	Putative gene predicted by FgeneshB
215	2764	7544	911819	-	hypothetical protein
215	2764	7545	911819	-	hypothetical protein
215	2764	7550	911819	-	Putative gene predicted by FgeneshB
215	2764	7551	911819	-	hypothetical protein
215	2769	8642	911819	-	Putative gene predicted by FgeneshB
215	2769	8794	911819	-	Putative gene predicted by FgeneshB
215	2769	8911	911819	-	hypothetical protein
219	2742	5420	2435494	-	peptidase M50
219	2742	5423	2435494	-	Putative gene predicted by FgeneshB
219	2742	5424	2435494	-	hypothetical protein
219	2742	5425	2435494	S	protein of unknown function DUF181
219	2742	5426	2435494	-	Nitroreductase
219	2742	5427	2435494	V	ABC transporter related
219	2742	5428	2435494	-	Putative gene predicted by FgeneshB
219	2744	5455	2435494	-	Putative gene predicted by FgeneshB
219	2744	5456	2435494	-	Putative gene predicted by FgeneshB
219	2744	5457	2435494	-	hypothetical protein
219	2744	5458	2435494	-	Putative gene predicted by FgeneshB
219	2744	5459	2435494	-	hypothetical protein
219	2744	5460	2435494	-	hypothetical protein
219	2744	5461	2435494	-	Putative gene predicted by FgeneshB
219	2744	5462	2435494	-	Putative gene predicted by FgeneshB
219	2744	5463	2435494	-	hypothetical protein
219	2744	5464	2435494	-	hypothetical protein
219	2744	5465	2435494	L	phage integrase
219	2744	5470	2435494	-	Putative gene predicted by FgeneshB
219	2744	5471	2435494	-	Uncharacterized protein conserved in bacteria
219	2744	5472	2435494	S	conserved hypothetical protein
219	2744	5473	2435494	-	Putative gene predicted by FgeneshB
219	2744	5495	2435494	-	hypothetical protein
219	2744	5496	2435494	-	Putative gene predicted by FgeneshB
219	2744	5497	2435494	-	hypothetical protein
219	2744	5498	2435494	-	hypothetical protein
219	2744	5499	2435494	-	hypothetical protein
219	2744	5500	2435494	-	phage-related conserved hypothetical protein
219	2749	5640	2435494		Predicted transcriptional regulators
219	2749	5641	2435494	K	helix-turn-helix motif
219	2749	5667	2435494	-	Putative gene predicted by FgeneshB
219	2754	5970	2435494	-	Putative gene predicted by FgeneshB
219	2754	5971	2435494	-	Putative gene predicted by FgeneshB
219	2754	5974	2435494	-	Acetyl CoA acetyltransferase
219	2754	5975	2435494	Q	acetone carboxylase gamma subunit

219	2754	6031	2435494	R	modification methylase
219	2754	6032	2435494	V	protein of unknown function DUF450
219	2755	6081	2435494	-	Putative gene predicted by FgeneshB
219	2755	6082	2435494	E	hydantoin racemase
219	2755	6083	2435494	R	hypothetical protein
219	2755	6084	2435494	C	L-carnitine dehydratase/bile acid-inducible protein F
219	2755	6085	2435494	E	pyruvate carboxyltransferase
219	2755	6086	2435494	R	conserved hypothetical protein
219	2755	6087	2435494	R	exporters of the RND superfamily
219	2755	6088	2435494	-	putative regulatory protein, LysR
219	2755	6093	2435494	Q	isochorismatase hydrolase
219	2755	6094	2435494	R	Alcohol dehydrogenase, zinc-binding
219	2755	6332	2435494	-	Putative gene predicted by FgeneshB
219	2757	6234	2435494	R	SMC protein-like
219	2757	6235	2435494	-	conserved hypothetical protein
219	2757	6254	2435494	S	protein of unknown function DUF125, transmembrane
219	2757	6333	2435494	-	hypothetical protein
219	2757	6334	2435494	L	putative ISXo8 transposase
219	2757	6335	2435494	-	conserved hypothetical protein
219	2757	6336	2435494	-	putative regulatory protein, LysR
219	2759	6472	2435494	-	Hypothetical protein
219	2759	6473	2435494	-	Putative gene predicted by FgeneshB
219	2759	6474	2435494	-	Putative gene predicted by FgeneshB
219	2759	6501	2435494	-	hypothetical protein
219	2759	6502	2435494	R	beta-lactamase-like
219	2759	6503	2435494	R	Alcohol dehydrogenase, zinc-binding
219	2759	6504	2435494	K	LysR, substrate-binding
219	2759	6505	2435494	-	Putative gene predicted by FgeneshB
219	2759	6506	2435494	S	Carboxymuconolactone decarboxylase
219	2759	6653	2435494	-	Putative gene predicted by FgeneshB
219	2759	6654	2435494	-	Putative gene predicted by FgeneshB
219	2759	6655	2435494	-	Putative gene predicted by FgeneshB
219	2759	6656	2435494	-	hypothetical protein
219	2759	6657	2435494	-	Putative gene predicted by FgeneshB
219	2759	6771	2435494	-	Putative gene predicted by FgeneshB
219	2762	7035	2435494	-	hypothetical protein
219	2762	7036	2435494	-	hypothetical protein
219	2762	7037	2435494	D	hypothetical protein
219	2762	7038	2435494	-	hypothetical protein
219	2762	7077	2435494	-	Putative gene predicted by FgeneshB
219	2762	7078	2435494	-	Putative gene predicted by FgeneshB
219	2762	7079	2435494	-	Putative gene predicted by FgeneshB
219	2762	7080	2435494	-	Putative gene predicted by FgeneshB
219	2762	7084	2435494	-	Putative gene predicted by FgeneshB
219	2762	7085	2435494	-	hypothetical protein
219	2762	7086	2435494	-	hypothetical protein
219	2762	7121	2435494	-	Putative gene predicted by FgeneshB
219	2762	7208	2435494	-	hypothetical protein
219	2763	7209	2435494	-	Putative gene predicted by FgeneshB
219	2763	7210	2435494	-	hypothetical protein
219	2763	7220	2435494	-	Putative gene predicted by FgeneshB
219	2763	7246	2435494	-	Putative gene predicted by FgeneshB
219	2763	7253	2435494	-	Putative gene predicted by FgeneshB
219	2763	7288	2435494	-	hypothetical protein
219	2763	7289	2435494	-	hypothetical protein
219	2765	7567	2435494	-	Putative gene predicted by FgeneshB
219	2765	7596	2435494	-	Putative gene predicted by FgeneshB
219	2765	7597	2435494	-	Putative gene predicted by FgeneshB
219	2765	7598	2435494	S	conserved hypothetical protein
219	2765	7599	2435494	-	Putative gene predicted by FgeneshB
219	2765	7600	2435494	S	hypothetical protein
219	2765	7604	2435494	-	Putative gene predicted by FgeneshB
219	2765	7605	2435494	-	Transposase and inactivated derivatives
219	2765	7606	2435494	-	ISPsy26, transposase orfB
219	2765	7607	2435494	-	conserved hypothetical protein
219	2765	7672	2435494	S	protein of unknown function DUF6, transmembrane
219	2765	7673	2435494	Q	DSBA oxidoreductase
219	2765	7674	2435494	S	Alkylhydroperoxidase AhpD core
219	2765	7675	2435494	O	OsmC-like protein
219	2765	7676	2435494	K	regulatory protein, TetR
219	2765	7763	2435494	-	hypothetical protein
219	2765	7764	2435494	-	Putative gene predicted by FgeneshB
219	2765	7765	2435494	-	Putative gene predicted by FgeneshB
219	2766	7822	2435494	-	Putative gene predicted by FgeneshB
219	2766	7849	2435494	-	Putative gene predicted by FgeneshB
219	2766	7854	2435494	E	Transcriptional regulator of met regulon-like
219	2766	8053	2435494	-	Putative gene predicted by FgeneshB
219	2769	8642	2435494	-	Putative gene predicted by FgeneshB
219	2769	8794	2435494	-	Putative gene predicted by FgeneshB
219	2769	8911	2435494	-	hypothetical protein
219	2769	8926	2435494	-	Putative gene predicted by FgeneshB
219	2770	8927	2435494	L	Resolvase-like
219	2770	8928	2435494	-	Putative gene predicted by FgeneshB
219	2770	8929	2435494	L	Recombinase
219	2770	8968	2435494	K	MT-A70
219	2770	8969	2435494	-	Putative gene predicted by FgeneshB
219	2770	8970	2435494	-	hypothetical protein
219	2770	9264	2435494	-	Putative serine protease
219	2770	9300	2435494	R	Patatin
219	2770	9301	2435494	-	hypothetical protein
219	2770	9302	2435494	-	Putative gene predicted by FgeneshB

219	2770	9303	2435494	K	response regulator receiver
219	2770	9319	2435494	-	hypothetical protein
219	2770	9320	2435494	-	Putative gene predicted by FgeneshB
219	2770	9321	2435494	-	hypothetical protein
219	2770	9322	2435494	-	Putative gene predicted by FgeneshB

Table S6. Mobile element distribution in FW106 compared to *Xanthomonas* species. Numbers of transposons for *Xanthomonas* species were obtained from the IMG database based on COG assignment.

Organism	Mobile Elements/Mb (COG)[a]	%LGT[b]
FW106 Metagenome (All Contigs)	10.4	-
FW106 Metagenome (Major Scaffolds[3])	11.7	7.1
X. axonopodis pv. *citri* str. 306	11.0	10.1
X. campestris pv. *campestris* str. 8004	9.1	12.2
X. campestris pv. *campestris* str. ATCC 33913	9.8	11.5
X. campestris pv. *vesicatoria* 85-10	17.2	9.9
X. oryzae KACC 10331	63.1	5.6
X. oryzae MAFF 311018	79.8	5.2

[a] COG functional groups used for counting: 0675, 1662, 1943, 2801, 2826, 2842, 2963, 3039, 3293, 3316, 3328, 3335, 3385, 3415, 3436, 3464, 3547, 3666, 3676, 3677, 4584, 4644, 5421, 5433, 5558 and 5659

[b] %LGT is the percentages of genes classified as putative alien genes by Colombo (Waack et al, 2006)

[c] Scaffolds > 100 kb

Table S7. Makeup of genomic islands identified on the major scaffold 224 by discriminant analysis (see also Figure S6). Shaded entries indicate the putative acetone carboxylase operon.

Contig	IMG GOID[a]	Length (bp)	Gene Function	PUTAL[b]	Donor Taxon[c]
Peak 1					
2742	2005742489	792	ABC-type multidrug transport system, permease component	N	-
2742	2005742490	1125	Two-component system sensor kinase	N	-
2742	2005742491	603	Two-component system LuxR-family response regulator	N	-
2742	2005742492	1011	Hypothetical peptidase	Y	-
2742	2005742493	588	Integrase	N	-
2742	2005742494	324	Insertion element IS401	N	-
2742	2005742495	168	Peptidase M50	Y	Chloroflexi
2742	2005742496	909	Putative gene predicted by FgeneshB	Y	-
2742	2005742497	1422	Hypothetical protein	Y	Actinobacteria
2742	2005742498	705	Protein of unknown function DUF181	Y	Actinobacteria
2742	2005742499	963	Nitroreductase	Y	-
2742	2005742500	738	ABC transporter related	Y	-
Peak 2					
2743	2005742527	891	Signal peptidase I	N	-
2743	2005742528	384	Hypothetical protein	N	-
2743	2005742529	660	Ribonuclease III	N	-

2743	2005742530	933	GTPase	N	-
2743	2005742531	729	DNA repair protein RecO	N	-
2743	2005742532	726	OmpR-Family response regulator	N	-

Peak 3

2754	2005743057	261	Putative gene predicted by FgeneshB	Y	Bacilli
2754	2005743058	192	Putative gene predicted by FgeneshB	Y	Actinobacteria
2754	2005743059	1029	NADP-dependent oxidoreductase	N	-
2754	2005743060	1566	Acyl-CoA synthetase	N	-
2754	2005743061	204	Acetyl-CoA acetyltransferase	Y	Actinobacteria
2754	2005743062	507	Acetone carboxylase γ-subunit	Y	Bacilli
2754	2005743063	2325	Acetone carboxylase α-subunit	N	-
2754	2005743064	2181	Acetone carboxylase β-subunit	N	-
2754	2005743065	1065	Hypothetical protein	N	-
2754	2005743066	1305	Hypothetical protein	N	-
2754	2005743067	2022	Transcriptional activator of acetoin/glycerol metabolism	N	-
2754	2005743068	1071	Transposase	N	-

[a] IMG Gene Object Identifier

[b] Putative alien genes (outlined in bold) as determined by SIGI-HMM

[c] Putative donor taxon as determined by SIGI-HMM

Supplementary Figures
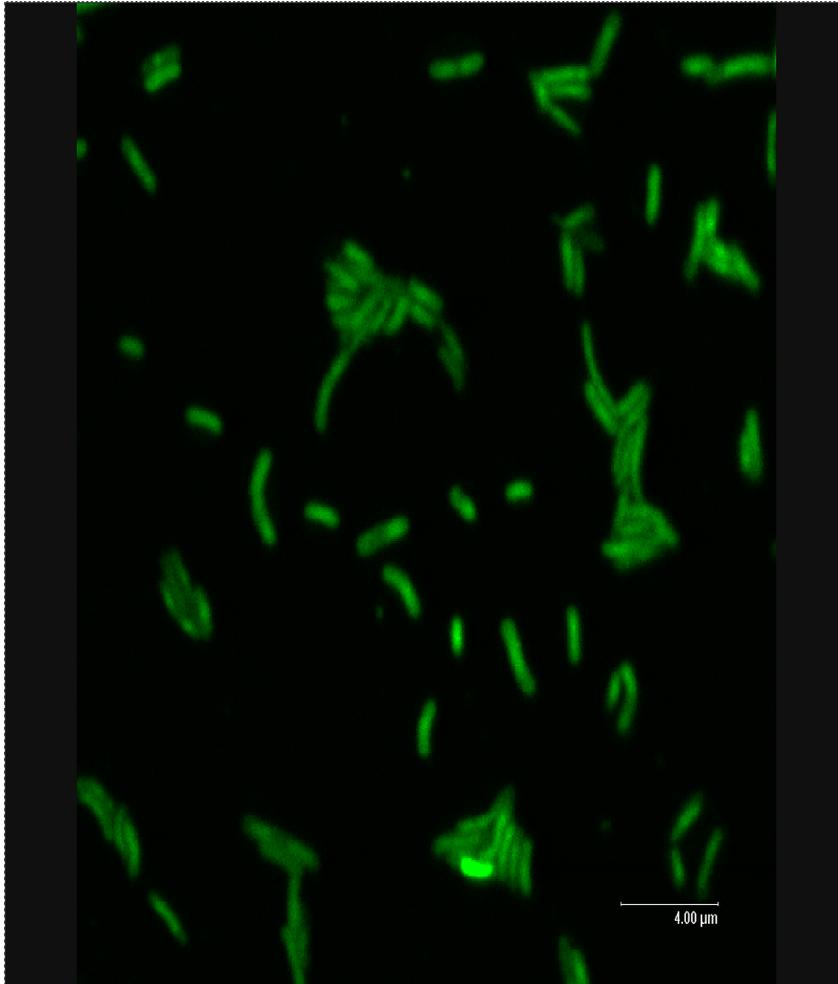
Figure S1. Confocal microscopy image of microbial cells obtained from FW106 groundwater.
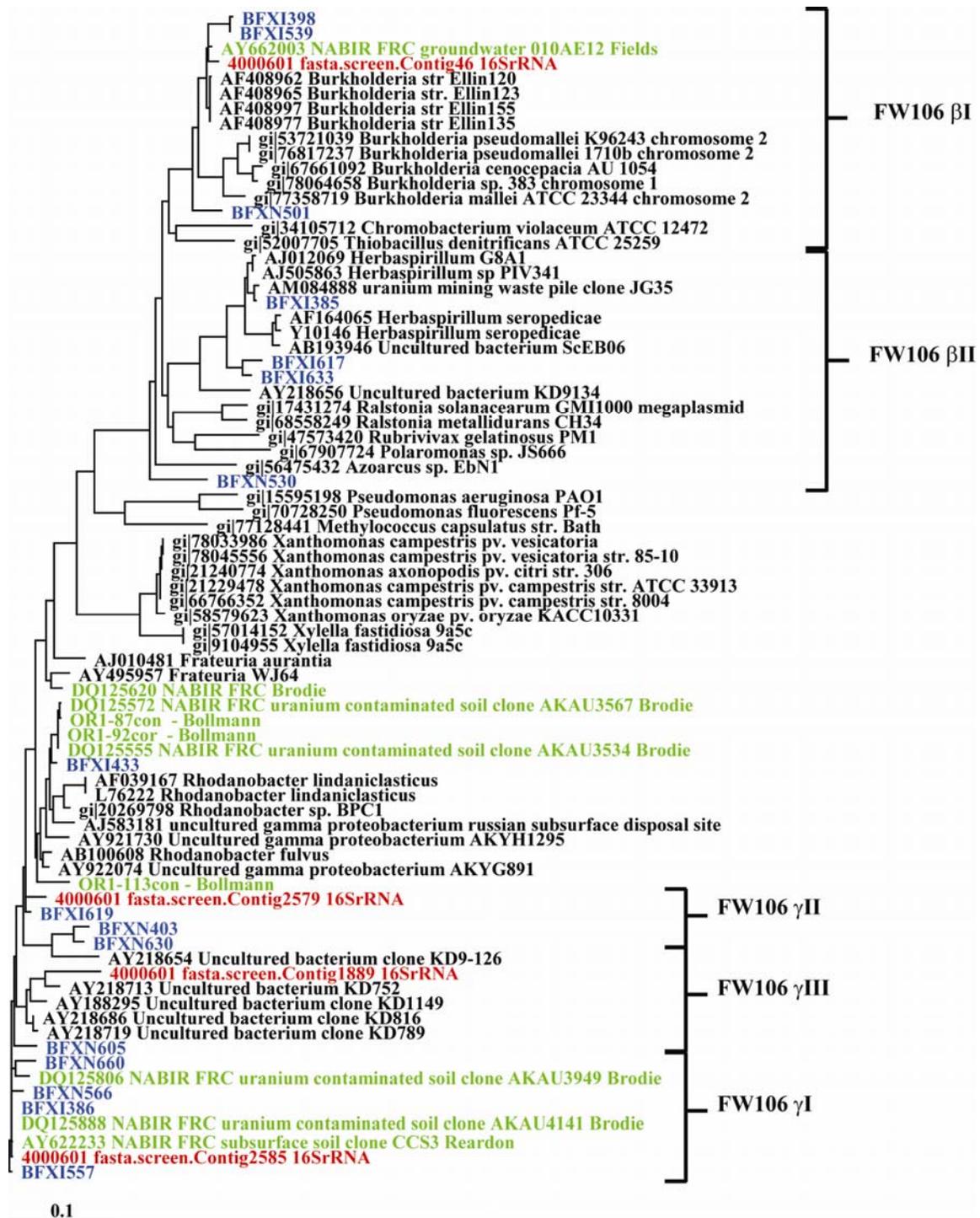
Figure S2

Figure S2. 16S rRNA phylogeny of FW106 used to define metagenomic bins. Entries are colored as follows: red, 16S rRNA gene fragments identified from the assembled metagenome; blue, FW106 16S rRNA genes identified from the OTU analysis described in Materials and Methods; green, 16S rRNA genes independently isolated from contaminated FRC sites or cloned from FRC isolates. The initial tree was constructed from the GreenGenes 16S rRNA dataset using the neighbor-joining methods of ARB and metagenomic fragments were added to the tree using the ARB parsimony insertion method (Ludwig et al, 2004). Phylogeny supports previous results suggesting dominance of the community by γ- and β-proteobacterial populations. A single 16S rRNA fragment was identified corresponding to an *Afipia*-like α-proteobacterial species (not shown) but no other sequences could be assigned to this bin. No evidence was found in the sample for sulfate-reducing or iron-reducing δ-proteobacterial species.
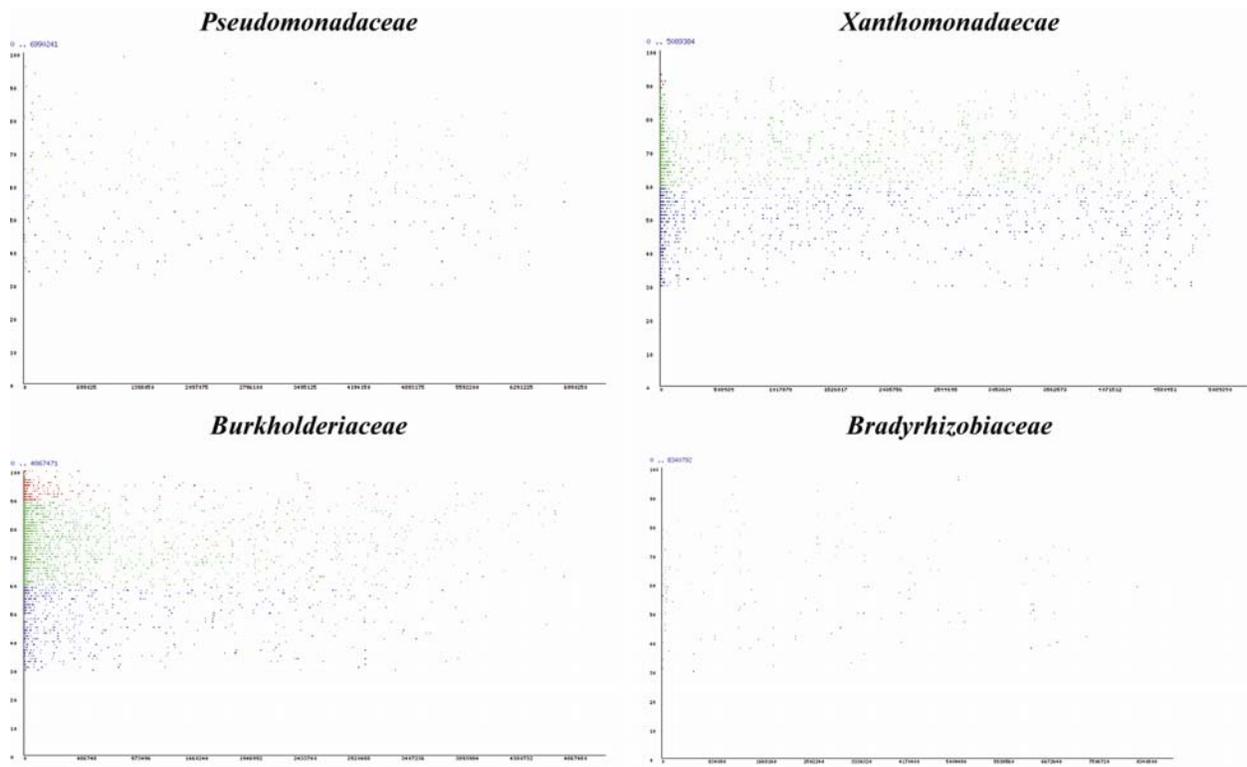
Figure S3

Figure S3. Protein recruitment plots for FW106. Colors are as follows: Red, >90% amino acid identity; Green, 60-90% aa identity; Blue, 30-60% aa identity.
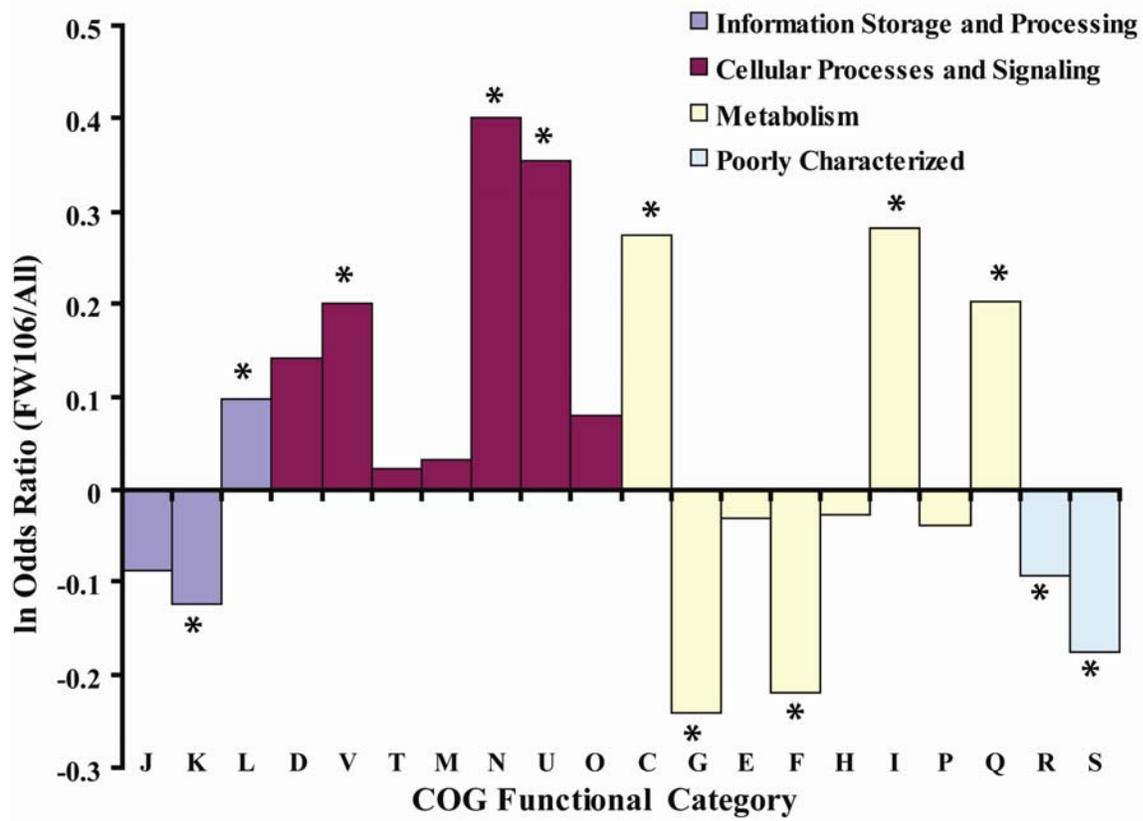
Figure S4

Figure S4. 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). COG categories are as follows: J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication, recombination and repair; D, Cell cycle control, cell division and chromosome partitioning; V, Defense mechanisms; T, Signal transduction mechanisms; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; U, Intracellular trafficking, secretion and vesicular transport; O, Posttranslational modification, protein turnover and chaperones; C, Energy production and conversion; G, Carbohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown.
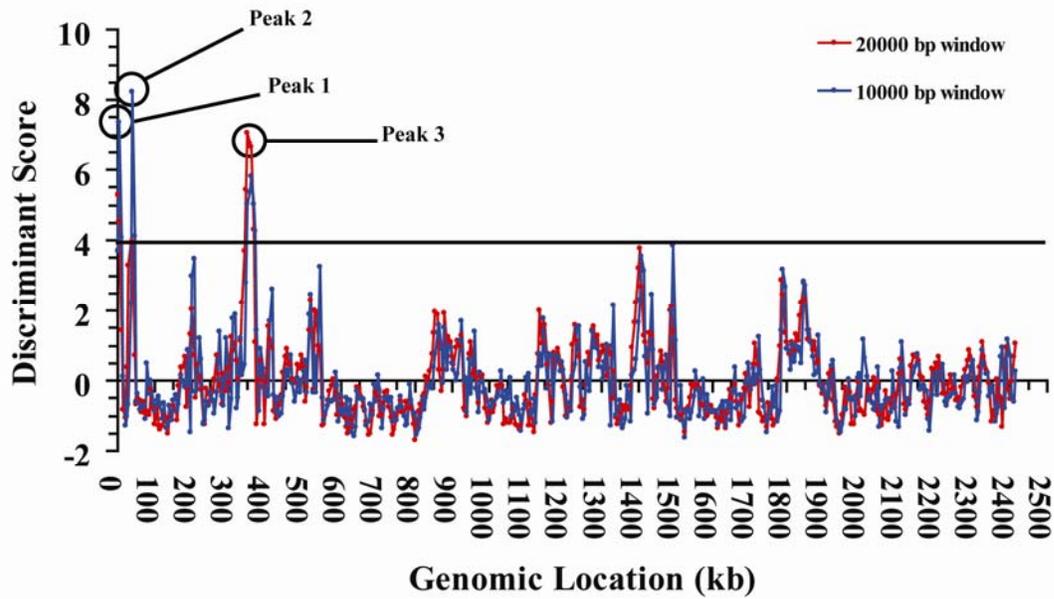
Figure S5

Figure S5. 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. COG categories are as follows: J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication, recombination and repair; D, Cell cycle control, cell division and chromosome partitioning; V, Defense mechanisms; T, Signal transduction mechanisms; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; U, Intracellular trafficking, secretion and vesicular transport; O, Posttranslational modification, protein turnover and chaperones; C, Energy production and conversion; G, Carbohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown.

Figure S6. Identification of genomic islands (GI) in the major scaffold 219 (FW106 γ1). GI's were determined by iterative discriminant analysis (Tu & Ding, 2003). Analyses were conducted using sliding windows of length 10000 (blue) and 20000 (red) bp advanced in 2500 bp steps. A discriminant scores cutoff of 3.9 was used to determine significance (Tu & Ding, 2003). Peak 3 corresponds to acetone carboxylase Operon A (Fig. 4).
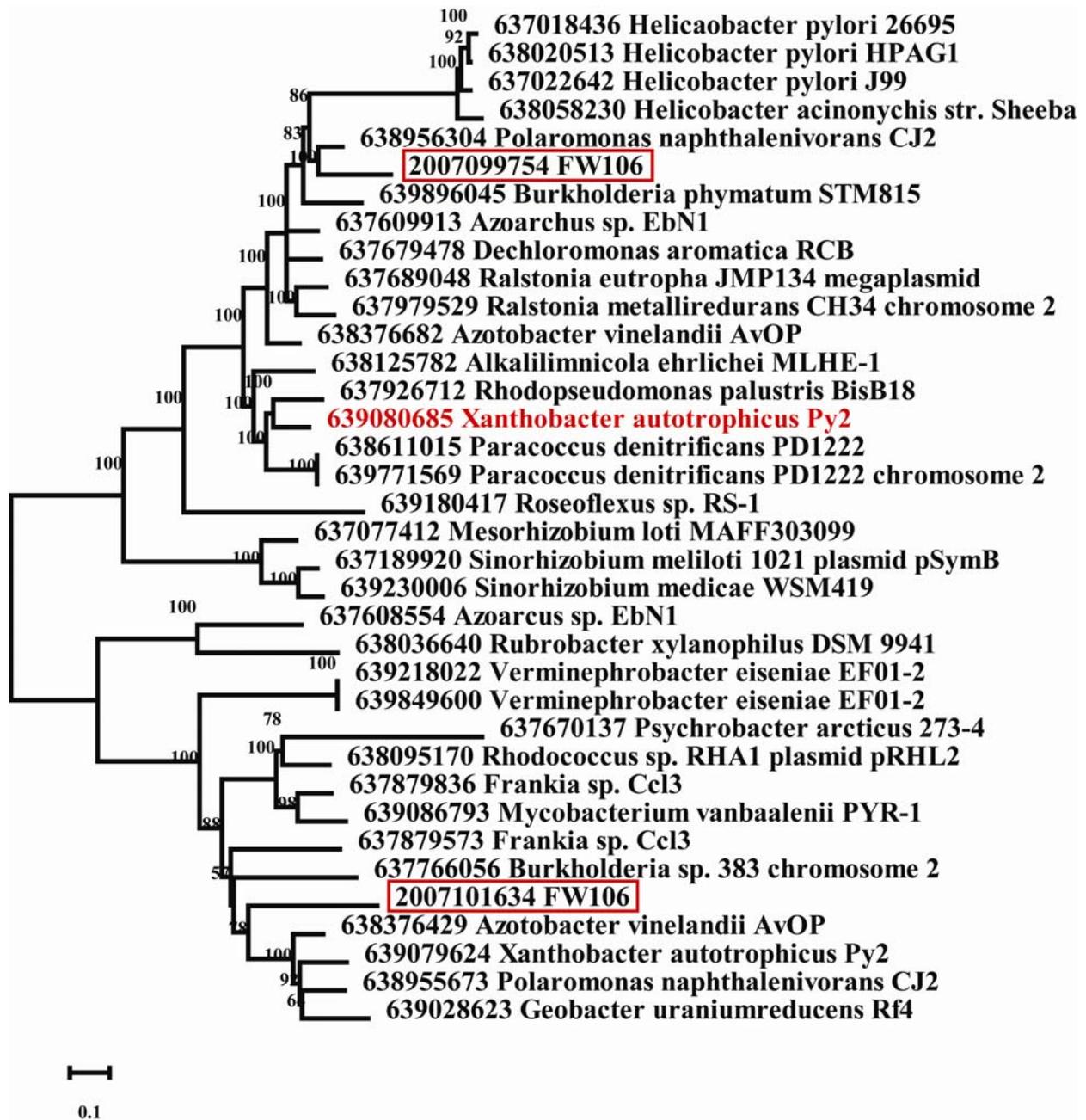
Figure S7

Figure S7. Phylogeny of concatenated nucleotide sequences of the α-, β- and γ- subunits of acetone carboxylase. Entries outlined in red represent FW106 operons. The functionally verified *Xanthobacter autotrophicus* acetone carboxylase in indicated in red text and is orthologous to Operon A (Fig. 4A). Protein alignments were converted to codon-based nucleotide alignments prior to tree construction. The evolutionary history was inferred using the Neighbor-Joining method (Saitou & Nei, 1987). The bootstrap consensus tree inferred from 500 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al, 2004) and are in the units of the number of base substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 2). The differences in the composition bias among sequences were considered in evolutionary comparisons. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 5400 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 (Tamura et al, 2007). Each branch is labelled with the IMG Gene Object Identifier of the α-subunit.
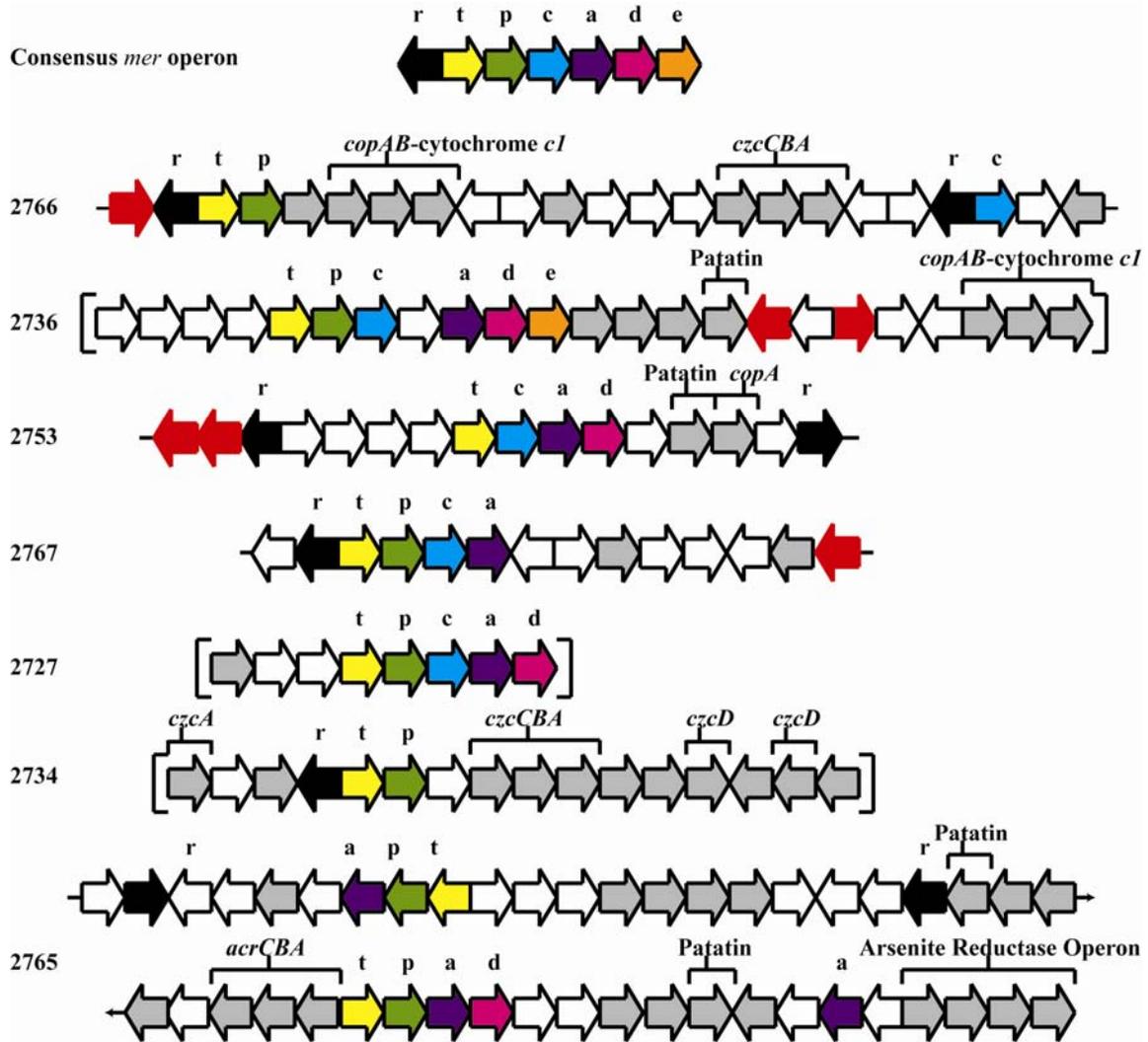
Figure S8

Figure S8. Presence and distribution of mercuric resistance genes of FW106. The consensus *mer* operon is shown at top with FW106 contigs aligned below (contig numbers are listed on the left). Genes are colored as follows: red, mobile elements; white, hypothetical proteins; grey, other functional protein-encoding genes. Orthologous FW106 *mer* genes are colored as in the consensus operon.
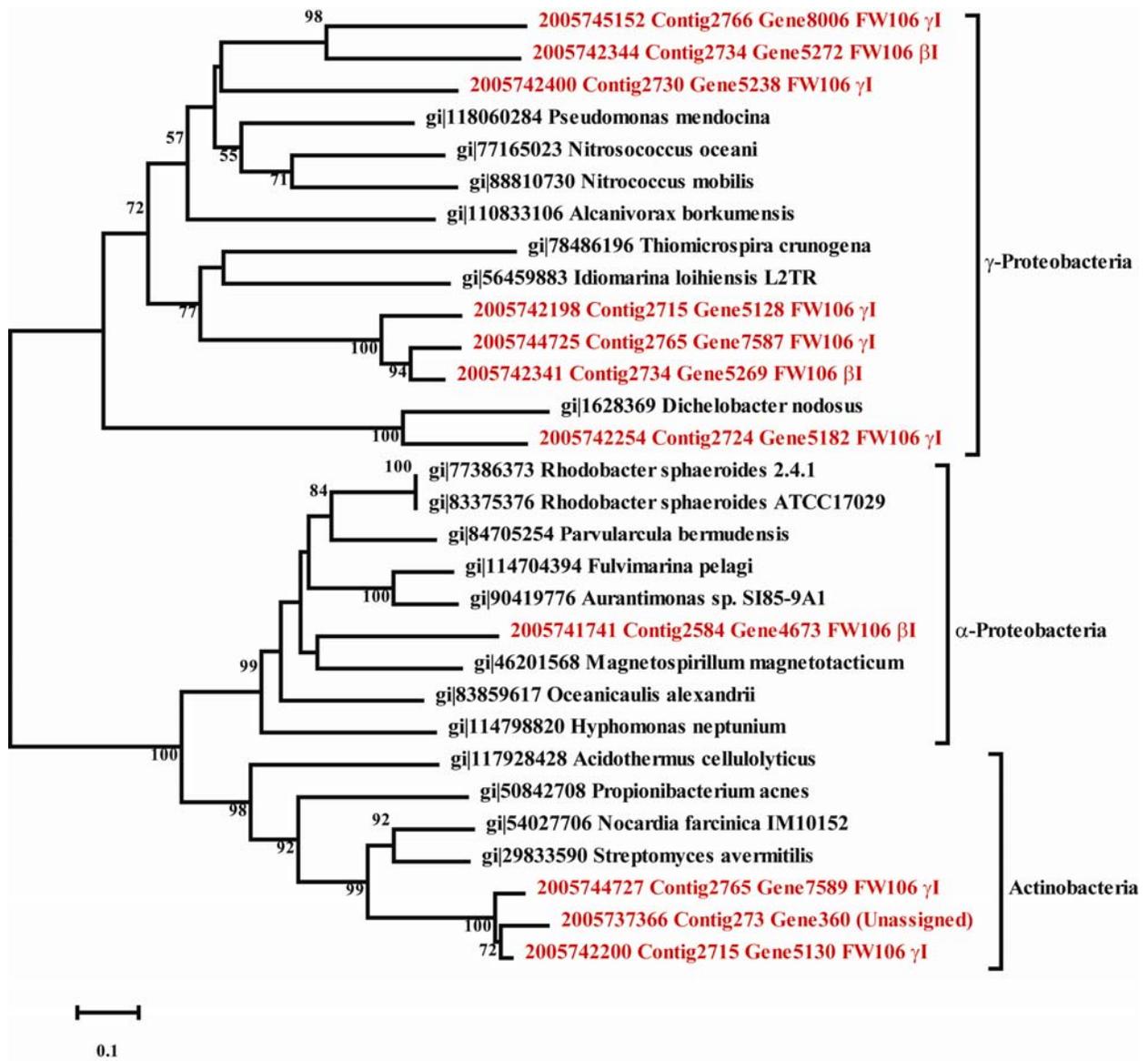
Figure S9

Figure S9. Phylogeny of CzcD-like metal efflux transporter protein sequences from FW106. Reference sequences are labelled with Genbank gi numbers. Additionally, FW106 sequences are colored red and are labelled with the IMG Gene Object Identifier number. The evolutionary history was inferred using the Neighbor-Joining method (Saitou & Nei, 1987). The bootstrap consensus tree inferred from 500 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl & Pauling, 1965) and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 144 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 (Tamura et al, 2007).

Supplementary References

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

Berriman M, Rutherford K (2003) Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4:** 124-132

Charlesworth D, Charlesworth B, Morgan MT (1995) The Pattern of Neutral Molecular Variation Under the Background Selection Model. *Genetics* **141:** 1619-1632

Chou H-H, Holmes MH (2001) DNA Sequence Quality Trimming and Vector Removal. *Bioinformatics* **17:** 1093-1104

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol* **72:** 5069-5072

Elshahed MS, Senko JM, Najar FZ, Kenton SM, Roe BA, Dewers TA, Spear JR, Krumholz LR (2003) Bacterial Diversity and Sulfur Cycling in a Mesophilic Sulfide-Rich Spring. *Appl Environ Microbiol* **69:** 5609-5621

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783-791

Hall T (2001) BioEdit version 5.0.6. *Department of Microbiology, North Carolina State University*

Kim Y (2006) Allele Frequency Distribution Under Recurrent Selective Sweeps. *Genetics* **172:** 1967-1978

Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. *Briefings in Bioinformatics* **5:** 150-163

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Forster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Konig A, Liss T 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, Chen IMA, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC (2007) IMG/M: a data management and analysis system for metagenomes. *Nucl Acids Res***:** gkm869

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

Merkl R (2004) SIGI: score-based identification of genomic islands. *BMC Bioinformatics* **5:** 22

Nei M, Gojobori T (1986) Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Mol Biol Evol* **3:** 418-426

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

Rosner B (2005) *Fundamentals of Biostatistics*, 6 edn.: Duxbury Press.

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

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-425

Schloss PD, Handelsman J (2004) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology* **71:** 1501-1506

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD 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***:** msm092

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences* **101:** 11030-11035

Tu Q, Ding D (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiology Letters* **221:** 269-275

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43

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

Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL (2006) Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. *PLoS Biol* **4:** e188

Yang Z (1997) PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood. *Computer Applications in BioSciences* **13:** 555-556

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62:** 316-322

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins, pp 97-166.